



JADT 2026 Proceedings

Volume II

Sessions of 9 and 10 July 2026

18th International Conference on Statistical Analysis of Textual Data

Palermo, Italy

8–10 July 2026

Conference Proceedings

Editors: Antonella Plaia, Mariangela Sciandra, Alessandro Albano

University of Palermo

<https://jadt2026.unipa.it>



Conference information

JADT 2026

18th International Conference on Statistical Analysis of Textual Data

Palermo, Italy
8–10 July 2026

<https://jadt2026.unipa.it>

These proceedings collect the contributions presented at JADT 2026, the 18th International Conference on Statistical Analysis of Textual Data, held in Palermo, Italy, from 8 to 10 July 2026.

The conference brings together scholars working on statistical analysis of textual data, text mining, natural language processing, corpus linguistics, digital humanities, social sciences, economics, official statistics, health-related text analytics, and related methodological fields.

The proceedings are organized in two volumes according to the conference programme:

- **Volume I:** sessions of 8 July 2026.
- **Volume II:** sessions of 9–10 July 2026.

Printed ISBN: 978-88-5509-882-3

Online ISBN: 978-88-5509-883-0

Scientific Committee

- Ramón Álvarez Esteban, Universidad de León, Spain
- Valérie Beaudouin, EHESS / Telecom Paris, France
- Marie Chandelier, Université Côte d'Azur, France
- Francesca Della Ratta, Istituto nazionale per l'analisi delle politiche pubbliche, Italy
- Anne Dister, UCLouvain Saint-Louis Bruxelles, Belgium
- Cédric Fairon, UCLouvain, Belgium
- Bénédicte Garnier, Institut national d'études démographiques, France
- Francesca Greco, University of Udine, Italy
- Serge Heiden, ENS de Lyon / CNRS IHRIM, France
- Domenica Fioredistella Iezzi, University of Rome Tor Vergata, Italy
- Margareta Kastberg, Université de Franche-Comté, France
- Ludovic Lebart, CNRS / Telecom Paris, France
- Jean-Marc Leblanc, Université Paris-Est Créteil, France
- Alain Lelu, Université de Franche-Comté, France
- Dominique Longrée, Université de Liège, Belgium
- Véronique Magri, Université Côte d'Azur / CNRS BCL, France
- Pascal Marchand, Université de Toulouse, France
- Damon Mayaffre, Université Côte d'Azur / CNRS BCL, France
- Michelangelo Misuraca, University of Salerno, Italy
- Bénédicte Pincemin, ENS de Lyon / CNRS IHRIM, France
- Céline Poudat, Université Côte d'Azur / CNRS BCL, France
- Pierre Ratinaud, Université de Toulouse, France
- Maria Spano, University of Naples Federico II, Italy
- Arjuna Tuzzi, University of Padua, Italy
- Mathieu Valette, Institut national des langues et civilisations orientales, France
- Laurent Vanni, Université Côte d'Azur / CNRS BCL, France

Organizing Committee

University of Palermo, Italy

- Antonella Plaia, Chair
- Mariangela Sciandra, Chair
- Antonio Abbruzzo
- Giada Adelfio
- Alessandro Albano
- Luigi Augugliaro
- Giovanni Boscaino
- Marcello Chiodi
- Nicoletta D'Angelo
- Chiara Di Maria
- Vincenzo Giuseppe Genova
- Angelo Mineo
- Vito Michele Rosario Muggeo
- Roberto Pirrone
- Andrea Simonetti
- Irene Siragusa
- Gianluca Sottile
- Michele Tumminello
- Erasmo Vassallo

Introduction

The International Conference on the Statistical Analysis of Textual Data, JADT (*Journées d'Analyse Statistique des Données Textuelles*), reached its 18th edition in 2026. For the first time, the conference was held in Palermo, Italy, from 8 to 10 July 2026, organized by StatPanormus and by the Department of Economics, Business and Statistics of the University of Palermo.

This biennial conference has continuously gained importance since its first edition in Barcelona in 1992, followed by the editions of Montpellier (1994), Rome (1996), Nice (1998), Lausanne (2000), Saint-Malo (2002), Louvain-la-Neuve (2004), Besançon (2006), Lyon (2008), Rome (2010), Liège (2012), Paris (2014), Nice (2016), Rome (2018), Toulouse (2020), Naples (2022), and Bruxelles (2024).

Every two years, JADT presents the state of the art concerning theories, problems, methods and applications in several domains sharing a quantitative approach to the study of lexical, textual, pragmatic and discursive features of information expressed in natural language.

In this edition, as already done in Naples in 2022, special attention was paid to young researchers through the 2026 VADISTAT – Per Simona Balbi Young Award. The award follows the attitude of our beloved colleague Simona Balbi, who strongly supported students and early-career researchers interested in the Statistical Analysis of Textual Data.

These proceedings collect 125 contributions delivered by more than 300 scholars from different countries. The papers address a wide range of issues open to scholars and researchers working in the field of textual data analysis, including textometry, lexicography, corpus linguistics, text mining, information science, digital humanities, social sciences, political discourse analysis, health-related text analytics, official statistics, economics, computational linguistics, embeddings and large language models.

The keynote speakers focused on central and emerging issues for textual data analysis, including open science in computational literary studies, the role of statistical analysis at a time of methodological transformation, and the relationship between linguistic competence, knowledge representation and large language models.

The proceedings follow the conference programme. Each volume includes the contributions associated with the corresponding conference sessions. Some contributions may appear in the programme but not in the proceedings, according to the authors' publication preferences or editorial requirements.

Acknowledgements

We would like to express our gratitude to all the authors, participants, reviewers, keynote speakers, specialized session organizers, session chairs and members of the Organizing Committee who contributed to the scientific quality and success of JADT 2026.

Keynote speakers

We warmly thank the keynote speakers for their valuable contribution to the conference:

- Christof Schöch, University of Trier, Germany
- Guido Vetere, Università Guglielmo Marconi and ISAGOG, Italy
- Arjuna Tuzzi, University of Padua, Italy

Specialized session organizers

Special thanks are due to the organizers of the specialized sessions, whose work contributed significantly to the richness and diversity of the scientific programme:

- Antonio Scalia, Banca d'Italia
- Orietta Luzi, ISTAT
- Davide Furceri, University of Palermo / International Monetary Fund
- Vincenzo Genova, University of Palermo
- Massimo Aria, University of Naples Federico II
- Francesca Greco, University of Udine
- Domenica Fioredistella Iezzi, University of Rome "Tor Vergata"

Reviewers

We are deeply grateful to the reviewers who offered their assistance in the selection and evaluation of the papers included in the conference programme and in these proceedings:

Antonino Abbruzzo
Alessandro Albano
Luigi Augugliaro
Alessandra Belfiore
Giovanni Boscaino
Maria Carmela Catone
Manuel Cobo
Nicoletta D'Angelo
Domenico De Stefano
Francesca Della Ratta
Anne Dister
Cédrick Fairon

Giada Adelfio
Massimo Aria
Valérie Beaudouin
Marion Bendinelli
Antonio Calcagni
Marie Chandelier
Alex Cucco
Luca D'Aniello
Emiliano Del Gobbo
Chiara Di Maria
Sascha Diwersy
Lara Fontanella

Augusto Gamuzza
Vincenzo Genova
Walter Giordano
Francesca Greco
Domenica Fioredistella Iezzi
Ludovic Lebart
Alain Lelu
Virginie Léthier
Pascal Marchand
Damon Mayaffre
Michelangelo Misuraca
Bénédicte Pincemin
Pierre Ratinaud
Annalina Sarra
Christof Schöch
Andrea Simonetti
Gianluca Sottile
Agostino Stavolo
Matilde Trevisani
Laurent Vanni
Ramón Álvarez Esteban

Bénédicte Garnier
Giuseppe Giordano
Agostino Gnasso
Serge Heiden
Margareta Kastberg
Jean-Marc Leblanc
Dominique Longrée
Véronique Magri
Raffaele Mattera
Rocco Mazza
Pasquale Pavone
Céline Poudat
Valentina Rizzoli
Germana Scepi
Andrea Sciandra
Irene Siragusa
Maria Spano
Venera Tomaselli
Arjuna Tuzzi
Emma Zavarrone

Institutional support

JADT 2026 was organized by StatPanormus and the Department of Economics, Business and Statistics of the University of Palermo, under the patronage and with the support of the institutions and associations represented in the conference materials.

We are particularly grateful to the Department of Economics, Business and Statistics of the University of Palermo, the University of Palermo, the Sicilian Region Assessorato dell'Istruzione e della Formazione Professionale, the Municipality of Palermo and the Italian Statistical Society for their support.

Finally, as regards the organization of the conference, we wish to thank all the members of the local Organizing Committee for their work and commitment.

Programme at a Glance

8 July 2026

Day 1

08:30

Welcome

09:15

Opening Session

Conference Chairs; Head of Department; SIS President; representatives of UNIPA, the Sicilian Region Assessorato dell'Istruzione e della Formazione Professionale, and the Municipality of Palermo.

10:00–11:00

Keynote talk

Arjuna Tuzzi

The Analysis of Textual Data at a Crossroads Between Tradition and Transformation: Are Statistics and Statisticians Still Crucial?

11:00–11:30

Coffee Break

11:30–13:00

Aula 11

S01

GLASS Group: Approaches to Textual Data in the Social Sciences

Aula 12

C15

Online communities, conflict and harmful discourse

Aula Multimediale C

C12

Corporate, finance and market intelligence

13:00–14:30

Lunch

14:30–16:00

Aula 11

S02

Emotional Text Mining

Aula 12

C09

Transformers, embeddings and LLM evaluation

Aula Multimediale C

C14

Political discourse and democratic communication

16:00–16:30

Coffee Break

16:30–18:00

Aula 11

S03

Embedding-based Approaches
to Text Categorization

Aula 12

C04

Institutional, literary and
conceptual corpora

Aula Multimediale C

C10

Multimodal, web and
user-generated content
analytics

19:00

Talea Orto Botanico

Welcome Cocktail

9 July 2026

Day 2

09:00–10:30

Aula 11

S04

Advanced Text Analytics for
Official Statistics: Methods,
Models, and Applications

Aula 12

C08

Topic modeling, classification
and interpretability

Aula Multimediale C

C17

Translation, stylistics and
literary variation

10:30–11:00

Coffee Break

11:00–12:00

Christof Schöch

Open Science in Computational Literary Studies

Keynote talk

12:00–13:30

Aula 11

S05

Quantitative Text Analysis
Methods for Scientific
Literature

Aula 12

C16

Andrea Camilleri and Sellerio
corpora

Aula Multimediale C

C11

Science, technology and
knowledge networks

13:30–15:00

Lunch

15:00–16:30

Aula 11

S06

Text Analysis in Economics

Aula 12

C03

Public, political and media
discourse

Aula Multimediale C

C02

French NLP, lexicon and
contrastive neology

16:30–17:00

Coffee Break

17:00–18:30

Aula 11

C01

Textometry, corpus methods
and visualization

Aula 12

C05

Italian institutional and
cultural corpora

Aula Multimediale C

C13

Health, clinical and
accessibility text analytics

19:45

Grand Hotel Piazza Borsa

Social Dinner

10 July 2026

Day 3

09:30–11:00

Aula 11

S07

Finance and Textual Analysis
Specialized session organized
with Banca d'Italia

Aula 12

C06

Digital tools, annotation and
corpus infrastructures

Aula Multimediale C

C07

Statistical methods and
pattern extraction

11:00–11:30

Coffee Break

11:30–12:30

Guido Vetere

Linguistic Competence and Large Language Models: Extracting Interpretable Lexical Knowledge

Keynote talk

12:30–12:45

2026 VADISTAT – Per Simona Balbi Young Award

Award Ceremony

12:45–13:00

Final remarks

Closing Ceremony

Contents

Conference information	ii
Scientific Committee	iii
Organizing Committee	iv
Introduction	v
Acknowledgements	vi
Programme at a Glance	viii
S04 – Specialized Session	
Advanced Text Analytics for Official Statistics: Methods, Models, and Applications	1
From textual agreements to structured data: using Large Language Models to enhance Official Statistics on anti-violence networks — <i>Elena Catanese, Mauro Bruno, Paolo Pizzo, Giuseppina Muratore and Claudia Villante</i>	2
Using AI techniques to support cause of death coding for Official Statistics — <i>Angela Pappagallo, Francesco Pugliese, Chiara Orsi, Tania Bracci, Francesco Grippo and Simone Navarra</i>	12
Semantic Search Approach to Occupation Coding in Official Statistics — <i>Giulio Massacci and Mauro Bruno</i>	22
Exploring User–Chatbot Communication: Methods, Insights, and Channel Differences — <i>Samanta Pietropaoli, Salvatore Agrillo, Paola Bosso, Barbara D’Amen and Gabriella Fazzi</i>	31
A Methodological Framework for Knowledge Graph Design Using Large Language Models — <i>Adriana Fidanza, Raffaella Maria Aracri, Alessio Buonocore and Roberta Radini</i>	40
C08 – Topic modeling, classification and interpretability	50
Editorial Transitions and Thematic Change in the American Journal of Sociology: A Structural Topic Modeling Analysis — <i>Stefano Sbalchiero, Valentina Rizzoli and Arjuna Tuzzi</i>	51
Titles, Abstracts, or Full-Texts? A Comparative Study of LDA and BERTopic Performance across Text Levels — <i>Francis Lareau and Christophe Malaterre</i>	60
A Multi-Method Validation Framework for Large-Scale Multilingual Text Analytics — <i>Stefano Blando and Domenica Fioredistella Iezzi</i>	70
Transforming UN Internal Audit Reports into Oversight Knowledge: Lexicometric Perspective — <i>Li Chen and Domenica Fioredistella Iezzi</i>	80

Comparing textometric approaches: Reinert-style Descending Hierarchical Classification versus probabilistic and neural topic modeling on a web-sourced corpus of French rap lyrics — <i>Robin Quillivic, Jules Dilé-Toustou and Pierre Ratinaud</i> . .	90
C17 – Translation, stylistics and literary variation	100
Demonstratives in human and machine translation between English and French — <i>Daniel Henkel</i>	101
Beyond Questions: A Corpus-Based Analysis of Interrogatives and Their Responses in Spoken French — <i>Yeo Jun Yun, Iris Eshkol-Taravella and Marianne Desmets</i> . .	111
The French Drama Revolution: Political Economy and Literary Production (1700-1900) — <i>Thiago Dumont Oliveira</i>	121
From Lyrics to Values: A Computational-Sociological Analysis of Cultural Change in Italian Popular Music — <i>Ciro Clemente De Falco and Emilia Romeo</i>	131
A Computational Authorship Attribution in the Children’s Tales of Oscar and Constance Wilde: The Case of “The Selfish Giant” — <i>Liviu P. Dinu, Alina Iacob and Cosmin Ciotloș</i> [not included in proceedings]	
S05 – Specialized Session	
Quantitative Text Analysis Methods for Scientific Literature	141
Mapping the Conceptual Landscape of Funded Computer Science Research in Spain — <i>Manuel Jesus Cobo Martín, Verónica Duarte Martínez, Luis M. Castillo-Chávez and Jose A. Moral-Munoz</i>	142
Modeling Emotion Constructs in Consumer Reviews: A Word2Vec-ELIta Hybrid Method Combined with PLS-SEM — <i>Cosima Valentino</i>	152
Mapping Communicative Purposes in Academic Discourse: designing an analytic linguistic pattern — <i>Alessandra Belfiore, Walter Giordano, Massimo Aria and Corrado Cuccurullo</i>	159
Enhancing leader-based community detection in scientific collaboration using publications metadata — <i>Geremia Sara and Domenico De Stefano</i>	168
Assessing BERTopic stability on a specialized scientific corpus: seed sensitivity, outlier reduction, and temporal semantic drift — <i>Carolina Vasconcelos, Sandro Mendonça and Bruno Damasio</i>	177
C16 – Andrea Camilleri and Sellerio corpora	185
Camilleri’s Sicilian: From Stopwords To Topic Modelling — <i>Vincenzo Giuseppe Genova and Andrea Sciandra</i>	186
Assunti teorico-metodologici su linguaggio letterario esclusivo e dati testuali — <i>Pietro Mazzarisi</i>	194
Classification of Andrea Camilleri’s works using bottom-up and top-down approaches. — <i>Andrea Sciandra and Michele A. Cortelazzo</i>	204
Distinctive Lexical Traits in Andrea Camilleri: A Keyness Measures Analysis of a Corpus of Sellerio Crime Novels — <i>Michele A. Cortelazzo, Matilde Trevisani and Arjuna Tuzzi</i>	213
A New Investigation for Sellerio — <i>Arjuna Tuzzi</i>	223
C11 – Science, technology and knowledge networks	233
Language as a Predictor of Scholarly Citation in European Policy Documents — <i>Francis Lareau, Jérémie Dion, Louis Renaud-Desjardins, Malvina Barra, Luisa Ripoll-Alberola and François Claveau</i>	234
Mapping the languages of science: a comparative cross-media textual analysis of scientific expertise. — <i>Gabriella Punziano, Rosanna Cataldo, Barbara Saracino and Ferdinando Iazzetta</i>	244

Discourse structure across technological maturity levels: a textometric approach — <i>Mélusine Caillard and Gaël Lejeune</i>	254
Structuring sectorial knowledge: automated detection of lexical relations — <i>Claudia Brunini, Eugenio Concetti and Giampiero D'Alessandro</i>	264
A hybrid rule-based and data-driven pipeline for author name normalization in IRIS institutional repositories — <i>Pasquale Pavone, Simone Righi and Sara Colombini</i>	274
Identifying Knowledge Gaps in Textual Networks: A Structural Holes Approach to Literature-Based Discovery — <i>Andrej Kastrin, Bojan Cestnik, Alja Nike Kastrin and Nada Lavrač</i>	284
S06 – Specialized Session	
Text Analysis in Economics	294
From Fed Speeches to Signals Embedding Central Bank Communication in Macroeconomic Forecasting — <i>Angelo Fama</i> ’	295
AI Meets Fiscal Policy: Mapping Government Spending Actions Across 64 Countries — <i>Davide Furceri</i> [not included in proceedings]	
Global and local effects of Federal Reserve communication: uncertainty and real activity — <i>Antonio Pietro Maria Morreale</i> [not included in proceedings]	
BILATERAL CONFLICT RISK AND TRADE: MILITARY WARS, TRADE WARS, AND DIPLOMATIC NOISE — <i>Jamel Saadaoui</i> [not included in proceedings]	
C03 – Public, political and media discourse	305
Détection des isotopies du discours présidentiel d’Emmanuel Macron : de la cooccurrence à la self-attention — <i>Damon Mayaffre and Laurent Vanni</i>	306
Les formes du futur en planification urbaine. Analyse textométrique des temps verbaux en planification régionale (Île-de-France, 1919-2024) — <i>Isabelle Chesneau</i>	316
Investigation de l’évolution des discours médiatiques à propos des réseaux socio-numériques à partir de la distance de Labbé et de sa version améliorée — <i>Pierre Ratinaud and Brigitte Sebbah</i>	326
Caractériser les commentaires en ligne : les émojis dans la loi « Duplomb » — <i>Pascal Marchand and Pierre Ratinaud</i>	336
Étudier les vidéos TikTok : comparaison lexicométrique des plus-values et des limites des retranscriptions automatiques versus les descriptions des vidéos. — <i>Panos Tsimpoukis, Lucie Loubere, Nikos Smyrnaïos, Pierre Ratinaud and Charlis Papaeangelou</i>	346
C02 – French NLP, lexicon and contrastive neology	354
Vers une correction automatique des formes inconnues : application du Masked Language Modeling de CamemBERT au corpus des Cahiers citoyens — <i>Sami Guem-bour, Solomiia Korol and Catherine Dominguès</i>	355
Séjour à l’étranger et développement des marqueurs discursifs en FLE : étude sur corpus chez les apprenants japonophones — <i>Kaori Sugiyama</i>	365
Le potentiel lexicographique des néologismes des TIC dans une perspective franco-slovaque : analyse par le biais de Sketch Engine — <i>Lucia Rackova</i>	375
Vers un chatbot basé sur la recherche sémantique et le raisonnement à base de règles pour l’optimisation des opérations de la Sécurité Sociale française — <i>Wenxiang Zhou, Touria Aït El Mekki, Frédéric Lardeux, Marc Legeay and Jean-Olivier Boesmier</i>	385
L’extraction semi-automatique de termes métaphoriques grâce à la spécificité sémantique : étude appliquée à un corpus français-italien de presse spécialisée dans le domaine de la bijouterie-joaillerie — <i>Marius François and Maria Grazia Massimo</i>	395

Utilisation des emprunts à l'anglais de type V-ING en espagnol, français et italien : une étude en diachronie courte basée sur les corpus du web du Wortschatz de Leipzig — <i>François Maniez, Maria Belén Villar Diaz and Sidonie Larato</i>	405
C01 – Textometry, corpus methods and visualization	415
Articuler méthodes quantitatives et épistémologies critiques en linguistique : enjeux, tensions et perspectives méthodologiques — <i>Tiago Joseph</i>	416
A la croisée des disciplines informatique et psychanalytique : une application sémio- tique de la méthode Reinert — <i>Claire Tissot</i>	426
De la mesure des paragraphes à la classification des visualisations : portée heuristique et démonstrative — <i>Claude Patricia Tardif</i>	436
De l'analyse de données en cartes : Visualisations géographiques et chronologiques de corpus textuels — <i>Jean-Marc Leblanc, Anne-Lise Humain-Lamoure and Marie Pérès-Leblanc</i>	446
Etude des cooccurrences du verbe Dire dans un corpus d'entretiens à propos d'un événement traumatique — <i>Anna Colli, Delphine Battistelli and Marie Chagnoux</i>	457
Europarser: exploiter de larges corpus de presse — <i>Marceau Hernandez, Gaël Lejeune, Félix Alié, Émile Rebours and Julien Bezançon</i>	467
C05 – Italian institutional and cultural corpora	477
La sfida del lavoro da remoto. Analisi di focus group a responsabili delle risorse umane — <i>Francesca della Ratta</i>	478
Un'analisi testuale dei discorsi parlamentari italiani — <i>Matteo Farnè and Marco Nannetti</i>	488
A Multilevel Analysis of Italy's Three-Year School Planning Documents Using a Lexical and Distributional-Semantic Framework — <i>Pasquale Pavone, Maria Teresa Sagri, Elettra Morini and Eniko Tolvay</i>	496
The Lexical Construction of Women's 'Image': Party Specificity and Gender Asym- metry in Italy's Constituent Assembly — <i>Francesca Aulino, Rosario D'Agata, Domenico De Stefano and Amin Gino Fabbrucci Barbagli</i>	505
Perceptions of Emilia-Romagna: a cross-linguistic analysis of keywords in English and French travel literature — <i>Rita Gramellini and Daniel Henkel</i>	515
C13 – Health, clinical and accessibility text analytics	525
Living with osteoporosis across contexts: a lexicometric comparison of Italian and Spanish narratives — <i>Chiara Tedesco, Ercole Vellone, Rosaria Alvaro, Gianluca Pucciarelli, Vicente Bernalte-Martí and Mariachiara Figura</i>	526
Self-care behaviours in people with osteoporosis: a multidimensional analysis of textual data — <i>Chiara Tedesco, Ercole Vellone, Rosaria Alvaro, Gianluca Pucciarelli, Vicente Bernalte-Martí and Mariachiara Figura</i>	536
Text Mining Applications in the Evaluation of AIFA Drug Innovativeness Reports in Italy — <i>Daniele Pirotta, Andrea Marletta, Carlotta Galeone and Paolo Mariani</i>	546
Characterization of texts written by deaf people through syntactic markers — <i>Mireille Esther Gettler Summa, Caroline Bogliotti and Anne Lacheret Dujour</i>	554
S07 – Specialized Session	
Finance and Textual Analysis	564
Interpretable Natural Language Concept Bottleneck Models for Default Risk estima- tion — <i>Andrea Conti and Giacomo Morelli</i> [not included in proceedings]	
Web Automatic Investigator (WAI): an Agentic Framework for Regulatory Multimodal Analysis of Web Contents — <i>Andrea Coletta, Miriam Corrado, Gabriele De Capoa, Oliver Giudice, Alessandro Maggi, Marco Mori, Anna Russo Russo and Valerio Signore</i>	565

Word2Prices: embedding central bank communications for inflation prediction — <i>Douglas Araujo, Nikola Bokan, Fabio Alberto Comazzi and Michele Lenza</i> [not included in proceedings]	
The cyber risk of non-financial firms — <i>Francesco Columba, Manuel Cugliari, Marco Orlandi and Federica Vassalli</i> [not included in proceedings]	
Processing instability and the value of information: from individual noise to common risk — <i>Rudraksh Kansal and Taejin Park</i> [not included in proceedings]	
C06 – Digital tools, annotation and corpus infrastructures	575
A Text analysis application: a library-led introductory course open to all University users — <i>Deborah Grbac and Valentina Schiariti</i>	576
JADT Anthology: decades of textual data analysis in one corpus. — <i>Laurent Vanni</i>	586
ActiveTigger: An open source collaborative text annotation software for computational social sciences — <i>Emilien Schultz, Julien Boelaert, Axel Morin, Emma Bonutti D’Agostini, Annina Claesson, Étienne Ollion and Arnault Chatelain</i>	596
ParaTaxe: a Suite of Tools for Processing and Publishing Multi-parallel Corpora — <i>Olivier Kraif, Cyrille François, Elnaz Jalilian, Natacha Rimasson Fertin and Pascale Roux</i>	606
Analyse de textualités multiples avec TXM : du texte brut aux éditions TEI, transcriptions synchronisées, traductions alignées et micro-textes — <i>Serge Heiden</i>	614
C07 – Statistical methods and pattern extraction	624
The UMCC: Usage and Measurement of Connectors in Context. A methodology and interface for measuring the use of connectors in different genres of corpora — <i>Corinne Rossari, Claudia Ricci, Linda Sanvido, Jonathan Schaber and Gerold Schneider</i>	625
The Motif Work Bench Pipeline: From Sequential Pattern Mining to Multidimensional Statistics — <i>Hugo Dumoulin and Timothée Premat</i>	635
The Moving Border Model: Applying Chronological Specificities to the Tuning of a Time Boundary — <i>Bénédicte Pincemin, Céline Guillot-Barbance and Alexei Lavrentiev</i>	645
Validation of Simultaneous Additive Trees in Textual Data Visualization — <i>Ludovic Lebart</i>	655
Pulling strings: scalable symbolic methods for Multi-Word Units detection in text corpora — <i>Antonio Balvet</i>	665
Estimating the typical-source distribution in imbalanced corpora — <i>Aris Xanthos</i>	675

S04

Specialized Session
Advanced Text Analytics for Official Statistics:
Methods, Models, and Applications

09-lug
09:00 - 10:30
Aula 11

From textual agreements to structured data: using Large Language Models to enhance Official Statistics on anti-violence networks

Elena Catanese¹, Mauro Bruno², Maria Giuseppina Muratore³, Paolo Pizzo⁴,
Claudia Villante⁵

¹Istituto Nazionale di Statistica – elena.catanese@istat.it

²Istituto Nazionale di Statistica – mauro.bruno@istat.it

³Istituto Nazionale di Statistica – muratore@istat.it

⁴Istituto Nazionale di Statistica – paolo.pizzo@istat.it

⁵Istituto Nazionale di Statistica – claudia.villante@istat.it

Abstract

This paper investigates the potential of Large Language Models (LLMs) to enhance official statistics on territorial anti-violence networks by complementing survey-based data with information extracted from formal agreements. Using evidence from the 2024–2025 Istat survey on networks combating violence against women in Italy, we compare structured questionnaire responses with data derived from protocols, memoranda, and agreements regulating local anti-violence governance.

We develop a multi-stage LLM-based information extraction pipeline to identify institutional actors, classify them according to a predefined taxonomy, and assign governance and operational roles based exclusively on explicit textual evidence. The extracted information is then systematically compared with survey responses to assess convergence, discrepancies, and implications for data quality.

Findings show that survey data tend to underestimate the number and diversity of actors involved in territorial networks, particularly in complex multi-actor agreements, while the LLM-based approach identifies a broader set of formal participants and governance structures. At the same time, the conservative design of the extraction pipeline reduces false positives by assigning key roles only when explicitly documented. The comparison highlights how text-based extraction can identify omissions, improve internal consistency, and enrich statistical representations of governance networks.

The study demonstrates the value of integrating LLM-based text analysis into official statistics workflows, showing how unstructured documentary sources can complement traditional survey instruments, reduce information gaps, and support more accurate monitoring of anti-violence governance systems.

Keywords: Large Language Models; Official Statistics; Information Extraction; Survey Data Integration; Violence Against Women; Territorial Governance

1. Introduction and research question

This paper arises from the intersection of two complementary methodological perspectives applied to the same object of study: territorial networks combating gender-based violence in Italy, with a specific focus on the subjects that promote these networks and the actors who participate in them.

The governance mandate introduced by Law 119/2013 has progressively translated, at both national and territorial levels, into a structured system of formal agreements that regulate anti-violence networks (Istat, 2025). These networks are established through formal acts—such

as protocols, agreements, memoranda, and conventions—through which different institutions define shared plans of action.

In 2025, Istat collected information on these formal acts through a structured survey administered to Regions and Autonomous Provinces, in partnership with the Department for Equal Opportunity (EOD) of the Italian Presidency of the Council of Ministers. The questionnaire gathers self-reported information from regional respondents on the composition, objectives, and formal characteristics of the networks operating on their territory. Each questionnaire is associated with a corresponding formal agreement, acquired by Istat as a scanned PDF document.

An initial textual analysis was performed to geo-locate the institutions involved in these networks. The present study extends this approach by aiming to identify and classify institutions in terms of their roles, functions, and attributes. In particular, the analysis distinguishes between *promoting subjects*—those who initiate the network and assume formal responsibility, shaping its governance structure—and *involved actors*—those who participate operationally without necessarily having promoted the agreement.

This work explores whether it is possible to automatically extract named institutions, their typologies, and their roles from unstructured textual documents using artificial intelligence techniques. It compares traditional survey data with an innovative text analysis approach based on LLMs, consistent with recent initiatives in OS (UNECE High-Level Group for the Modernisation of Official Statistics, 2023; United Nations Economic Commission for Europe, 2022).

Focusing on *promoting subjects* and *involved actors* is particularly relevant, as these categories directly reflect the capacity of the multi-level governance system to understand its own structure—namely, who is effectively part of the network, with what role, and with what level of responsibility. In this perspective, textual analysis can provide significant added value by complementing survey data, especially in contexts where monitoring the implementation of formal agreements is crucial.

Comparing how these two categories are represented across different data sources allows us to assess the ability of survey instruments to capture a distinction that is both analytically and policy-relevant. existing official statistics rely on self-reported survey data, but these may omit actors and governance roles embedded in formal agreements; therefore, we test whether LLM-based extraction can complement and validate survey information.

The study is guided by the following research questions:

1. Can LLM-based methods reliably extract actors and their roles from formal agreements?
2. How do the extracted data compare with survey-based information?
3. To what extent can this approach improve data quality and reduce respondent burden?

The remainder of the paper is structured as follows. Section 2. presents the background, Section 3. describes the data, Section 4. outlines the methodology, Section 5. discusses the results, and Section 6. concludes with implications and future developments.

2. Background: territorial networks as a tool for anti-violence governance

Strengthening assistance and support for women and their children is a priority objective of the Extraordinary Plan against sexual and gender-based violence and of Law no. 119 of 15 October

2013, which implements the principles of the Istanbul Convention.¹ Among the objectives set out in the law is the definition of a structured governance system across all levels of government, building on experiences and good practices already developed within local and territorial networks.²

In response to this regulatory framework, Italy has progressively developed an articulated system of formal agreements—including memoranda of understanding, conventions and framework agreements—that establish and regulate territorial anti-violence networks. The survey conducted by Istat in 2024–2025, in collaboration with the Department for Equal Opportunities (DPO), the Regions and the Autonomous Provinces, produced for the first time a systematic mapping of this system, identifying 251 territorial networks across 18 responding regions. This represents an unprecedented information asset, enabling observation of the morphology of the Italian anti-violence governance system in its aggregate dimension.

Yet mapping formal acts alone risks providing only a partial picture of this system. A memorandum of understanding is a legal act that establishes commitments, defines subjects and allocates responsibilities; however, signing a protocol does not necessarily guarantee either the effective participation of all listed subjects or the operational vitality of the network that originates from it. The gap between formal governance and substantive governance is widely discussed in the literature on anti-violence networks and represents one of the most critical issues in evaluating gender-based violence policies (Cimagalli, 2015; Deiana and Pellegrino, 2019; Kantola and Squires, 2012). Addressing this gap requires analytical tools that go beyond the collection of self-reported data and engage directly with the content of the formal acts produced by the networks.

3. Data Sources

The analysis relies on two main data sources.

The first data source is a structured questionnaire completed by representatives of the Regions and Autonomous Provinces as part of the Istat 2025 survey on territorial anti-violence networks. Not all regions participated in the survey; overall, 251 responses were collected. The questionnaire, implemented using a CAWI (Computer-Assisted Web Interviewing) approach, collects a wide range of information on the characteristics of the networks.

In this study, the analysis focuses specifically on the identification of institutions involved in the networks, their roles, and their typologies. For each network, respondents were asked to identify the subjects that promoted the formal agreement (i.e., those initiating the network), using a predefined list of 30 categories adopted for the Istat Antiviolence Centers Survey. Respondents were also asked to report the total number of actors involved and, where possible, to list them explicitly. Additional questions concerned the identification of actors, their classification according to the same categories, and the subjects responsible for coordination, monitoring, and management activities.

¹ Chapter II of the Istanbul Convention highlights the relevance of integrated policies and data collection in combating violence against women, recommending multi-level governance models in which “all relevant actors, including public authorities, human rights institutions and civil society organisations”, contribute to integrated and coordinated policies. Council of Europe. (2011). Council of Europe Convention on preventing and combating violence against women and domestic violence, CETS No. 210. <https://rm.coe.int/168008482e> In Italy, it was ratified by Law No. 77 of 27 June 2013, which authorised ratification and gave full effect to the Convention in the Italian legal order.

² Article 5 of Decree-Law no. 93 of 14 August 2013, converted, with amendments, into Law no. 119 of 15 October 2013.

The second data source consists of the formal documents attached to the questionnaire—protocols, agreements, memoranda, and conventions—which are typically provided as scanned images of signed acts. These documents exhibit a high degree of heterogeneity in terms of structure and editorial standards, as no common template is adopted. In some cases, the document explicitly lists the parties involved (e.g., through expressions such as “this protocol is between the following subjects”), sometimes the promoting subject is in “bold” character, while in others the roles of the actors are not clearly specified, making it difficult to distinguish between promoting and participating entities.

Textual analysis of these documents offers a richer level of detail compared to survey data. In particular, it enables the identification and geolocation of entities, as well as the extraction of additional attributes, such as the addresses of anti-violence centres or other institutions—information that is not collected through the questionnaire.

It must be noted that characterising these entities—i.e., understanding their role (promoting, involved, or supervising) and typology (CAV or RUNTS third-sector organizations)—may be challenging. In some cases, texts also refer to potential actors that are not listed by respondents. For example, statements such as “in the case of minors, the victim should go to the court” introduce entities that are not always reported in the questionnaire, but may nonetheless be considered part of the network.

4. Methodology

4.1. Survey carried out involving Italian Regions

The survey instrument was developed through a participatory process involving the Regions themselves, which preceded the administration of the questionnaire. The primary objective of this process was to establish a shared nomenclature articulated around two dimensions: the classification of network promoters — that is, the institutional and non-institutional actors assuming a driving and coordinating role in the establishment of territorial networks — and the classification of the actors involved in network activation, encompassing public institutions, third-sector organizations and specialist bodies.

The co-construction of a shared nomenclature made it possible to overcome the terminological and classificatory differences present across the various regional experiences, ensuring the comparability of the data collected and their interpretability within a systemic framework at the national level.

4.2. Survey-control data editing

Even though the questionnaire was developed through an extensive collaborative process involving stakeholders and institutional actors engaged in violence against women interventions, limitations inherent to self-reported survey data remain. Respondents may interpret classification categories differently, omit information, or introduce inconsistencies between related items — limitations that are particularly salient in a domain as complex and institutionally varied as anti-violence network governance.

To address these quality concerns, a second-stage validation procedure was implemented using a Large Language Model (LLM)-based pipeline. The formal agreements submitted by respondents as scanned PDF documents within the CAWI system were processed through an extraction pipeline built on OpenAI models. Using semantic search and entity extraction — guided by prompts derived directly from the questionnaire’s classification modalities — the pipeline

identified actors, subjects and their roles as described in the original texts, and cross-referenced these against the corresponding survey responses. Where discrepancies were detected between the textual content of the formal agreements and the self-reported answers, these were flagged for review, enabling the research team to resolve inconsistencies and complement item non-responses with information extracted directly from the source documents.

This approach serves a dual purpose: it functions simultaneously as a quality assessment tool for the survey data and as a validation framework for the LLM-based extraction methodology itself, demonstrating the potential of AI-assisted text mining to enhance data quality in official statistics on gender-based violence.

4.3. LLM-based information extraction pipeline

To complement survey data, an automated text analysis approach was developed to extract structured information from formal agreements. The method relies on a multi-step information extraction pipeline combining semantic search techniques and transformer-based language models (Devlin, 2019), which have recently shown strong capabilities in processing unstructured text and supporting information extraction tasks (Brown et al., 2020).

The goal of the pipeline is to identify relevant institutional actors, classify them according to predefined categories, and assign roles based on explicit textual evidence. The extraction process is designed to prioritise precision and interpretability, avoiding implicit inferences and relying only on information explicitly stated in the documents, in line with recent approaches emphasising controllability and reliability in LLM-based systems (Bommasani, 2021).

In the first stage, the pipeline identifies (i) the formal signatories of the protocol and (ii) all institutional and organisational actors mentioned in the text. Signatories are defined as entities explicitly listed as parties formally subscribing to the agreement, typically appearing in the opening or closing sections of the document (e.g., lists introduced by expressions such as “between”, “the parties”, or “signed by”).

All remaining entities are extracted as *subjects*, including public institutions, service providers, associations, and other organisations mentioned in any capacity. Each subject is recorded using a structured schema (name, type, municipality, address, lead role, and notes).

Classification relies on a predefined controlled vocabulary (the 30 categories adopted in the survey) combined with a hierarchical set of rules to ensure consistency across cases. To maximise precision, the extraction process is restricted to clearly identifiable organisations: only entities explicitly mentioned in the text are retained, no attributes are inferred, and duplicates are removed. Entities that cannot be reliably interpreted as organisations (e.g., individuals, legal references, or abstract concepts) are excluded.

The system also addresses heterogeneity in naming conventions by normalising alternative designations referring to the same entity (e.g., “ASL”, “Azienda Sanitaria Locale”, or specific names of territorial health authorities).

In the second stage, the subset of *proposing actors* is identified. These are defined as entities that explicitly initiate or promote the protocol. Selection is restricted to the previously extracted subjects and relies solely on explicit textual evidence (e.g., expressions such as “promoted by” or “initiated by”). No inference is applied at this stage: entities are not classified as proposing actors based on their role as signatories or coordinators unless this function is clearly stated in the text.

In the third stage, *actors involved* in the operational implementation of the protocol are iden-

tified. These include entities participating in service delivery, collaboration networks, or intervention activities. Identification relies on explicit indicators of operational roles (e.g., participation, collaboration, service provision, case management), and is again restricted to the set of previously extracted subjects.

This step also supports the disambiguation between different organisational typologies, particularly between CAV and third-sector organisations listed in the RUNTS. In cases where an organisation is formally classified as a non-profit entity but is described in the text as responsible for providing support to victims of violence, it is reclassified as an anti-violence centre based on its functional role. This rule-based reinterpretation ensures consistency between formal classifications and the operational roles described in the protocols.

In the final stage, an enrichment step assigns roles to each extracted subject by matching them against the lists of signatories, proposing actors, and involved actors. This matching process combines normalized string comparison, rule-based alias generation (to account for variations in naming conventions), and optional fuzzy matching with a conservative threshold to minimise false positives.

Within this framework, a further analytical distinction is introduced between *actors involved* and *governance actors*. The former includes all entities participating in operational activities, whereas the latter refers to entities exercising coordination, strategic oversight, or institutional responsibility (e.g., lead organisations or coordinators when explicitly indicated). These roles are captured through specific attributes (e.g., “lead entity”) and annotations in the structured data, rather than inferred automatically.

Across all stages, conservative extraction rules are applied to minimise false positives: entities are retained only when clearly supported by the text, duplicates are removed, and no assumptions or implicit inferences are introduced. The final output consists of structured lists of signatories, subjects, proposing actors, and involved actors for each protocol. All extracted entities are assigned a role based on explicit textual evidence; in the absence of such evidence, they are classified as involved actors. This structured representation enables a systematic comparison between survey-based and text-derived information, supporting the analysis presented in the next section.

5. Results

This section presents the main findings of the analysis, focusing on areas of convergence and divergence between survey-based and text-derived representations of territorial networks. The results provide insights into both the structure of anti-violence governance and the quality of the information collected through different data sources.

5.1. Results from survey

Figure 1 displays the distribution of network actors by institutional category, distinguishing between promoting bodies — those who take the initiative in establishing territorial anti-violence networks — and involved actors, who are subsequently called upon to participate in and contribute to the network once constituted.

The data reveals that Anti-violence Centres and shelters represent the most frequent promoters (203), confirming their central role as drivers of network governance at the local level, as also recommended by the Istanbul Convention.

The distribution of actors involved in the protocols reveals a hierarchy of presence that reflects,

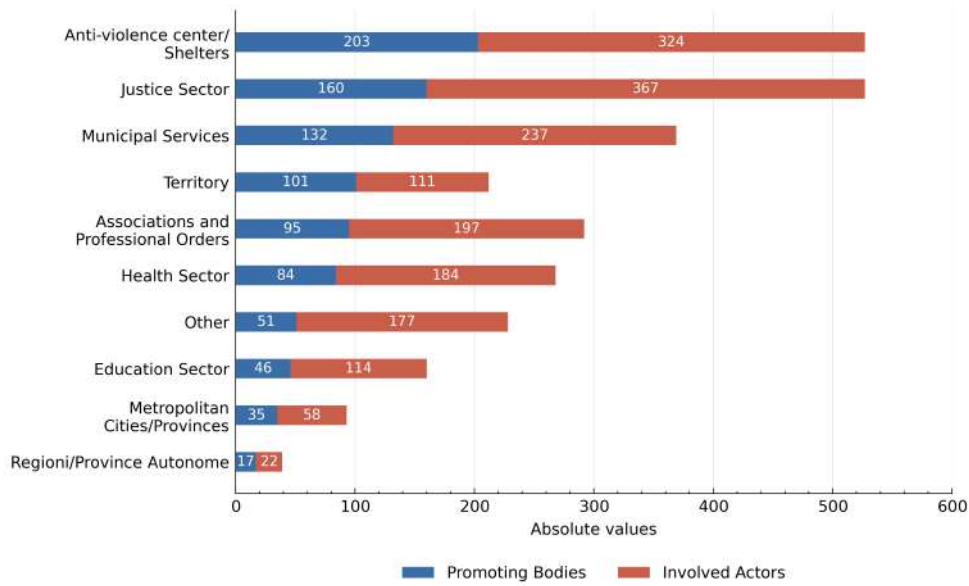


Figure 1: Promoting bodies and actors involved in territorial protocols/agreements by type. Year 2025 (Absolute values)

but does not perfectly replicate, that of the promoting subjects. The judicial sector is the most broadly represented actor, with 367 presences out of 1,791 actors overall recorded (20.5% of the total): Prefectures, Courts, Public Prosecutors’ Offices, and bar associations build around the networks a legal protection architecture that goes beyond their role as promoters alone. CAV and Shelters, taken jointly, account for 18.1% of actors, confirming their centrality not only as subjects that initiate networks but as operational nodes actively involved in the care of victims. The gap between the position of municipal services among the promoters (132 occurrences) and their presence as involved actors (237) signals a broadening effect of the network at the implementation stage: many bodies enter the operational ecosystem of the network without having signed the founding agreement. This finding is methodologically significant: it suggests that the number of signatory subjects systematically underestimates the plurality of actors that effectively inhabit territorial networks. More information is provided in Figure 2.

Figure 2 shows the regional distribution of promoting bodies — those who take the initiative in establishing territorial anti-violence networks — and involved actors, those subsequently called upon to participate in them. In the large majority of regions involved actors consistently outnumber promoting bodies — a structural feature consistent with the logic of anti-violence network governance, in which coordinating institutions activate a broader constellation of public agencies, service providers and civil society organizations around a shared governance framework. It has to be mentioned that Umbria, Valle d’Aosta and Sardegna have only one proponent body.

5.2. Comparison between survey data and LLM-based extraction

Figure 3 shows the distribution of entities by type and role identified through the LLM-based extraction pipeline.

The results show that the number of protocol signatories identified in the texts (2,689) exceeds the total number of promoting and involved actors reported in the survey (2,413). This suggests that survey responses may underestimate the full set of stakeholders formally participating in territorial agreements.

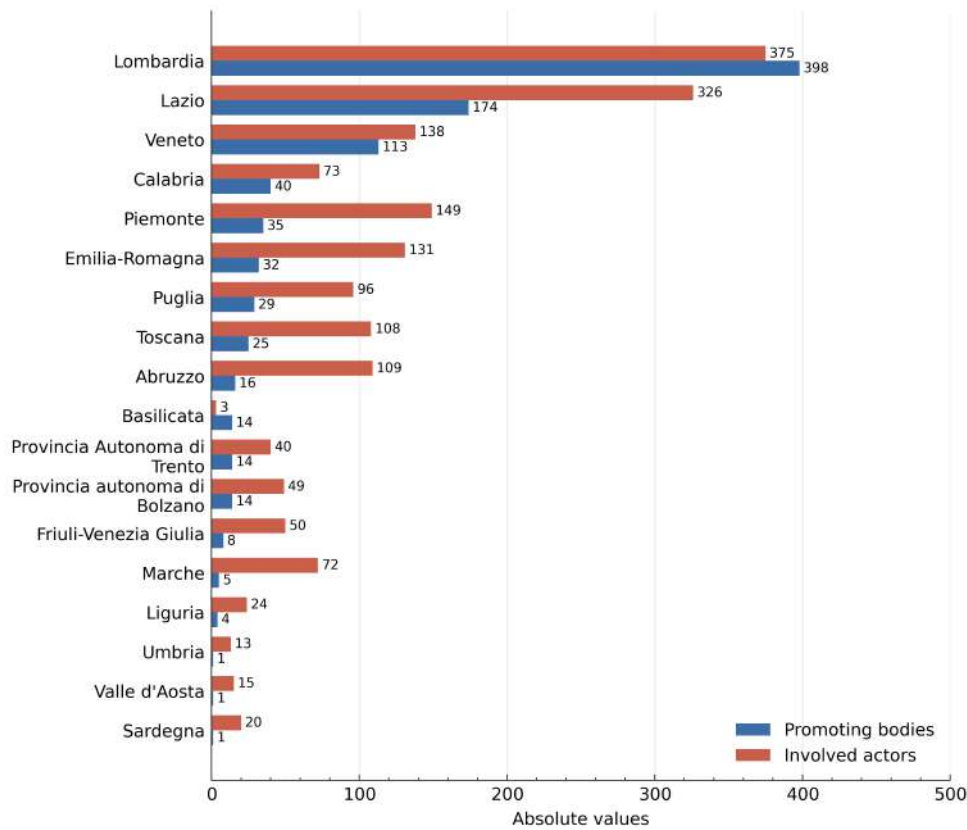


Figure 2: Source: Istat, *Survey on Territorial Networks against Violence against Women, 2025*

More broadly, the total number of actors identified in the protocols is substantially higher than that recorded through the questionnaire. The largest discrepancies concern municipalities and local public bodies. A plausible explanation is that, when a large number of municipalities are involved in the same agreement, respondents tend to report aggregated counts rather than listing each entity individually.

By contrast, the number of promoting subjects identified through the extraction pipeline (586) is lower than the survey estimate (808). This result is consistent with the conservative design of the method, which classifies promoters only when explicit textual evidence is available, thereby minimising false positives.

The number of governance actors identified (367 across 251 agreements) is consistent with the expectation that, in most cases, a single lead institution is responsible for coordination. Here, governance actors refer to entities responsible for coordination, monitoring, or strategic oversight when such roles are explicitly stated in the text. For some typologies, such as CAV and the residual category “other”, the number of governance actors exceeds that of explicitly identified promoters. This suggests that coordinating institutions may effectively play a promoting role even when this function is not formally stated in the protocol.

To further investigate these differences, signatories identified in the texts (2,689) were compared with the total number of actors reported in the survey (2,413) at the regional level (Figure 4). The extraction-based estimates are slightly lower than survey counts only in Lazio and Veneto, while higher values are observed for Lombardia, Emilia-Romagna, Toscana, Puglia, Abruzzo, Liguria and Marche. For the remaining regions, the two sources show a high degree of consistency.

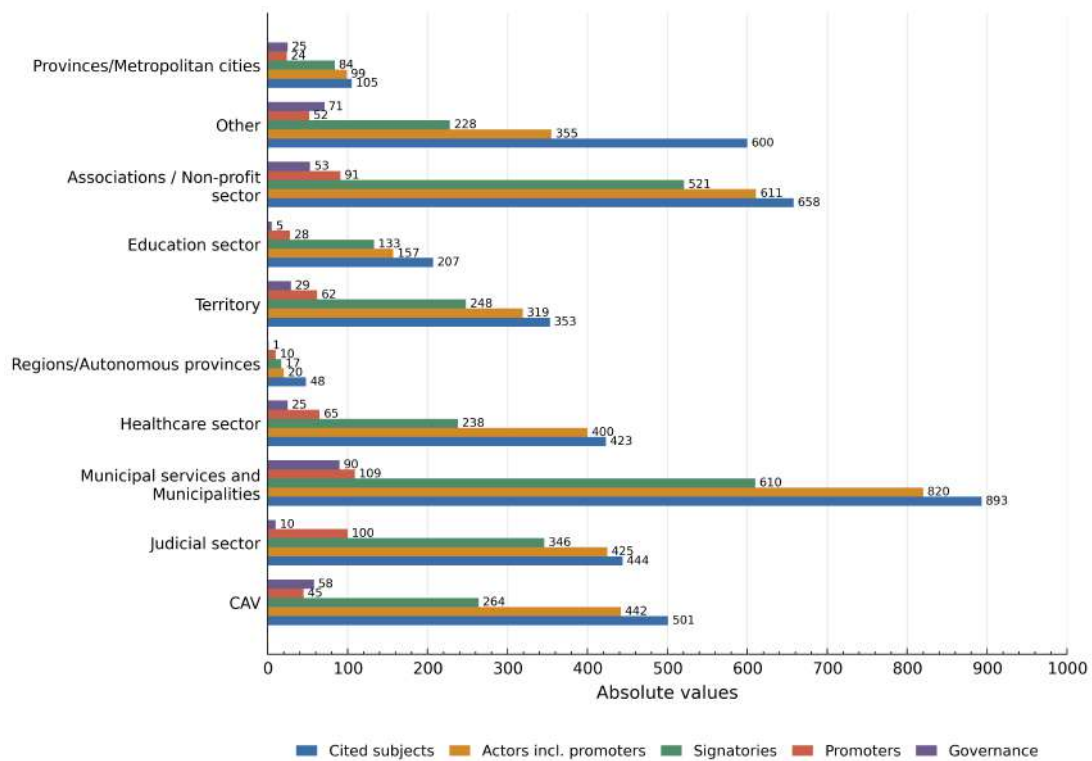


Figure 3: Entities identified in territorial protocols and agreements by type and role through the LLM-based extraction pipeline. Year 2025. Absolute values.

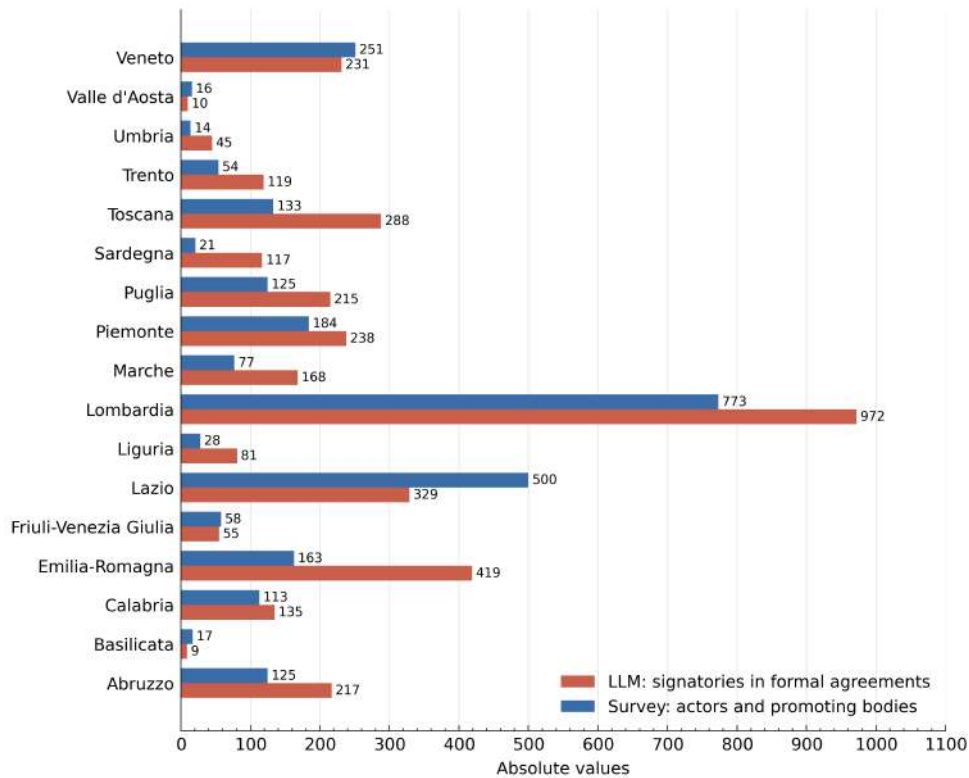


Figure 4: Comparison between signatories identified in formal agreements and actors reported in the survey, by region. Year 2025. Absolute values.

6. Conclusions

This paper examines the extent to which survey-based representations of territorial anti-violence networks align with those obtained through direct analysis of formal agreements, and what divergences between the two sources reveal about network structure and data quality. The findings provide both substantive and methodological contributions. From a substantive perspective, the study offers new evidence on the governance of anti-violence networks in Italy. The 2024–2025 ISTAT survey represents the first systematic national mapping of these networks, highlighting the plurality of institutional and non-institutional actors involved.

From a methodological perspective, the paper demonstrates the potential of LLMs for OS. The proposed extraction pipeline shows how complex and heterogeneous textual documents can be transformed into structured statistical information. The comparison between extracted entities and survey responses suggests that text-based approaches can complement traditional data collection by identifying omissions, reducing inconsistencies, and capturing a broader set of actors.

Overall, integrating survey data with documentary text analysis helps bridge the gap between formal and observed governance. While structured questionnaires ensure comparability and standardisation, text analysis adds contextual depth that would otherwise remain inaccessible.

Future research will focus on protocol-level micro-analysis to better understand discrepancies between sources, improve classification accuracy, and refine the extraction process, with a specific focus on network monitoring activities. Additional developments include more precise georeferencing of all network actors (including full addresses), extending the pipeline to other types of agreements, and conducting longitudinal analyses of network evolution over time.

References

- Bommasani R. e. a. (2021). On the opportunities and risks of foundation models. *arXiv:2108.07258*.
- Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., and Amodei D. (2020). Language models are few-shot learners. 33:1877–1901.
- Cimagalli F., editor (2015). *Le reti anti violenza: Esperienze, metodologie e prospettive*. Carocci.
- Deiana G. and Pellegrino S. (2019). Le reti territoriali contro la violenza di genere: Struttura, funzioni e criticità. *La Rivista delle Politiche Sociali*, (2):87–104.
- Devlin J. e. a. (2019). Bert: Pre-training of deep bidirectional transformers. *NAACL*.
- Istat (2025). Prevenire e combattere la violenza contro le donne: le reti territoriali. anno 2025.
- Kantola J. and Squires J. (2012). From state feminism to market feminism? *International Political Science Review*, 33(4):382–400.
- UNECE High-Level Group for the Modernisation of Official Statistics (2023). Large language models for official statistics.
- United Nations Economic Commission for Europe (2022). Machine learning for official statistics.

Using AI techniques to support cause of death coding for Official Statistics

Angela Pappagallo¹, Francesco Pugliese², Chiara Orsi³, Tania Bracci⁴,
Francesco Grippo⁵, Simone Navarra⁶,

¹Italian Institute of Statistics (Istat) – angela.pappagallo@istat.it

²Italian Institute of Statistics (Istat) – francesco.pugliese@istat.it

³Italian Institute of Statistics (Istat) – chiara.orsi@istat.it

⁴Italian Institute of Statistics (Istat) – tania.bracci@istat.it

⁵Italian Institute of Statistics (Istat) – francesco.grippo@istat.it

⁶Italian Institute of Statistics (Istat) – simone.navarra@istat.it

Abstract

The Italian National Institute of Statistics (Istat) annually produces cause-specific mortality statistics based on death certificates, completed by certifying physicians for each death occurring in Italy. Medical expressions reported on each certificate are coded using the World Health Organization's International Classification of Diseases (ICD-10), and the underlying cause of death is selected for official statistics and international comparisons, in accordance with the provisions of ICD-10. Coding process and selection of the underlying cause are performed using the Iris software. A first module of Iris assigns the corresponding ICD-10 code to each medical expression reported on the certificate, performing text recognition based on a dictionary of medical terms. An additional module selects the underlying cause. Each year, approximately 670,000 death certificates are processed. About 80% are fully coded automatically by Iris, while approximately 135,000 require manual coding. Most cases not fully coded by Iris are due to failures in medical text recognition (the first module). This inspired the idea of leveraging Artificial Intelligence (AI) techniques to improve the recognition and coding of medical terminology on death certificates, supported also by the successful experience of AI application to cause-of-death (CoD) data production in France. The objective of the present study was to identify AI models suitable for short and noisy medical text. The models were designed to support the dictionary-based step, and their performance was evaluated in terms of coding accuracy and automatic coverage.

A curated selection of supervised machine learning models was implemented to perform the coding task, with consideration for state-of-the-art deep learning models (e.g., BiLSTM, BERT) and "open-weight" Large Language Models (LLMs), including LLaMA 1B. The experimental results indicate that AI approaches offer a promising pathway towards more efficient and reliable CoD, although there is room for improvement, particularly for terms and expressions absent from the training set.

Keywords: causes of death, natural language processing, deep learning, artificial intelligence

1. Introduction

Cause-of-death statistics represent one of the most important sources of information for monitoring the health status of the population and guiding public health policies. At the core of their production lies the classification of causes of death, a step that is critical for ensuring both the quality and the comparability of statistics across time and across countries.

In Italy, the National Institute of Statistics (Istat) is responsible for collecting, processing, and publishing cause-specific mortality statistics on an annual basis. The information is derived from the Istat paper-based form, also referred to as death certificate, which follows the standard recommended by the World Health Organization (WHO). The death certificate includes several lines on which a certifying physician reports the sequence of conditions leading to death, as well as any other contributing conditions. For each line, the physician should also report the time elapsed from the advent of the conditions described and the event of death. Each paper form is then digitized and the causes of death reported undergo the complex steps of coding according to the provisions of the International Classification of Diseases, ICD-10 (WHO, 2019). The coding includes three main phases. In the first, an ICD-10 code is attributed to each reported cause; in the second, codes assigned in the first phase can be modified on the basis of other information reported on the certificate (multiple causes); in the third, the underlying cause of death — defined as the condition that initiated the chain of events leading directly to death — is selected in accordance to rules described in the ICD-10. Both multiple and underlying cause are released as official cause-of-death statistics.

The coding process and the selection of the underlying cause are carried out using Iris, a software tool adopted in many European and non-European countries for both automated and manual coding of causes of death (www.iris-institute.org). A first module of Iris assigns the corresponding ICD-10 code to each medical expression reported on the certificate. To this end, Iris performs text recognition based on a dictionary of medical terms developed by Istat experts. Regular expressions built into Iris are also used to standardize the text reported on the certificate, thereby improving matching performance. The second module is a rule-based system that applies ICD-10 rules to determine multiple causes and the underlying cause (Istat, 2023). Each year, Istat records approximately 670,000 death certificates. Of these, about 80% are fully automatically coded by Iris. The remaining certificates, around 135,000 per year, require manual coding, involving a substantial investment of highly trained personnel.

Most automatic coding failures are due to the inability of Iris to recognize medical terminology (the first module), caused by incompleteness of the dictionary, incorrect reporting by certifying physicians, or inadequate recording of the certificates.

2. Related Works

Deep learning models have achieved substantial results in text classification (Bruno M. et al., 2024). Specifically, in the health-care domain, the application of AI for the automatic coding of clinical texts, according to standardized classification systems, such as ICD-10, has been widely studied in medical informatics. As noted in a systematic review (Kaur et al. 2021), the methods have evolved progressively from traditional machine learning based on sparse text representations to deep learning and transformer-based architectures, which are designed to handle semantic complexity and large label spaces more effectively. Early work on clinical text classification and ICD coding primarily relied on methods such as Support Vector Machines, Naïve Bayes, and Random Forest, which are based on bag-of-words or TF-IDF techniques. In this context, Random Forest is often used as a robust and interpretable baseline due to its

effectiveness in high-dimensional feature spaces. However, these models rely heavily on high-level lexical features, which limits their ability to capture semantic similarity, contextual meaning, and variation in medical terminology. Kaur et al. highlight that, across the literature surveyed up to 2020, traditional machine learning methods generally perform worse than more recent context-based models such as artificial neural networks. A first substantial improvement was achieved using deep learning models such as convolutional and attention-based neural networks (Mullenbach, J. et al., 2018). These models replaced sparse features with dense representations learned directly from text. While these architectures improved the modelling of local patterns in clinical narratives and generally outperformed traditional baselines, their gains were still limited by the challenges of processing long documents, rare labels and broader contextual dependencies.

More recently, transformer-based models, especially BERT (Devlin et al., 2019), have become the leading approach for automated ICD coding. BERT-XLM (Zhang et al., 2020) is a notable example, adapting BERT to extreme multi-label classification and outperforming earlier neural baselines on large-scale ICD coding tasks. Pascual et al. (2021) further confirmed the potential of BERT-based systems, while also highlighting that constraints on input length remain a significant limitation for long clinical notes. Overall, the literature clearly shows a progression from traditional methods to deep neural models, and finally to BERT-based approaches. These approaches generally achieve the strongest results, often with micro-F1 scores above 0.6 in large-scale settings. However, direct comparisons remain difficult because of differences in datasets, label sets, and evaluation protocols. This evolution is particularly relevant for cause-of-death coding, where models must process short and heterogeneous descriptions while complying with official coding rules. In this context, BERT-based approaches appear better suited than Random Forest and other traditional classifiers for capturing the semantic content and improving ICD-10 assignment. Nevertheless, explainability, rare-label performance and integration into real production workflows remain challenges, especially in institutional contexts such as official statistics.

There is also evidence that AI can support cause-of-death coding with ICD-10 even in production settings. Névél et al. (2018) reported the results of a shared task in which fourteen research teams developed systems for the automatic ICD-10 coding of death certificates in French, Hungarian and Italian, achieving F-measures above 0.83 across all three languages. Zambetta et al. (2024) described the French experience of integrating AI in cause-of-death data production combining three methods: traditional rule-based automated coding with Iris, transformer-based ICD-10 coding trained on previously coded data, and targeted manual coding. Their results suggest that these approaches are complementary, although manual review remains necessary for some ICD categories. In summary, most automatic coding failures are concentrated in the text recognition phase, whereas the subsequent rule-based coding phases achieve consistently high performance. Evidence from other national contexts therefore suggests that AI methods are most effective when integrated with traditional coding systems and supported by targeted manual supervision.

Motivated by recent advances in automated coding, this work extends the methodological landscape of ICD-10 cause-of-death coding. It systematically compares approaches ranging from conventional machine learning to BERT-based models and Large Language Models (LLMs) on Italian medical records. The analysis provides new evidence on their potential to improve medical text recognition within the Iris-based workflow, reduce manual intervention, and preserve the comparability of official mortality statistics.

3. Methodology

In this section, we describe the used datasets, the data preparation process, and the models used to perform the classification of medical texts according to the ICD-10.

3.1 Dataset

For model development, datasets were prepared composed by the whole lines of the death certificates. Each line can contain a single or several medical expressions separated by semicolons. A single expression may correspond to one or more ICD-10 codes. In more complex cases, conjunctions, prepositions, or other separating phrases within a single expression lead to multiple diagnostic entities, each requiring a separate code. The elapsed time between the emergence of the conditions and the subsequent occurrence of death event, as documented in the death certificates lines, has not been considered in this study. From the full set of lines referring to deaths occurring in 2022 and 2023 (total available lines: 4,327,927), completely coded either automatically by Iris or manually by expert coders, 300,000 were randomly extracted. Each line included the medical text and the corresponding ICD-10 code (or codes). Furthermore, additional 950 death certificate lines, partially or completely unrecognized by Iris, have been extracted. The coding of these lines has been validated by expert coders, and the resulting dataset has been used as “gold standard”. This set includes only lines not fully coded by Iris, as the study specifically aims to apply AI methods to the text that Iris fails to recognize, and is used to evaluate the models’ ability to generalize to previously unseen data.

3.2 Models

To the aim of classifying Italian medical texts according to the ICD-10, a heterogeneous set of models was evaluated to compare approaches with different levels of representational complexity. The experimental setup started including Random Forest as a traditional machine-learning baseline. Subsequently, we adopted a BiLSTM (Schuster and Paliwal, 1997) implemented in PyTorch (Mikolov et al., 2013) with Word2Vec embeddings (Paszke et al., 2019). Furthermore, we made use of several transformer-based architectures. In particular, XLM-RoBERTa Base and XLM-RoBERTa Large were identified as multilingual encoder-only models (Conneau et al., 2020); while BioBIT and MedBIT-r3-plus (Buonocore et al., 2023; Crema et al., 2023) were included to assess the contribution of their Italian biomedical internal knowledge. Finally, LLama 3.2 1B (Meta, 2024) was considered to explore the potential of a compact large language model in order to perform the ICD-10 classification. Overall, this set of models enabled a comparison of conventional ensemble learning, recurrent neural networks, multilingual transformers, Italian biomedical transformers and lightweight LLMs within a unified classification framework.

Before the training, we fine-tuned the experimented models for seeking the best hyperparameters. With regard to Random Forest model, the number of estimators was set to two, the Gini index was used as the splitting criterion, and the minimum number of samples required to split an internal node was set to two. For deep learning models, training was performed for 70 epochs and the previously optimized hyperameters, such as the learning rate and batch size, were set, as reported in Table 1.

Table 1 – *Deep Learning models and LLM model hyperparameters*

Model Type	Learning	
	Rate	Batch size
XLM-Roberta - Large	$2 \cdot 10^{-5}$	512
LLama Base 1B	$2 \cdot 10^{-5}$	128
XLM-Roberta - Base	$2 \cdot 10^{-5}$	512
Bert Bio ITA	$2 \cdot 10^{-5}$	512
MedBIT-r3-Plus	$2 \cdot 10^{-5}$	512
Pytorch - BiLSTM - W2V	10^{-3}	256

In some cases, we performed re-training of the language model prior to training the actual classification model. The purpose of this procedure was to enrich the existing embedding space with new medical terms present in the dataset and in the gold standard. This was done only when we made use of Word2Vec for Deep Learning models based on BiLSTM classifiers and XLM-RoBERTa (Base and Large). Regarding the BiLSTM with Word2Vec embedding, we choose an embedding vector of size 300, an embedding window of size 8, skip-gram (Mikolov et al., 2013) as embedding algorithm and 30 epochs for the training of the embedding. For XLM-RoBERTa the embedding retraining hyperparameters that we chose are: 1 epoch, learning rate of $5 \cdot 10^{-5}$ and a batch size of 32.

The quality of the classification performed by all the models was measured harnessing the weighted F1-score (Christen P. et al., 2023). This performance metric was chosen as the primary evaluation metric since the ICD-10 classification is significantly imbalanced. Indeed, if compared to accuracy, F1-score is more resilient to situations with a strong unbalance of classes in the dataset. Compared to macro F1-score, the weighted F1-score reflects the real-world distribution of labels rather than treating extremely rare and frequent codes in the same way. For these reasons, it was considered the most appropriate metric for assessing performance in a production-oriented setting.

3.3 Data Preparation

In this work, we designed a rather complex and structured preprocessing pipeline aimed at transforming raw textual descriptions of causes of death into a standardized format suitable for serving as both training and test sets for the models we compared during the model selection phase. As a first step, data integrity checks were performed to identify and remove incomplete or inconsistent records. In particular, a missing values management procedure was implemented by removing entries containing missing values from both the textual field and the associated label (ICD-10 codes). These entries were then excluded from subsequent processing stages, where models frequently encounter failure when dealing with missing values, especially in the labels.

A preliminary Exploratory Data Analysis was also conducted, revealing that 3.99% of labels in the dataset contained more codes than the number of textual diagnostic entities separated by semicolon, which is conventionally used as the separator for both codes and texts. Optionally, “/” is also used as a label separator, but this is less common. Additionally, in 0.80% of cases, there were more diagnostic entities than codes. The remaining 95.21% of the data exhibits a

one-to-one correspondence between the number of medical texts and the number of codes in the labels. These calculations were performed on a sample of 300,000 records of the dataset of 4 million lines, as previously anticipated in section 3.1. However, since this sample was obtained via simple random sampling, we believe these percentages are likely representative of the entire dataset.

After data cleaning and preliminary analysis, all multiple-text entries of the form “text1;text2” and “code1;code2” were split, where possible, into multiple rows, resulting in entries such as “text1” with “code1” and “text2” with “code2” on a subsequent row. These cases, where the number of texts equals the number of codes, were estimated at 34,274 out of 300k records. This procedure led to an initial reduction in the number of distinct classes from 54,200 to 26,446, since instead of treating combined codes like “code1; code2” as distinct classes, they were treated as atomic classes “code1” and “code2”, which may repeat. On the other hand, the number of text entries increased from 300,000 to 343,973. The same splitting procedure was applied to the gold standard dataset, increasing it from 950 to 1,154 entries. The only cases in which the texts were not split were those where the number of codes exceeded the number of medical text segments in the entry. This occurred under two conditions: either because the text explicitly reported fewer diagnostic entities, possibly due to an error or omission, or because multiple diagnostic entities were linked within the same text segment by conjunctions such as “with” or “in”. We did not implement semantic splitting procedures for conjunction-based cases, an idea we reserve for future work. Conversely, cases with more pathologies than codes were discarded, although rare, due to ambiguity in code assignment. Furthermore, all medical texts whose class appeared only once in the entire dataset were removed. This decision was made because we performed stratified train-test splitting to preserve the same data distribution. A class with only one occurrence would prevent this and create severe imbalance, which would be detrimental to model training. This cut-off procedure reduced the number of classes in the training set from 16,714 to 9,732. Subsequently, a target cleaning procedure removed suffixes such as “years”, “months”, or “days” in expressions like “code1 (n years)” or “code1 (n months)”, normalizing them to “code1”. This resulted in a further reduction in the number of classes from 9,732 to 2,934 in the 300k dataset, and to 716 classes in the gold standard. The main dataset was then split into training and test sets using stratification, with an 80%–20% ratio. The creation of an internal test set is crucial for evaluating model performance both during training (e.g., for regularization purposes) and afterward, particularly for comparing models during model selection. After the split, the training set consisted of 272,564 data points and the test set of 65,452.

To re-train the language model for both Word2Vec with BiLSTM and XLM-RoBERTa, it was necessary to merge all texts from the datasets into a single large corpus during the preprocessing stage. This was due to the fact that both embeddings are generally trained in a fully unsupervised manner on textual corpora. For transformer-based models (e.g., LLaMA and BERT), the data were tokenized using the pre-trained tokenizers of the respective models and padded using dynamic padding.

4. Results

Table 2 reports, for each model, the F1 score obtained on the training set, the test set (meant as the 20% of the whole labelled dataset), and the gold standard, together with the training time.

As expected, all the models reach a very high F1 score when recoding the training set with values ranging from 0.967 for Pytorch - BiLSTM - No pretrained to 0.995 for XLM-Roberta – Large. The F1 score is slightly lower on the labelled test set, but still high with values ranging from 0.942 for Random Forest to 0.977 for XLM-Roberta - Large - embeddings retraining. A sharp drop in F1 values is observed when testing the gold standard: the highest score is reached by XLM-Roberta - Large with embeddings retraining with 0.535 which plummets to 0.357 in case of Pytorch - BiLSTM - W2V- No pretrained. Concerning the training time, faster training is associated with overall low performance: Pytorch and Random Forest with a training time of less than 13 minutes (only 4 minutes in case of Random Forest) show also the overall lowest performance. On the other hand, XLM-Roberta, especially the variety with embedding retraining, needing 4 hours of training, reaches the best overall performance both on the labelled test set and on the gold standard.

Table 2 – F1 score on different test sets and training time for different models tested.

Model Type	F1 SCORE			
	Training set	Test set	Gold Standard	Training Time
XLM-Roberta – Large - embeddings retraining	0.994	0.977	0.535	4h
XLM-Roberta - Large-LLama Base 1B	0.995	0.976	0.519	3h e 49 min
	0.985	0.953	0.513	50 h
XLM-Roberta - Base	0.991	0.973	0.502	1h e 36 min
Bert Bio ITA	0.990	0.972	0.493	1h e 44 min
MedBIT-r3-Plus	0.990	0.971	0.487	1h e 42 min
Random Forest	0.974	0.942	0.418	3.69 min
Pytorch - BiLSTM - W2V 300 - No pretrained - FULL Corpus	0.970	0.948	0.370	13 min
Pytorch - BiLSTM - W2V - No pretrained	0.967	0.946	0.357	13 min

Focusing on the results of the best-performing model on the gold standard (XLM-RoBERTa Large, with embeddings retraining), we compared the codes attributed by the model with those assigned by expert coders to each string reported in the rows. Rows can contain more than one string separated by a semicolon, yielding a total of 1,154 strings derived from 950 rows. The predicted code exactly matches the gold standard in 55.7% of cases (642 strings). Examining the probability of code attribution, 19.8% of strings (229) were coded with a probability of 100%; among these, for 218 the predicted code is the same as the gold standard, meaning that the agreement rate is 95.2%. Moreover, 34.2% of strings (395) were coded with a probability $\geq 95\%$ and $< 100\%$; among these, the agreement rate between the predicted and the gold standard codes is 74.7%, compared to only 24.3% when the probability falls below 95%. A further factor influencing agreement is the complexity of the strings. For strings requiring more

than one code, the agreement rate drops to 8.7%, as the model tends to predict only a single code in such cases.

5. Discussion

The best-performing model for the automatic coding of texts, found on Italian cause-of-death certificates, is XLM-RoBERTa large with embeddings retraining. This model yielded an F1 score of 0.535 and an agreement rate of 55.7% with the gold standard. The model performs much better on the test set, which represents 20% of the full dataset, than on the gold standard set. This difference is expected because the gold standard contains rows not fully coded by Iris. These rows are usually harder to classify, which makes the gold standard a more challenging test set. The confidence probability of code attribution can be used to identify cases where AI coding is reliable. However, a portion of certificates would still require manual revision, in line with studies recommending the integration of such methods with human supervision (Zambetta et al. 2024; Név     et al. 2018). Analysis of disagreement cases shows that discrepancies stem mostly on complex medical texts requiring several codes. For these texts the model predicts fewer codes than the gold standard, with agreement dropping sharply to 8.7%. Results should also be interpreted in light of the known error rates in manual coding, particularly for causes such as infectious diseases and traffic accidents (Harteloh et al., 2010; Grippo et al., 2015).

A key limitation is that the model was applied to entire rows, including those partially coded by Iris. An alternative approach could be to provide the model with individual diagnostic entities rather than certificate lines, this would likely improve performance. However, it is necessary to evaluate methods that allow for proper separation of diagnostic entities. As previously highlighted, in many cases entities are separated by semicolons, but in others by words or expressions that are not always easily identifiable automatically.

6. Conclusions

The study's findings are relevant both for the national and the broader international community, as integrating such methods with existing automated coding systems like Iris could yield further performance gains. Contrary to expectations, generative LLMs such as LLaMA do not outperform encoder-only transformer models such as XLM-RoBERTa, which appear particularly well suited to automatic coding tasks. Moreover, generative LLM models, containing billions of parameters, are computationally expensive to train, as reflected in the results reported in Table 2. Future work could therefore explore hybrid approaches in which such models are used selectively, for example, to support the correction of the most challenging coding cases, without requiring full model retraining.

References

- Bruno, M., Catanese, E., Ortame, F., & Pugliese, F. (2024, June). Measuring social mood on economy during covid times: A BiLSTM neural network approach. In *International Conference on Learning and Intelligent Optimization* (pp. 305-317). Cham: Springer Nature Switzerland.
- Zambetta E., Razakamanana N., Robert A., Clanché F., Rivera C., Martin D., Hebbache Z., Flicoteaux R. and Coudin E. (2024). Combining deep neural networks, a rule-based expert system and targeted manual coding for ICD-10 coding causes of death of French death certificates from 2018 to 2019. *International Journal of Medical Informatics*, 188, 105462. DOI: 10.1016/j.ijmedinf.2024.105462
- Kaur, R. and Ginige, A. and Obst, O. (2023). AI-based ICD Coding and Classification Approaches Using Discharge Summaries: A Systematic Literature Review. In *Expert Systems with Applications*. Elsevier.
- Mullenbach, J., Wiegrefe, S., Duke, J., Sun, J., & Eisenstein, J. (2018, June). Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)* (pp. 1101-1111).
- Devlin J., Chang M.W., Lee K., and Toutanova K. 2019. BERT: Pre-training Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zhang, Z., Liu, J., & Razavian, N. (2020, November). BERT-XML: Large scale automated ICD coding using BERT pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop* (pp. 24-34).
- Pascual, D., Luck, S., & Wattenhofer, R. (2021, June). Towards BERT-based automatic ICD coding: Limitations and opportunities. In *Proceedings of the 20th workshop on biomedical language processing* (pp. 54-63).
- World Health Organization. (2019). International classification of diseases, 10th revision (ICD-10), version 2019. <https://icd.who.int/browse10/2019/en>
- Névél A., Robert A., Grippo F., Morgand C., Orsi C., Pelikan L., Ramadier L., Rey G. and Zweigenbaum P. (2018). CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. *Conference and Labs of the Evaluation Forum (CLEF)*.
- Istat. (2023). Codifica delle cause di morte con l'ICD-10, versione 2020. Istat, *Lecture Statistiche – Metodi*. <https://www.istat.it/produzione-editoriale/codifica-delle-cause-di-morte-con-licd-10-versione-2020/>
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*
- Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32
- Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, 8440–8451.
- Buonocore, T. M., et al. (2023). Localizing in-domain adaptation of transformer-based language models for the Italian biomedical domain. *Journal of Biomedical Informatics*, 142, 104398.

- Crema, C., et al. (2023). Advancing Italian biomedical information extraction with large language models: Methodological insights and multicenter practical application. *Journal of Biomedical Informatics*, 145, 104473.
- Meta. (2024). Llama 3.2 model card.
- Christen P., Hand D.J., and Kirielle N. 2023. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Comput. Surv.* 56, 3, Article 73 (March 2024), 24 pages. <https://doi.org/10.1145/36063>
- Harteloh P., de Bruin K. and Kardaun J. (2010). The reliability of cause-of-death coding in The Netherlands. *European Journal of Epidemiology*, 25, 531–538. DOI: 10.1007/s10654-010-9445-5
- Grippe F., Grande E., Simeoni S., Pennazza S., Cinque S., Bracci T. and Frova L. (2015). Reliability of causes-of-death statistics: the Italian experience from the ICD-10 training course. *Rivista di Statistica Ufficiale*, 17(3), 103–119

A Semantic Search Approach to Occupation Coding in Official Statistics

Mauro Bruno¹, Giulio Massacci²

¹Istituto Nazionale di Statistica – mauro.bruno@istat.it

²Istituto Nazionale di Statistica – giulio.massacci@istat.it

Abstract

Occupation coding is a relevant task in Official Statistics, especially when free-text job descriptions must be translated into formal occupational classifications. In the Italian context, this challenge emerges clearly in surveys where respondents describe their work activities in natural language, while statistical production requires coding into the official taxonomy (CP2021). This paper investigates whether semantic search can support this process by improving the alignment between user language and classification language. More specifically, the study formulates occupation coding as a retrieval task over the CP2021 label space and evaluates different multilingual embedding models on a benchmark of Italian occupation descriptions paired with gold-standard codes.

The paper also examines a central methodological question: how the textual representation of occupational categories affects retrieval quality. The results show that semantic retrieval is a promising approach for occupation coding and, above all, that richer textual representations of occupational categories substantially improve performance across systems. These findings suggest that the design of occupational category representations is a key component in developing more effective semantic tools for coding support in Official Statistics.

Keywords: Artificial Intelligence; Official Statistics; Occupation Classification; Semantic Search; Automatic Coding

1. Introduction

In recent years, National Statistical Offices (NSOs) and International Organizations have increasingly invested in Machine Learning and Artificial Intelligence (AI) to support the modernization of statistical production processes (Bruno et al., 2025; HLG-MOS, 2023; UNECE, 2022). These approaches are particularly relevant for tasks involving unstructured textual data, where flexible methods are needed to bridge the gap between user-generated descriptions and formal classification systems.

In this context, occupation coding represents a relevant task in Official Statistics, as it requires translating free-text job descriptions into formal occupational classifications. In Italy, this function is performed by CP2021, the national classification of professions, which represents the labour market through a hierarchical organization of occupational categories with increasing levels of specificity (Istat, 2021). As the Italian version of ISCO, CP2021 also ensures coherence with the broader international framework for occupation statistics.

At the same time, translating real-world descriptions of work activities into this formal classification remains a challenging task, as such descriptions are often heterogeneous, ambiguous, and expressed in natural language rather than in the institutional terminology of the taxonomy. This issue is particularly relevant in survey questionnaires that include open-ended questions, whose responses must be coded into complex statistical classifications, such as those used for

occupations or economic activities. The Labour Force Survey provides a clear example: respondents describe their main work activity in natural language, and these descriptions must then be translated into the official profession classification. Such descriptions are often short, uneven in specificity, and expressed in everyday language rather than in the formal wording of CP2021. The central bottleneck is therefore semantic rather than merely lexical: the system must recover the intended occupational meaning even when overlap with the official class labels is weak or absent.

Istat already supports this task through the *Navigatore delle professioni*, a tool that combines hierarchical exploration of the classification with textual search over occupational entries¹. Its methodology is essentially lexical and rule-based, as retrieval depends on matching normalized user queries against a controlled vocabulary built from CP2021 labels and the associated *voci professionali*, that is, example occupation titles linked to each 5-digit category. These examples enrich the search vocabulary with expressions that are closer to user language, thereby improving the effectiveness of query matching. This design highlights the importance of enriching the textual representation of occupational categories, a principle that also underlies more recent semantic approaches. The *Navigatore delle professioni* therefore represents an important institutional baseline, but it also highlights a structural limitation of lexical matching approaches: user queries and classification language do not always coincide.

More generally, the description of an individual's work may refer to several closely related dimensions, such as employment status, position in the profession, and economic activity. For this reason, these dimensions must be considered explicitly when designing a coding methodology. Moreover, user-provided descriptions do not always contain sufficient information to unambiguously identify a specific occupational category, as they may capture only partial aspects of the underlying occupational profile. In this context, occupation coding depends on two complementary elements: the quality of the textual representation of occupational categories and the information content of user queries. The present paper formulates occupation coding as a semantic retrieval problem over a structured label space. More specifically, it evaluates whether the alignment between natural-language occupation descriptions and CP2021 classes can be improved by comparing different textual representations of 5-digit occupational codes.

2. Data Sources

The empirical setting of this study is based on two complementary data sources. The first is the CP2021 occupational classification and an auxiliary datasets used to enrich the official class descriptions with 6-digit occupational entries. The second is a labelled dataset of Italian free-text occupation descriptions paired with gold-standard CP2021 codes. In the paper, the classification datasets define the target occupational categories and their official textual descriptions, while the labelled dataset is used to evaluate retrieval performance and to enrich the textual representation of the 5-digit codes.

2.1. CP2021 classification and 6-digit enrichment resource

CP2021 is the occupational classification used in this study as the target coding framework. The classification is organized hierarchically and includes titles and descriptions at multiple levels of granularity. In this study, occupation coding is performed at the 5-digit level, since the labelled dataset contains 5-digit CP2021 codes and statistical production requires occupation data at this level of detail. Table 1 summarizes the hierarchical structure of CP2021 in the source dataset.

¹ *Navigatore*, available at: <https://professioni.istat.it/sistemainformativoprofessionioni/cp/>
JADT 2026: 18th International Conference on Statistical Analysis of Textual Data

Level	Number of codes	Description
1-digit	9	Highest aggregation level, consisting of 9 major professional groups.
2-digit	40	Composed of 40 professional groups.
3-digit	130	Composed of 130 professional classes.
4-digit	510	Composed of 510 professional categories.
5-digit	813	Finest level of detail, consisting of 813 professional units.

Table 1: Hierarchical structure of the CP2021 classification

To enrich the official classification, the study also uses an auxiliary data source containing a more detailed occupational inventory linked to CP2021. The resource includes 6,813 complete rows associated with 6-digit occupational entries and is used to augment each 5-digit class with a set of more concrete occupational denominations that instantiate the broader class.

This enrichment resource is methodologically important because it provides lexical and semantic detail that is not always available in the official 5-digit description. While the official classification captures the institutional definition of the occupational class, the 6-digit entries offer a finer-grained representation of the occupations that fall under that class. In the retrieval setting adopted here, these 6-digit labels act as semantically informative examples that can improve the alignment between user queries and the textual representations of occupational codes.

2.2. Labelled occupation dataset

The second core data component is a labelled dataset of Italian occupation descriptions paired with gold-standard CP2021 codes assigned through prior annotation. The labelled dataset contains 61,881 records and serves as the benchmark of the study. Each record is interpreted as a realistic occupational description that the model must map to the appropriate CP2021 class. Each annotated occupation description is converted into a query, and the gold-standard CP2021 code is used to determine whether the correct occupational category is retrieved within the top k or not retrieved at all.

This use of the dataset provides a direct evaluation setting: the benchmark measures how effectively multilingual embedding models can retrieve the correct CP2021 occupational category from naturally occurring free-text descriptions.

3. Methodology

The objective of this study is the automatic assignment of CP2021 occupational codes to Italian free-text job descriptions. Rather than formulating the task as a standard multiclass classification problem, we cast it as a semantic retrieval problem. This choice is motivated by the structure of the task itself: both the input and the label space are textual, and the quality of the assignment depends on the semantic correspondence between a user-generated occupation description and a textual representation of the target occupational category, hereafter referred to as a *descriptor*.

Let $q \in \mathcal{Q}$ denote a query describing an occupation, and let $D = \{d_1, d_2, \dots, d_N\}$ be the set of candidate class descriptors. Each descriptor d_j is a natural-language representation of one CP2021 occupational class. A sentence-embedding model defines a mapping

$$f : \mathcal{T} \rightarrow \mathbb{R}^m,$$

where $\mathcal{T}$ is the space of input texts and $\mathbb{R}^m$ is a dense semantic vector space, as commonly done in recent semantic representation approaches (Gao et al., 2021; Ni et al., 2021; Reimers and Gurevych, 2019). Query texts and class descriptors are projected into the same space:

$$\mathbf{q} = f(q), \quad \mathbf{d}_j = f(d_j).$$

Relevance is estimated through a similarity function $s(\mathbf{q}, \mathbf{d}_j)$, implemented here using cosine similarity. For each query, the system ranks all candidate descriptors in decreasing order of similarity and returns the top- k results:

$$\pi_k(q) = \text{TopK}_{d_j \in D} s(\mathbf{q}, \mathbf{d}_j).$$

Under this formulation, occupational coding becomes a ranking problem over a semantically structured label space. The correct code is not predicted through a classification step; instead, it is recovered if the corresponding descriptor is ranked sufficiently high.

This formulation is especially suitable for occupational coding for two reasons. First, it preserves the textual nature of the classification and makes the representation of classes analytically explicit. Second, it naturally supports top- k evaluation and human-in-the-loop coding workflows, where a small set of highly ranked candidates is often more useful than a single forced prediction.

3.1. Sentence embedding and semantic search

The central methodological assumption of the study is that semantically similar occupation descriptions and occupational categories should be mapped to nearby points in the embedding space. In this framework, the effectiveness of the coding methodology depends on the quality of both the embedding function f and the descriptor text d_j .

This second point is central to the study. In OS classifications, official category descriptions are often concise, formal, and lexically distant from the language used in survey responses or user-generated text. As a consequence, even strong multilingual embedding models may fail to align queries and classes if the descriptor representation is too narrow or too formalized.

The study therefore treats descriptor construction as a substantive methodological variable rather than a purely technical preprocessing step. The main hypothesis is that retrieval quality can be improved not only by selecting a stronger embedding model, but also by designing richer descriptors that reduce the semantic mismatch between user language and classification language.

3.2. Descriptor Design

The study evaluates two descriptor configurations, both defined at the 5-digit level.

- **Official Descriptor:** The first configuration, denoted *official*, uses only the official 5-digit class representation. Each descriptor combines the official occupational title and the official class description. This condition acts as the conceptual baseline, as it relies exclusively on the wording provided by the classification scheme.
- **Enriched Descriptor:** The second configuration, denoted *enriched*, augments each 5-digit descriptor with occupational examples derived from the associated 6-digit entries. The rationale is that 6-digit labels provide a richer and more concrete lexical surface than the official 5-digit description alone. While the 5-digit description captures the conceptual definition of the class, the 6-digit entries expose the occupational vocabulary that typically instantiates that class in practice.

3.3. Benchmark Design

The evaluation compares a set of state-of-the-art multilingual sentence-embedding models under a consistent retrieval protocol, including architectures from recent open-source families. Detailed model-specific results are presented in Section 4.. For each combination of embedding model and descriptor configuration, the full set of class descriptors is embedded once to create the retrieval index, and each query is then encoded and matched against the indexed descriptors. Retrieval is truncated at the top-100 candidates.

The present methodology should be understood as a retrieval benchmark rather than a complete production evaluation. Its purpose is to measure how embedding-based semantic search behaves under controlled variations in descriptor design and model choice. It does not yet include significance testing, subgroup-level error analysis, or qualitative inspection of failure cases. These elements would be valuable extensions, as well as the integration of uncertainty quantification techniques, such as conformal prediction, to provide statistically valid confidence measures (Angelopoulos and Bates, 2021; Ortame and Isidori, 2026; Vovk et al., 2005). However, the current study is primarily concerned with establishing whether descriptor enrichment produces a clear and robust gain in retrieval-based occupational coding.

4. Results

Table 2 summarizes the benchmark results across all evaluated models and descriptor configurations. Taken together, the benchmark shows two main patterns. First, the strongest overall configuration is Qwen/Qwen3-Embedding-4B (Zhang et al., 2025) combined with the enriched descriptors. Second, the most stable improvement does not come from model choice alone, but from enriching the textual representation of the target occupational classes.

To examine the ranking behavior more closely, Figure 1 reports the Accuracy@k curves obtained under the enriched descriptor configuration. All models improve regularly as k increases, but their relative ordering is already visible at Top-1 and remains broadly stable across the full ranking profile. In this setting, Qwen/Qwen3-Embedding-4B is the strongest model overall, with Top-1 = 0.424, Top-3 = 0.647, and Top-5 = 0.722. BAAI/bge-m3 (Chen et al., 2024) and Qwen/Qwen3-Embedding-8B (Zhang et al., 2025) follow closely, but neither surpasses the leading configuration at any of the main cutoffs. This pattern indicates not only that the best-

Model	Descriptor	Top-1	Top-3	Top-5
BAAI/bge-m3	official	0.333	0.519	0.595
BAAI/bge-m3	enriched	0.415	0.600	0.672
intfloat/multilingual-e5-base	official	0.314	0.481	0.549
intfloat/multilingual-e5-base	enriched	0.393	0.585	0.657
intfloat/multilingual-e5-large	official	0.336	0.524	0.598
intfloat/multilingual-e5-large	enriched	0.389	0.585	0.660
Alibaba-NLP/gte-multilingual-base	official	0.310	0.495	0.576
Alibaba-NLP/gte-multilingual-base	enriched	0.350	0.541	0.623
Qwen/Qwen3-Embedding-0.6B	official	0.224	0.380	0.454
Qwen/Qwen3-Embedding-0.6B	enriched	0.252	0.425	0.509
Qwen/Qwen3-Embedding-4B	official	0.348	0.554	0.635
Qwen/Qwen3-Embedding-4B	enriched	0.424	0.647	0.722
Qwen/Qwen3-Embedding-8B	official	0.379	0.594	0.670
Qwen/Qwen3-Embedding-8B	enriched	0.402	0.625	0.703

Table 2: Performance comparison across models and descriptor configurations.

performing models retrieve the correct occupational code more often, but also that they place it more consistently in the top portion of the ranked list.

Figure 2 focuses on the effect of descriptor design. The pattern is clear: moving from official to enriched improves Top-1 accuracy for all models. The largest gain is observed for BAAI/bge-m3 (+8.2 percentage points), which allows it to overtake Qwen/Qwen3-Embedding-8B under the enriched setting. It is followed by intfloat/multilingual-e5-base (Wang et al., 2024) (+7.9 points), which surpasses its own larger version, intfloat/multilingual-e5-large, suggesting that the additional contextual information provided by descriptor enrichment can matter more than model scale alone. Qwen/Qwen3-Embedding-4B also records a substantial gain (+7.6 points), preserving its leading position in the ranking. The main result is therefore not tied to a single architecture. Rather, it emerges from a more informative verbalization of the target occupational classes, which improves the semantic alignment between real occupation descriptions and the CP2021 descriptor space.

Future work will further investigate the role of descriptor enrichment and its impact on retrieval performance. In particular, this includes exploring alternative strategies for enhancing descriptor representations, such as incorporating additional lexical evidence derived from annotated data or leveraging synthetic data to better capture the variability of user language.

4.1. Performance–complexity trade-off

The benchmark results can also be interpreted in terms of the trade-off between retrieval performance and model complexity. While larger embedding models generally achieve higher accuracy, they also entail greater computational requirements, which may limit their applicability in practice.

From an analytical perspective, this trade-off highlights the importance of jointly considering effectiveness and computational efficiency when evaluating semantic retrieval approaches. In this study, the number of model parameters is used as a practical proxy for model complexity.

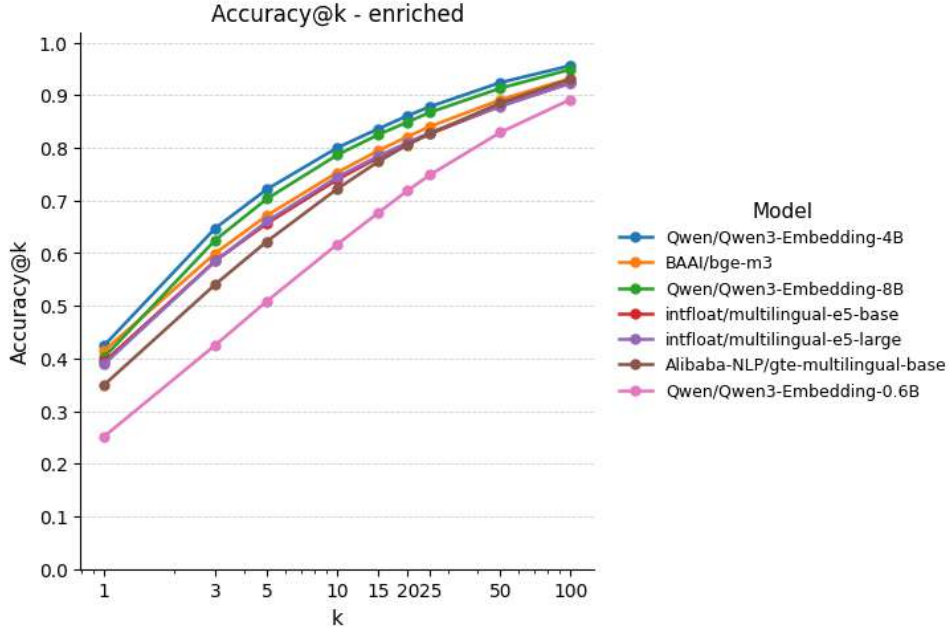


Figure 1: Accuracy@k for different multilingual embedding models using enriched descriptors. All models show a consistent increase in performance as k grows, with Qwen3-Embedding-4B achieving the highest accuracy across all values of k.

In particular, we define a compactness indicator as:

$$C = \log \left(\frac{N_{\max}}{N} \right),$$

where N denotes the number of parameters of a given model and $N_{\max}$ is the maximum number of parameters observed in the benchmark. Figure 3 illustrates this relationship by comparing Top-1 accuracy (under the enriched descriptor configuration) with this indicator. The two quadrant boundaries are defined by the median values of the two variables considered, namely normalized compactness on the horizontal axis and Top-1 accuracy on the vertical axis.

The figure shows that models such as intfloat/multilingual-e5-base achieve a favorable balance between performance and complexity, whereas Qwen/Qwen3-Embedding-4B and Qwen/Qwen3-Embedding-8B offer higher accuracy at the cost of increased computational burden. Conversely, Alibaba-NLP/gte-multilingual-base exhibits lower complexity but also comparatively lower performance.

These results suggest that model selection in semantic retrieval tasks should not rely on accuracy alone, but should instead account for the broader balance between effectiveness and computational cost. Future work will extend this analysis by considering a wider range of models and more refined measures of computational complexity, in order to provide a more comprehensive assessment of the performance–complexity trade-off.

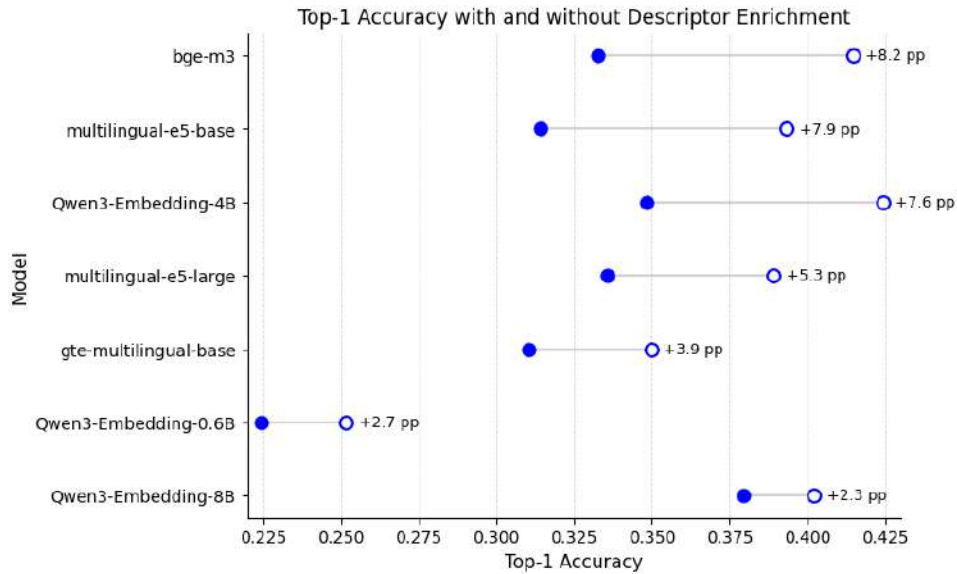


Figure 2: Top-1 accuracy under progressive descriptor enrichment (official, enriched). All models benefit from descriptor augmentation, with consistent improvements across stages. Gains are particularly pronounced for smaller models, suggesting that descriptor quality plays a crucial role in compensating for limited model capacity.

5. Conclusions

This paper examined occupation coding in OS as a semantic retrieval problem over the CP2021 label space. Starting from the institutional difficulty of mapping free-text occupation descriptions to official categories, the study compared multilingual embedding models under richer textual representations of the target classes. The benchmark shows that semantic retrieval is a viable approach for this task and that the strongest overall configuration is obtained with Qwen/Qwen3-Embedding-4B and the enriched descriptor setting.

More importantly, however, the main contribution of the paper does not lie in model comparison alone. The most robust result is that descriptor enrichment systematically improves retrieval quality, highlighting that the textual formulation of occupational classes is a central methodological component of retrieval-based coding rather than a secondary implementation detail. These findings should be interpreted in line with the scope of the study. The present paper is a benchmark-oriented methodological contribution, not yet a full production evaluation. It shows that enriching official descriptors with occupational examples can substantially improve the semantic alignment between user language and classification language.

At the same time, several extensions remain necessary before moving toward a broader assessment, including error analysis, subgroup-level evaluation, qualitative inspection of difficult cases, explicit performance–complexity comparisons across models, and comparison with stronger baselines or assisted coding workflows. In this sense, the present results provide a compact but solid basis for further research on semantic tools that can support occupation coding while remaining compatible with the methodological requirements of OS.

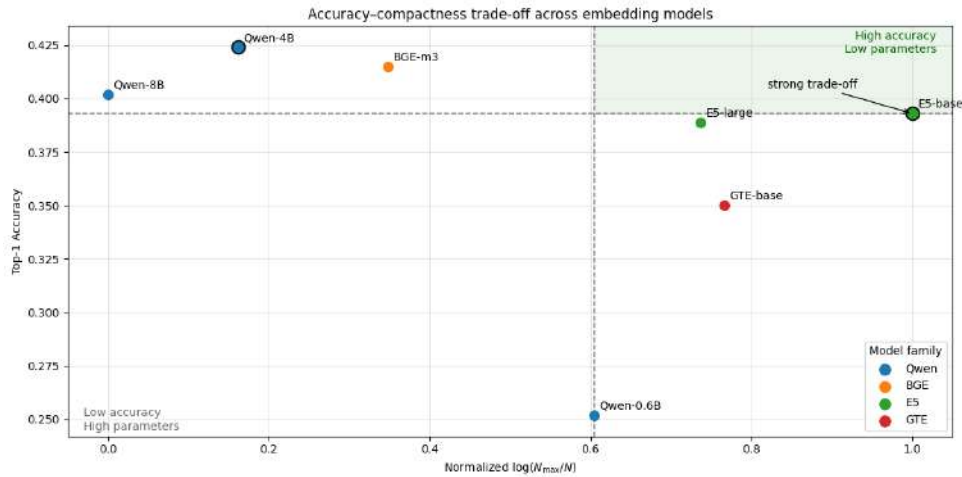


Figure 3: Accuracy-compactness trade-off across embedding models. Compactness reflects model size (number of parameters). Quadrants are defined by median values. The top-right region highlights models with above-median accuracy and compactness.

References

- Angelopoulos A. N. and Bates S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
- Bruno M., Catanese E., Cerasti E., De Cubellis M., De Fausti F., Di Zio M., Grippo G., Lancioni G., Massacci G., Mugnoli S., Ortame F., Pappagallo A., Pugliese F., Righi A., Sabbi A., Sisti F., Summa D., and Valentino L. (2025). Big data and machine learning at istat. In *Foundations and Advances of Machine Learning in Official Statistics*, pp. 239–298. Springer Nature Switzerland, Cham.
- Chen J., Xiao S., Zhang P., Luo K., Lian D., and Liu Z. (2024). M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Gao T., Yao X., and Chen D. (2021). Simcse: Simple contrastive learning of sentence embeddings. *arXiv*.
- HLG-MOS (2023). Large language models for official statistics. White paper, UNECE, Geneva.
- Istat (2021). Classificazione delle professioni 2021. PDF resource.
- Ni J. et al. (2021). Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv*.
- Ortame F. and Isidori F. (2026). Statistical frameworks for reliable machine learning predictions and inference. *Statistical Journal of the IAOS*, 42(1):203–210.
- Reimers N. and Gurevych I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992.
- UNECE (2022). Machine learning for official statistics. Technical report, UNECE, Geneva.
- Vovk V., Gammerman A., and Shafer G. (2005). *Algorithmic Learning in a Random World*.
- Wang L., Yang N., Huang X., Yang L., Majumder R., and Wei F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Zhang Y., Li M., Long D., Zhang X., Lin H., Yang B., Xie P., Yang A., Liu D., Lin J., et al. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Exploring User–Chatbot Communication: Methods, Insights, and Channel Differences¹

Samanta Pietropaoli¹, Salvatore Agrillo¹, Paola Bosso¹, Barbara D’Amen¹,
Gabriella Fazzi¹

¹Italian National Institute of Statistics (Istat) – samanta.pietropaoli@istat.it, salvatore.agrillo@istat.it,
paola.bosso@istat.it, barbara.damen@istat.it, gabriella.fazzi@istat.it

Abstract

The increasing adoption of conversational AI within National Statistical Institutes is transforming how users access and interact with official statistics services. This paper describes a text analytics study of the conversations generated by the chatbot developed by the Italian National Institute of Statistics (ISTAT) and integrated into its institutional deployment channels, including Contact Centre and Business Portal. The goal of the study is to investigate how communication patterns, information needs and interaction dynamics vary across service channels while exploiting chatbot conversations as a valuable source of unstructured data for official statistics analysis.

At this aim, a first thematic classification of the users' requests is carried out for grouping conversations by specific channel and topic, such as access problems, operational questions, and clarification requests. Secondly, a comparative linguistic analysis is conducted by considering lexical features, vocabulary richness, and message structure differences between user inputs and chatbot outputs, for avoiding bias introduced by the highly standardised and polite nature of the automated responses. Finally, the study adopted sentiment analysis, based on AI and natural language processing tools, applied both in its standard form and in a customised version. This analysis seeks to detect recurring patterns in how users address the chatbot and evaluate the tone of interactions.

The expected results aim to deepen the understanding of communicative dynamics between users and virtual assistants, offering methodological reflections on the use of artificial intelligence applied to textual data and insights for developing more effective conversational agents.

Keywords: Human–Chatbot Interaction, Sentiment Analysis, Topic Modelling

1. Introduction

The Single Point of Contact (PUC) is Istat's internal centralised service for assisting businesses and households involved in statistical surveys. The service operates with an omnichannel approach, enabling users to submit support requests through multiple channels, including telephone (public utility number 1510), standard email, certified email and web platform. The PUC supports respondents in all phases of the reporting process, from general information to data entry, system access and regulatory compliance. The service operates on a two-level model: an external first level handles approximately 90% of requests – the simplest and most recurring ones – while specialised Istat staff at the second level address the remaining 10%, which requires deeper knowledge of survey content and processes. This structure enables the

¹ This paper is the result of collaborative work among all the authors. Sections 1 and 7 were written by Paola Bosso; sections 2 and 5 were written by Gabriella Fazzi; sections 3 and 4 were written by Barbara D’Amen; sections 6 and 6.1 were written by Samanta Pietropaoli; section 6.2 was written by Salvatore Agrillo.

PUC to support 60-80 surveys per year and to manage between 250,000 and 300,000 requests annually.

Despite progressive optimisations (process automation, expansion of contact channels), critical issues remain due to the discontinuous and variable nature of request flows. Survey launch and closure phases, reminders to non-respondents, and deadline extensions generate hard-to-predict spikes in request volume, complicating service management.

To address this challenge, in 2025, Istat introduced a virtual agent (chatbot) as a front-end level upstream of the resolution process, thereby further expanding its omnichannel offering. The virtual agent handles demand peaks and automatically handles routine cases, relieving human operators for higher-value activities. In October 2025, the chatbot was integrated into the Istat web portal; in 2026, chat functionality was extended to the Business Portal, the data collection system for enterprises. Work is currently underway to integrate the virtual agent into the telephone channel. The chatbot currently operates in user-less mode: it retains no user information and cannot escalate a request to a human operator.

This paper focuses on one specific aspect of this experience: the textual analysis of chatbot conversations. Based on a corpus of 12,017 conversations collected from November 2025 to March 2026, the study examines request volumes, types, recurring patterns, and their implications for service efficiency and quality.

Informed by the existing literature, this study addresses the following research questions:

- What lexical and stylistic differences characterise user and chatbot language, and how do these vary across request types?
- What is the effectiveness of the virtual agent in resolving users' problems, and through which communicative strategies is this effectiveness achieved?

Through a set of text analysis techniques (lexical analysis, LIWC for Language Style Matching, sentiment analysis with multilingual BERT, topic modelling with BERTopic, and rule-based behavioural analysis), the study aims to understand human-machine communication dynamics in an institutional context, offer methodological reflections on using AI for textual data analysis, and provide insights for designing more effective conversational agents.

2. Literature review

A chatbot is a computer program designed to simulate and process human conversation. The primary goal is to allow users to interact with digital devices as if they were communicating with a real person, enabling the system to understand user intent, process requests, and provide relevant responses through text or voice.

Early linguistic studies focused on comparing conversations between humans and those between a human and a chatbot. This research consistently revealed that people tend to alter their language when interacting with a machine. For instance, early findings showed that "people tend to use a more restricted vocabulary when they believe they are interacting with a computer" (Park et al., 2018). As chatbot technology, powered by advances in Natural Language Processing (NLP) and Large Language Models (LLMs), has grown more sophisticated, the focus of linguistic analysis has expanded. Today, such studies are not only used to improve user experience but have also become critical in distinguishing between human- and AI-generated text, a task relevant to combating the spread of misinformation.

In recent years, chatbots have become increasingly widespread, with customer service emerging as one of their most prominent applications, generating a vast amount of conversational data and providing a rich resource for the analysis of human-chatbot interactions. This analysis has revealed consistent linguistic and emotional differences between the language used by humans and that generated by chatbots.

When interacting with chatbots, human language exhibits greater variability and informality. Humans use more personal references and express a wider range of negative emotions. Their messages can be shorter and less structured, frequently containing spelling and grammatical errors, which reflects a more spontaneous and less filtered communication style, similar to instant messaging (Park et al., 2018; Sandler et al., 2024; Rodrigues et al., 2026). Humans use more personal references and express a wider range of negative emotions, such as anxiety and sadness.

AI-generated language is consistently more formal, structured, and, as some studies note, "more positive and motivational" than human language (Sandler et al., 2024). Chatbots tend to use more positive and motivational language, avoiding sensitive or negative emotional expressions. Their linguistic style is often more analytical and authoritative, marked by a higher use of words related to cognitive processes (e.g., "cause," "insight") and certainty. Studies have shown that models like ChatGPT score higher than humans in categories such as social processes, attentional focus, and positive emotional tone, leading some researchers to describe their language as being "more human than human" in its politeness and attentiveness (Sandler et al., 2024). However, this can also manifest as unnatural interactions; for instance, some chatbots may repeatedly change the topic to avoid making a mistake, while others can appear overly argumentative (Park et al., 2018). This lack of natural variability is a key differentiator, as AI models tend to produce more uniform and predictable linguistic patterns compared to the rich and often messy complexity of human expression.

Recent research has leveraged sentiment analysis to evaluate the quality of human-chatbot interactions. This analysis provides quantitative metrics for user experience. By retrospectively analysing chat logs, studies in customer service have measured user satisfaction, identifying specific "breaking points" where user sentiment turns negative. In mental health and emotional support, this method is used to validate a chatbot's efficacy: in this field, specific models have been developed to track a user's emotional state, aiming to build "robust emotion-aware chatbots that can understand and respond to users' emotions" (Chen et al., 2021). Longitudinal studies also apply this approach to "assess the long-term effects of continued use of ChatGPT on the emotional well-being of its users," exploring constructs like loneliness and emotional dependency over time (Reviriego et al., 2024). In comparative linguistics, it reveals that AI models maintain a more consistently positive emotional tone, whereas humans exhibit a broader emotional spectrum.

3. Users' conversations with the Istat chatbot: main characteristics and types of requests

As previously stated, this study is based on 12,017 conversations between the Istat chatbot and users (respondents), collected from November 2025 to 31 March 2026. The number of conversations per month remains fairly stable, ranging between 5% and 6%, with the exception

of February, when there was an increase in interactions, reaching 49.7%, followed by March (33.8%).

Considering the distribution across days of the week, most of the conversations are concentrated in the early part of the week, in particular on Mondays (22.8%) and Tuesdays (23.9%), before gradually decreasing over the following days until the weekend (17%), when the lowest interaction rates are recorded on Saturdays (2.6%) and Sundays (2.3%). Interactions between users and the chatbot peak in the morning and late morning, accounting for 29.7% and 27.7%, respectively, and gradually decrease to below 3% from evening through early morning.

The analysis provides insights into the topic of users' requests. Most of the chatbot conversations are related to business surveys, accounting for 74.5% of all conversations, even in the Contact center channel, while individual and household surveys represent 8.5% of the conversations, and a small fraction is incompletely classified (2.5%) or unclassified (14.5%). Moreover, nearly half (47.6%) of the conversations are focused on issues encountered by respondents in accessing the portal or questionnaires, while 15.1% requested information or clarification about the survey. Other concerns were less frequent and related to information on regulations, privacy, or response obligations (1.3%), information on assessments or violation reports (0.4%), and interview booking (0.1%), while 35.5% of the conversation was not specified.

Regarding the number of messages per user, most users send few messages: 85.7% send three or fewer, and nearly 93.8% send four or fewer, while only 2.7% of users send six or more messages. On average, users send approximately 1.72 messages each, confirming that typical engagement is low.

4. Corpus structure and lexical characteristics

As shown in Table 1, the analysed conversations consist of 25,743 messages from the chatbot and 25,985 messages from users. These messages were collected into two corpora to compare their main characteristics and to identify differences or similarities in lexical usage. For this purpose, the corpora were imported into the TaLTaC2 software and pre-processed through parsing, normalisation, and lexicalisation, to reduce redundancy and ensure consistency among forms. Then, a lexical analysis was carried out to identify the most frequent words and lemmas of the verbs. As regards the main characteristics of the corpora, the type-token ratio (TTR) is below 20% (Bolasco, 1999, p.203) in both corpora, showing that these can be subjected to quantitative analysis.

Table 1: corpus characteristics

	Number of messages	Word Type (V)	Word Token (N)	Type Token Ratio (V/N)*100	Hapax Percentage (V1/V)*100
Agent	25,743	7,734	2,101,377	0.36	28.30
Users	25,985	9,000	206,686	4.35	56.57

According to the literature, the chatbot tends to use a smaller number of distinct words (V) (Reviriego et al., 2024) than the users, but the higher frequency of these words (N) indicates more repetitions and redundant expressions (Yildiz Durak et al., 2025). A notable difference is related to the percentage of hapax legomena, words that occur only once in a text (Baayen, 2001). As shown in Table 1, this is higher in users' language than in chatbot output, reflecting

the presence of typographical errors or misspellings and, in general, a more variable language. Regarding lexical characteristics, the analysis highlighted some differences between the language used by the chatbot and users. In particular, the chatbot’s lexicon is characterised by verbs such as *utilise* (14,195), *contact* (11,338), *click* (7,185), *persist* (3,886), *verify* (3,590), *provide* (2,055), *control* (1,758), and *confirm* (1,570), reflecting its role in supporting procedural execution and system navigation (Chen et al., 2021). Accordingly, the most frequent nouns used by the chatbot reflect its technical role (*assistance*, 24,168; *username*, 19,024; *browser*, 6,445; *cache*, 2,645), accompanied by nouns such as *business portal* (12,284), *portal manager* (3,172), *informative letter* (5,661), referring to the Istat organisation and its institutional procedure. As regards users, the analysed messages included nouns linked to contextual domains such as *questionnaire* (4,336), *password* (2,282), *user code* (1,107), *mail* (1,045), *questionnaire completion* (744), *login credentials* (303), *company* (277) and *link* (246) expressly aimed to explain issues for finding solutions. Consistently, verbs such as *succeed* (2,791), *access* (2,206), *complete* (1,946), *find* (1,662), *retrieve* (1,108), and *answer* (634) are characteristics of users’ requests.

5. Linguistic Patterns in Human-Chatbot Conversation

To analyse linguistic patterns in human–chatbot interactions, we employed the Linguistic Inquiry and Word Count (LIWC) framework, a widely used tool that categorises words into psychologically and linguistically meaningful classes, enabling quantitative analysis of communication style and underlying cognitive processes. Within this framework, we focused on Language Style Matching (LSM), a metric that measures linguistic coordination at the lexical level (Pennebaker, 2022). LSM quantifies the degree of similarity between speakers based on the relative use of function-word categories (e.g. pronouns, articles, prepositions), which capture stylistic alignment independently of the topic. Prior research shows that higher LSM is associated with greater interpersonal rapport, with typical values in human conversations ranging higher than 0.83. In conversations between users and LLMs, LSM scores are usually below 0.7, and can be as low as 0.6 (Durandard et al., 2025).

In the present study, we compared the texts produced by the virtual agent with those produced by users. We extracted all Italian-language chats pertaining to data collection topics and removed the chatbot’s introductory message, which is identical across interactions. The resulting corpus was analysed using the Italian LIWC dictionary (Agosti and Rellini, 2007), and LIWC-based Language Style Matching (LSM) scores were computed via pairwise comparison.

Overall, LSM between the agent and users was low (approximately 0.50). Channel-specific estimates indicate an LSM of 0.56 for chats initiated from the Contact Centre general pages and 0.48 for those conducted via the Business Portal. Topic-level variation was also observed, with comparatively higher alignment in more technical exchanges, such as those concerning violation reports or survey questionnaires.

This finding is consistent with the Istat chatbot’s design: through prompt engineering, the system was configured to adopt a formal, institutional register and a constrained set of lexical choices. Moreover, due to privacy constraints, the chatbot does not retain or leverage information from prior interactions to adapt to individual users’ stylistic preferences. While these design decisions support consistency and clarity, they may also limit the system’s capacity

for dynamic stylistic accommodation, thereby contributing to the comparatively low levels of language style matching observed in the corpus.

Table 2: LSM by topic classification

Topic	LSM score
Difficulty accessing the portal or the questionnaire	0.50
Information on regulations, privacy, or response obligation	0.49
Information and clarifications on the survey or questionnaire	0.60
Booking an interview	0.48
Information on assessments and violation report	0.67
Not classified	0.53

6. Beyond Sentiment: From Communicative Patterns to Interaction Effectiveness

This section examines user–chatbot interactions from two complementary perspectives. First, it analyses lexical and semantic patterns across sentiment classes to understand how emotional polarity is distributed across different interaction types. Second, it evaluates the effectiveness of interactions through a behavioural approach, shifting the focus from expressed sentiment to actual problem resolution.

6.1 Interaction Types and Communicative Patterns Across Sentiment Classes

Sentiment analysis reveals a distribution strongly skewed toward negative polarity across all channels, with a predominance of “very negative” messages. This pattern is particularly pronounced in the Business Portal, where over half of the interactions (50.7%) fall into the most negative class, compared to 40.02% in the Contact Centre. Conversely, the Contact Centre shows a higher share of positive and very positive messages (27.79% vs 20.51%), while neutral messages remain substantial in both channels (approximately 26–29%), reflecting the coexistence of informational and emotionally connoted exchanges.

The analysis is based on a pre-trained transformer model (multilingual BERT) applied without domain-specific fine-tuning. While this ensures robustness and cross-lingual generalisation, it also introduces a systematic bias: in problem-oriented contexts, negative polarity is amplified, as the model captures the lexical surface of user frustration rather than the actual quality of the interaction.

A more granular perspective, based on thematic classification, highlights that sentiment is not uniformly distributed but closely tied to the nature of the request. Messages related to access difficulties consistently exhibit the highest levels of negativity, often exceeding 50% in the “very negative” class. Similarly, requests related to non-compliance with response obligations and associated sanctions are strongly associated with negative sentiment, confirming that constraint-driven interactions are perceived more critically. In contrast, procedural exchanges such as interview booking display a more balanced profile, with a predominance of neutral messages and limited negative intensity. Mixed-topic interactions tend to inherit the polarity of

their most critical component, especially when access issues are involved. Overall, sentiment emerges as a function of interaction type rather than a direct evaluation of the service.

To move beyond polarity alone, a topic-based lexical analysis was conducted using transformer-based embeddings and BERTopic clustering. This approach captures latent semantic structures and provides a more interpretable representation of communicative patterns.

The results show a clear functional segmentation of interactions. A first group of topics concerns access and authentication issues (e.g. login failures, password recovery, credential management), systematically associated with negative sentiment. These exchanges are characterised by repetitive and procedural language, indicating that frustration is driven by operational barriers rather than by explicit dissatisfaction with the service. A second group relates to system failures and service unavailability (e.g. HTTP errors, non-responsive channels), which also concentrates in the negative polarity class and reflects structural friction points in the interaction process.

In contrast, topics associated with neutral sentiment correspond primarily to administrative communication, metadata-like references (e.g. identifiers, codes, survey labels), and short acknowledgements. These interactions show low lexical diversity and minimal emotional content, reflecting a transactional communicative function. Positive sentiment, instead, is mainly linked to procedural guidance and interactional exchanges (e.g. requests for instructions, greetings, thanks), which signal smooth interaction rather than explicit user satisfaction.

The integration of sentiment and topic modelling leads to a key result: emotional polarity is structurally embedded in the type of interaction. Negative sentiment concentrates in contexts involving access constraints, authentication problems, and system errors, while neutral and mildly positive expressions dominate informational and procedural exchanges. This finding challenges the direct interpretation of sentiment as a proxy for service quality and highlights the limitations of relying on polarity alone.

A qualitative inspection of topic outputs further reinforces this interpretation. Courtesy and closure expressions (e.g. “thank you very much”, “perfect”) are consistently associated with positive sentiment and low informational content. In contrast, access-related expressions (e.g. “password”, “credentials”, “I cannot log in”) systematically co-occur with negative polarity. A distinct cluster captures system errors (e.g. “403 Forbidden”, “CRITICAL_ERROR”), often reported through copied technical messages and associated with high negativity. Procedural requests formulated as interrogatives (“how can I register”, “how do I proceed”) instead display a more balanced sentiment profile.

Overall, lexical evidence suggests that sentiment is primarily driven by operational friction, such as access barriers, procedural constraints, and system errors, rather than by an explicit affective evaluation of service quality. Combining sentiment analysis with topic modelling proves essential to disentangle functional and affective dimensions of communication, and to correctly interpret user-chatbot interactions in institutional contexts.

6.2 From Sentiment to Effectiveness: A Behavioural Evaluation of Interactions

The evaluation of user-chatbot interactions was further refined through a Rule-Based Behavioural Analysis, shifting the focus from purely lexical polarity to the functional success

of the exchange. To ensure an objective assessment and mitigate the bias inherent in the user's initial report of a technical problem, the analysis involved segmenting the conversations and excluding the opening message. This methodological distinction is crucial: it separates the frustration caused by a pre-existing service barrier from the user's actual reaction to the agent's support. The effectiveness of the assistance was measured using a 5-star rating scale. In this framework, a 1-star rating indicates explicit user anger or abandonment due to repeated failures, while a 5-star rating reflects explicit gratitude or confirmed problem resolution. The findings demonstrate the virtual agent's high level of effectiveness, reflected in an overall Sentiment Rating of 4 stars. Distribution data shows that 70% of interactions conclude with a positive outcome (4 or 5 stars) while critical failures (1-star) are non-existent in the analysed sample. This success is largely attributed to the agent's role as an efficient dispatcher.

By providing targeted resources, such as direct links or specific procedural codes, the agent effectively resolves operational *friction points*. Furthermore, the analysis highlights the agent's capacity to handle a diverse range of inquiries, from routine technical assistance to complex requests for statistical data. Ultimately, the transition from initial technical difficulty to a neutral or positive resolution confirms that the conversational agent successfully streamlines the support process, significantly reducing the burden on traditional assistance channels.

7. Conclusion and perspectives

The analysis of conversations between users and the Istat chatbot reveals a multifaceted picture. On the one hand, a linguistic distance emerges between the two interlocutors (repetitive vocabulary of the chatbot in contrast to users' lexical variability; LSM of 0.50), reflecting the agent's formal style and its limited capacity to adapt to users due to privacy constraints. On the other hand, the negativity detected by sentiment analysis does not reflect a negative evaluation of the service but is instead selectively associated with specific technical issues (e.g., portal access and system errors).

In terms of effectiveness, the chatbot achieves positive results (rating of 4 out of 5 stars; 70% positive outcomes), supported by a clear strategy: acting as a dispatcher, providing targeted resources, and maintaining a predictable register. Despite linguistic differences and initial frustration, the virtual agent resolves the majority of requests, thereby reducing the burden on human operators.

From a broader perspective, this study contributes to the literature on human–chatbot interaction by providing empirical evidence from a public institutional context, which remains relatively underexplored compared to private-sector applications. It shows that in regulated environments characterised by formal communication and strong privacy constraints, effectiveness does not necessarily depend on high levels of linguistic alignment or personalisation. Instead, clarity, procedural guidance, and reliability emerge as key drivers of successful interactions.

At the same time, the contribution of this study should be interpreted considering the specific characteristics of the case analysed and the available data. Future research may extend the analysis using a comparative approach, integrate user-centred evaluation methods, and explore forms of linguistic adaptation compatible with regulatory constraints.

References

- Agosti A. and Rellini A. (2007). *The Italian LIWC dictionary*. Austin, TX: <https://www.liwc.app>
- Baayen R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.
- Bolasco S. (1999). *L'analisi multidimensionale dei dati*. Carocci Editore.
- Chen D., Chen H., Yang Y., Lin A. and Yu Z. (2021). Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In Toutanova K., Rumshisky A., Zettlemoyer L., Hakkani-Tür D., Beltagy I., Bethard S., Cotterell R., Chakraborty T. and Zhou Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3002–3017.
- Durandard N., Dhawan S. and Poibeau T. (2025). Language Style Matching in Large Language Models, *Proceedings of the 26th Annual Meeting of SIGDIAL*, pages 620–636
- Yildiz Durak H., Eğin F. and Onan A., (2025). A Comparison of Human-Written Versus AI-Generated Text in Discussions. *Education and Information Technologies*, 30(1): 1–15.
- Park M., Aiken, M. and Salvador L. (2018). How do Humans Interact with Chatbots?: An Analysis of Transcripts. *International Journal of Management and Information Technology*, 14, 3338-3350.
- Pennebaker J. (2022). LIWC-22 Tutorial 5: Language Style Matching.
- Reviriego P., Conde J., Merino-Gómez E., Martínez G. and Hernández J. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. *Machine Learning with Applications*, 18(1): 100602.
- Rodrigues F. A., Sturm N. F. and Pinheiro F. L. (2026). A linguistic comparison between human- and AI-generated content. *iScience*, 27(6), 109860.
- Sandler M., Choung H., Ross A. and David P. (2024). *A Linguistic Comparison between Human and ChatGPT-Generated Conversations*. In *International Conference on Pattern Recognition and Artificial Intelligence*, Springer Nature Singapore, pp. 366-380.

A Methodological Framework for Knowledge Graph Design Using Large Language Models

Raffaella Maria Aracri¹, Alessio Buonocore², Adriana Fidanza³, Roberta Radini⁴

¹ISTAT — aracri@istat.it

²ISTAT — alessio.buonocore@istat.it

³ISTAT — adriana.fidanza@istat.it

⁴ISTAT — r.radini@istat.it

Abstract (in English)

This study introduces a comprehensive methodological framework for the design and development of knowledge graphs as a foundational phase of systematic ontology engineering. The approach leverages the linguistic, analytical, and generative capacities of large language models (LLMs) to support domain exploration, conceptual abstraction, and the formalization of entities, relationships, and constraints. By integrating LLMs into the ontology development lifecycle, the framework aims to reduce modeling complexity, enhance structural coherence, and improve the expressive adequacy of resulting artifacts while optimizing time and cognitive resources. The study further provides a systematic and critical evaluation of knowledge graphs generated by LLMs. Given that early-stage ontology construction requires the rigorous organization of conceptual hierarchies and explicit semantic specification, particular attention is devoted to assessing structural integrity, internal consistency, output stability, and reproducibility. The proposed evaluation criteria also measure alignment with expert-validated ontological references. Through controlled experimental analysis, the research examines both the strengths and limitations of LLMs in knowledge engineering, moving beyond anecdotal claims to assess their epistemic reliability in formal modeling tasks. The broader objective is to establish a robust methodological foundation for the responsible integration of generative models into knowledge representation practices, fostering dialogue between statistical modeling, artificial intelligence, and domain-specific conceptual modeling.

Keywords: Ontology - Large Language Models - Ontology Engineering

1. Introduction

This work frames the processing workflow and the tools that support the management of semantic assets required for interoperability among Public Administrations. Significant investment has been devoted to this area through the National Recovery and Resilience Plan, which led to the implementation of the “*National Data Catalog (NDC)*” for semantic interoperability and to the definition of a semantic network modeling data of the various administrative processes of the Public Administration. The development of the portal <https://schema.gov.it/> and its semantic products has been carried out under an agreement between ISTAT and the Department for Digital Transformation. The set of contents and the platform together constitute the Catalog, designed to support the discovery and reuse of semantic resources — such as ontologies, data schemas, and controlled vocabularies — in order to ensure semantic interoperability both at the structural level and in the meaning of exchanged information (Aracri et al., 2024; ISTAT, 2023). A key factor for interoperability is the extension of the semantic network to all sectors of the Public Administration. This requires substantial investment not only in ontology development but also in the availability of domain experts with specialized skills, which are currently

scarce within public administrations. To address this need, and the associated critical challenges, ISTAT — acting as the methodological authority specifically entrusted with the design, development, and implementation of the Catalog — has focused on exploring tools capable of supporting and partially automating knowledge modeling activities. In particular, this study investigates the potential role of Large Language Models (LLMs) as support systems for early-stage ontology engineering tasks, including concept extraction and preliminary structuring of knowledge graphs. Rather than positioning LLMs as autonomous ontology builders, this work adopts a more conservative perspective: LLM-generated outputs are treated as intermediate artifacts that require validation and refinement by domain and ontology experts. The contribution of this paper is twofold. First, it proposes a structured methodological framework for integrating LLMs into the ontology development workflow. Second, it presents an empirical evaluation of the extent to which LLM-generated knowledge graphs align with an expert-defined ontology, focusing primarily on the extraction of conceptual elements.

2. Background and Motivation

The increasing availability of unstructured textual data has intensified the need for methods that can transform such data into structured, machine-interpretable knowledge. In many application domains, this transformation is essential for enabling data integration, semantic interoperability, and advanced analytical workflows. Two key paradigms have emerged as particularly relevant in this context: ontologies, which provide formal and interpretable representations of knowledge, and Large Language Models (LLMs), which offer powerful capabilities for extracting and organizing information from natural language. This work is motivated by the complementary strengths and limitations of these two approaches, and explores their combined use within a unified framework.

2.1. Ontologies

Ontologies play a central role in knowledge representation by providing explicit, formal specifications of conceptualizations within a given domain. In the context of computer and information sciences, they define a set of representational primitives—such as classes, properties, and relationships—that can be used to model domain knowledge in a structured and semantically meaningful way. These primitives are accompanied by formal definitions and constraints that ensure logical consistency and enable automated reasoning (Gruber, 2009). Unlike database schemas, which operate primarily at the logical or physical level of data organization, ontologies are situated at a higher semantic level. This abstraction allows them to serve as a unifying layer across heterogeneous data sources, facilitating interoperability between disparate systems and supporting the integration of distributed knowledge bases. Standardized languages such as the Web Ontology Language (OWL), grounded in description logics, provide the formal foundations necessary for expressive and computationally tractable modeling (W3C, 2012). Despite the availability of mature tools and established standards, ontology engineering remains a complex and cognitively demanding task. It requires both domain expertise and familiarity with formal modeling principles, which can limit scalability and adoption. In this work, ontologies are treated as the reference semantic structures against which other representations are evaluated.

2.2. Large Language Models

Large Language Models (LLMs) are computational models designed to process and generate human language. They are trained on large collections of text and learn statistical patterns that

allow them to produce coherent and contextually relevant outputs. In recent years, LLMs have demonstrated strong performance across a wide range of language-related tasks, including text summarization, question answering, and information extraction (Vaswani et al., 2017). One of their most relevant capabilities in this context is the ability to transform unstructured text into more structured forms of information, such as lists of entities and relationships or graph-like representations. This makes them potentially useful for supporting knowledge extraction processes that would otherwise require significant manual effort (Bender et al., 2021). However, the use of LLMs in knowledge representation tasks presents important limitations. Their outputs are probabilistic in nature and can vary depending on how a task is formulated. They do not inherently enforce logical consistency, nor do they rely on formally defined semantic structures. As a result, the information they generate may be incomplete, ambiguous, or inconsistent when evaluated against formal representations such as ontologies (Lippolis et al., 2025). For these reasons, LLMs are not treated in this work as authoritative sources of knowledge, but rather as assistive tools that can support the initial stages of knowledge extraction. In particular, knowledge graphs generated with LLM support are considered preliminary approximations of the underlying semantic structure, which can subsequently be refined and validated through ontology-driven methods.

3. Data Sources

This work is based on two main datasets: (i) a complete set of domain textual documentation, which represents the primary source from which relevant concepts and relationships are extracted; (ii) an ontology formally modeled by human experts based on such documentation. All documentation files are available on the Github repository¹.

3.1. Textual Documentation

Within the scope of the National Data Catalog (NDC), the OpenCUP² ontological modeling project was considered particularly relevant for this study, as it is supported by rich, formal, and comprehensive textual documentation conducted by ISTAT (Aracri et al., 2025). This documentation served as the reference basis for ontology construction by ontology modeling experts, and was subsequently shared and validated by domain experts. To make the input manageable for LLM processing, the document was segmented into smaller self-contained units, defined by a chapter (*cap#*) and a paragraph (*def#*), named below as *cap#_def#*.

3.2. Expert-Driven Ontology

A manually constructed ontology serves as the ground-truth reference. It is formalized in OWL2 and validated by domain experts. This expert-driven ontology plays a crucial role in the experimental process: it is used as a validation reference to compare the knowledge graphs automatically generated by LLMs, enabling a systematic evaluation of how closely the automatically generated ontological structures match those manually constructed, as well as whether additional concepts are introduced beyond the expert-defined ontology from the same documentary basis. While this ontology provides a strong baseline, it should be noted that it reflects modeling choices, not an absolute ground truth, as alternative valid conceptualizations may exist.

¹ <https://github.com/istat-methodology/llm-ontology-designer>

² <https://www.opencup.gov.it/portale/web/opencup/home>

4. Data Processing Workflow

The proposed methodology is implemented through a structured data processing workflow designed to transform unstructured textual input into an intermediate graph-based representation. The workflow is composed of three main stages: (i) input acquisition and preprocessing, (ii) concept identification via Large Language Models (LLMs), and (iii) graph construction and visualization. In the first stage, textual data are ingested and prepared for analysis. This step ensures that the input is suitable for downstream processing while preserving its semantic content. In the second stage, LLMs are employed to extract relevant concepts and relationships from the text. This process leverages prompt-based interaction to guide the model toward identifying domain-relevant entities, attributes, and connections. The extraction phase represents the core of the methodology, as it translates unstructured knowledge into a structured form. In the final stage, the extracted elements are organized into a graph representation. Rather than producing fully formal ontologies, the system generates simplified graph structures in which classes and attributes (data properties) are represented as nodes, while relationships between classes (including subclass and object-property relations) are represented as edges enriched with labels and descriptions. This representation is intentionally designed as an intermediate modeling artifact. On the one hand, it enhances usability by providing an intuitive and accessible visualization that supports human understanding and interaction with the emerging ontology. On the other hand, it retains sufficient expressive capacity to capture the primary ontological constructs required for subsequent refinement. The resulting graphs are therefore intended to serve as a starting point for more complete and formal ontology development, potentially in collaboration with an AI-assisted designer. The workflow described above was implemented in a prototype web-based system to support experimentation and validation. During development, multiple LLMs (GPT-5-mini, GPT-4.1, GPT-5.1³) were tested to explore their suitability for the task. However, the evaluation does not aim to provide a systematic comparison of model performance under controlled conditions. Instead, a single optimized configuration was selected for the final experiments, prioritizing output consistency and alignment with the methodological objectives.

4.1. Prompt Design Strategy

Prompt engineering is the set of rules for crafting input texts for LLM interfaces. The prompt aims to guide the language model toward coherent and optimized responses aligned with user requirements. Interaction through prompts is typically an iterative process, starting with an input message that defines the operational context, key instructions, and desired output; subsequent prompts adapt language, wording, and instructions based on the model's responses to progressively refine results (De Gasperis, 2025; Wei et al., 2022). In our case, interactions performed to optimize the prompt were stored in order to select the most effective prompt while keeping the input constant and controlling the output. The prompt was designed using a hierarchical structure with explicitly defined sections, enabling clear readability for both humans and language models, increasing the chances that all components are respected and that the output meets expectations. To optimize the model's performance and ensure output consistency, the prompts were structured following the TCREI framework (Task, Context, References, Evaluate, Iterate). This methodology, widely recognized as an industry standard for systematic prompt design, provides a modular approach to minimizing ambiguity in Large Language Model (LLM) interactions (GoogleCloud, 2024). While this process improved output quality, it introduces a key limitation: results are dependent on prompt design, which may reduce generalizability. Table 1

³<https://developers.openai.com/api/docs/models>

describes this section-based prompt structure.

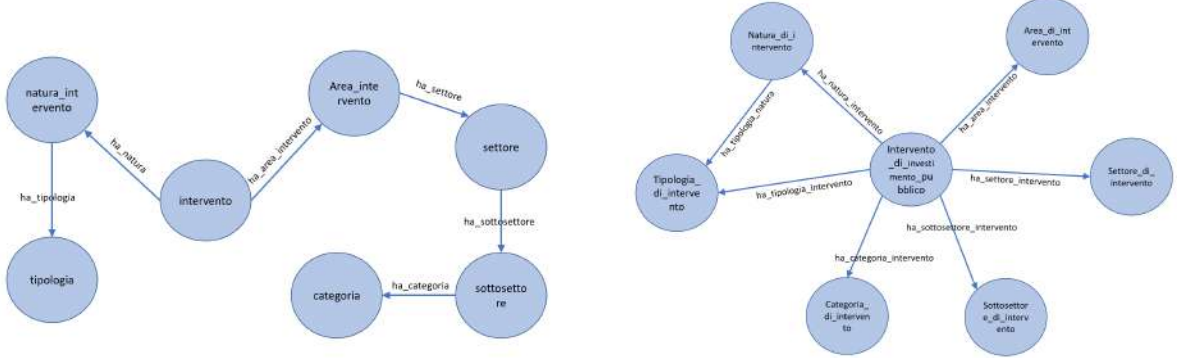
Section	Function	Implementation
Task	Describe exactly what you want. Include a persona (what expertise should the AI draw from?) and a format preference (how should the output look?).	<i>"Sei un assistente esperto di ontologie, modellazione semantica e tecnologie legate al web semantico (OWL, RDF, RDFS). L'output deve essere un JSON valido"</i>
Context	Provide necessary details to refine the output. What specific details would improve AI's understanding?	<i>"Il tuo compito è analizzare un testo di dominio ed estrarre una lista di classes, object properties e data properties nel linguaggio owl adottando le seguenti regole: (i) considera solo i concetti semanticamente rilevanti; (ii) Per ogni concetto riporta la sua descrizione in linguaggio naturale solo se presente nel testo; (iii) evidenzia le relazioni di ISA tra i concetti; (iv) non inventare nè classi nè descrizioni se non sono presenti nel testo; (v) i nomi delle classi devono essere scritti in notazione 'snake_case'."</i>
References	Include examples to guide AI's response.	single-shot prompting (one example)
Evaluate	Review the generated output against your initial requirements.	Identify gaps, errors, or tone inconsistencies, often asking the model to self-correct or critique its own response based on the criteria provided.
Iterate	Refine the prompt based on the evaluation. Prompt engineering is a cyclical process.	Tweak the instructions, add missing context, or clarify the task until the output meets your professional standards.

Table 1: Section-based prompt structure

5. Results

In the initial phase of the analysis, we adopted a systematic data collection protocol to ensure full traceability and reproducibility of all experimental trials. For each configuration, we retained the model specification, the prompt formulation, the input text, and the corresponding outputs, including predicted class labels and generated ontology elements. This design enabled a coherent comparison across models and prompt variants, while preserving the necessary information for subsequent analyses. Such a structured setup also supported an iterative refinement process, allowing prompt configurations to be adjusted in relation to the specific characteris-

tics of each model under consistent input conditions. The evaluation was primarily qualitative, based on visual inspection of the generated outputs, with particular attention to structural consistency and semantic alignment with the reference ontology. The comparison between obtained (figure 1a) and expected results (figure 1b) enabled iterative refinement of the prompt, leading to a final solution that was eventually used for all experiments.



(a) Portion of knowledge graph generated from LLM (based on cap3_def3 text file)

(b) Portion of Real Ontology about public investment intervention

Figure 1: Comparison between Generated Graph and expert driven ontology, about public investment intervention

5.1. Evaluation Methodology

The second part of the analysis focused on validating the classification results (graph nodes): it is framed as a semantic matching problem between two sets of labels. Sentence embeddings and cosine similarity are used to compute pairwise similarity, followed by optimal matching via the Hungarian algorithm. We evaluate the quality of label extraction by comparing the set of predicted labels generated by the LLM against a reference set of ground-truth labels. Since the two sets may differ in size and are not aligned one-to-one, we frame the problem as a set matching task under semantic similarity, rather than a standard classification problem. To account for lexical and semantic variability in label generation, each label is encoded into a dense vector representation using a pretrained sentence embedding model (specifically, Sentence Transformers with the *all-MiniLM-L6-v2* checkpoint⁴). Pairwise similarities between predicted and ground-truth labels are computed using cosine similarity, a standard metric for comparing embeddings in vector space (Cosine Similarity) (You, 2025). Given the resulting similarity matrix, we compute an optimal bipartite matching between the two sets using the Hungarian Algorithm (Kuhn, 1955), which maximizes the total semantic similarity across matched pairs. This step ensures a globally optimal alignment and avoids the suboptimality of greedy matching strategies. After obtaining the optimal assignment, we apply a similarity threshold to determine whether a matched pair is considered correct. Matches above the threshold are counted as True Positives (TP), while unmatched ground-truth labels are counted as False Negatives (FN), and unmatched predicted labels as False Positives (FP). True negatives are not explicitly defined due to the absence of a closed label space, which is typical in open-set or generative settings. From these quantities, we compute standard evaluation metrics including precision, recall, and F1-score. To ensure robustness, we also analyze the sensitivity of the results with respect to the similarity threshold and report the distribution of similarity scores for matched pairs. Additionally, we provide qualitative error analysis by inspecting representative true positive and false negative

⁴ More information here: <https://sbert.net/>

cases, highlighting borderline semantic mismatches. This evaluation framework is particularly suitable for LLM-based generation tasks, where outputs exhibit high lexical variability but may still be semantically correct. It aligns with recent practices in semantic textual similarity and embedding-based evaluation (Reimers and Gurevych, 2019).

5.2. Confusion Matrices Results

The choice of similarity threshold significantly affects results. While a threshold of 0.5 was selected based on F1 trends, this decision remains heuristic and may bias evaluation outcomes. The table 2 summarizes the results, with the optimal prompt fixed for all self-consistent paragraphs, as described in Section 3.1.]. This approach provides a practical approximation but introduces important assumptions: (i) semantic similarity is used as a proxy for ontological equivalence; (ii) results depend on the embedding model and threshold selection.

Text Title	TP	FP	FN	Precision	Recall	F1
<i>cap2_def1</i>	4	7	0	0.36	1	0.53
<i>cap2_def4</i>	1	6	2	0.14	0.3	0.20
<i>cap2_def8</i>	4	0	4	1	0.5	0.67
<i>cap2_def9</i>	5	13	0	0.27	1	0.43
<i>cap3_def1</i>	19	6	5	0.76	0.79	0.77
<i>cap3_def2</i>	9	26	4	0.25	0.69	0.37
<i>cap3_def3</i>	13	6	7	0.68	0.65	0.67
<i>cap3_def4</i>	13	9	7	0.59	0.65	0.62

Table 2: Summary table obtained with a similarity threshold of 0.5

The results show high variability across texts. Some cases exhibit high precision but low recall (conservative extraction); others show high recall but low precision (over generation). Only a subset achieves balanced performance, that suggests LLMs capture the semantic neighborhood of concepts but struggle with ontological boundaries. Performance differences between Chapter 2 ($\overline{F1} \sim 0,45$) and Chapter 3 ($\overline{F1} \sim 0,61$) suggest that descriptive and data-oriented text is easier to process than normative and definitional text, that poses greater challenges. This was also observed during the ontology modeling, since the Ontology designers themselves mainly based their work on Chapter 3.

5.3. Insight Analysis

Below, we report the results for a single text file, namely *cap3_def1*, including representative examples from the 19 True Positives (TP), 6 False Positives (FP), and 5 False Negatives (FN) identified in the classification. In tables 3 and 4, the *True value* corresponds to the expected class label as defined in the ontology, while the *Predicted value* refers to the class label extracted by the LLM. The similarity score is used to determine whether a predicted class matches the expected ontology class, based on a predefined threshold of 0.5.

Table 3 reports some samples of matched classes (i.e. True Positive). As shown some classes—such as *coerente_con_programma_spesa* and *completo*—are correctly matched to their corresponding ontology values despite relatively low similarity scores. Conversely, other classes, such as *codice_unico_progetto*, exhibit similarity scores close to 1, indicating a near-exact match.

Table 4 reports some samples of unmatched cases, where the predicted classes differ signif-

icantly from the expected ones, as reflected by their low similarity scores. It is important to note that the classification is defined from two complementary perspectives. *Predicted values*, reported in the table have similarity scores below the predefined threshold (0.5) and therefore do not match the expected ontology classes (*True values*). From the perspective of the extracted predictions, these unmatched instances (*Predicted values*) are classified as False Positives. Conversely, from the perspective of the ground truth, the corresponding expected classes (*True values*) that are not correctly retrieved are classified as False Negatives.

True values	Predicted values	Similarity score
Codice_unico_di_progetto	codice_unico_progetto	0.97
Stato_completo_di_cup	completo	0.59
Stato_coerente_programma_spesa_di_cup	coerente_con_programma_spesa	0.76
Stato_di_cup	stato_cup	0.93
Soggetto_concentratore	soggetto_concentratore	1

Table 3: Samples of matched classes for *cap3_def1* text file

True values	Predicted values	Similarity score
Stato_attivo_di_cup	generato_attivo	0.44
Stato_di_atto_di_finanziamento	normale	0.26
Stato_di_coerenza_di_atto_di_finanziamento	corredo_informativo	0.35
Stato_di_emanazione_di_atto_di_finanziamento	sostituzione	0.44

Table 4: Samples of Unmatched classes for *cap3_def1* text file

Overall, the proposed evaluation should be interpreted as a preliminary and coarse-grained analysis of the LLM’s capability to reproduce the ontology structure. The current assessment focuses exclusively on node-level alignment, i.e., the correspondence between extracted labels and ontology classes, without considering relational aspects of the graph. Results show a highly variable behavior across examples: while some cases (e.g., *cap2_def8*) achieve perfect precision but limited recall, indicating conservative predictions, others (e.g., *cap2_def1* and *cap3_def2*) exhibit high recall but very low precision, suggesting a tendency of the model to over-generate semantically related but ontologically incorrect classes. More balanced performance is observed in *cap2_def6*, *cap3_def1* ($F1 \approx 0,77$) and *cap3_def3* ($F1 \approx 0,67$), although errors are still present. These discrepancies highlight that the model is often able to capture the semantic neighborhood of the target concepts, but struggles to adhere strictly to the ontology boundaries. Importantly, this evaluation does not account for the relational structure of the ontology, which remains an open limitation. Extending the analysis to include object properties (relations between entities) and data properties (attributes of entities) would provide a more comprehensive assessment of the generated knowledge graph. Consequently, the current results should be seen as an initial step, useful for diagnosing general trends, but insufficient for fully validating the structural correctness of the generated representations.

6. Conclusion

The proposed framework demonstrates that LLMs can effectively support early-stage ontology development, particularly in tasks involving concept extraction from structured textual descriptions. However, the experimental results reveal substantial variability and sensitivity to both input characteristics and evaluation parameters. Overall, the findings indicate that LLMs are better suited for generating candidate conceptual structures rather than producing formally valid ontologies. As such, their outputs should be treated as intermediate artifacts requiring systematic validation, rather than as substitutes for expert-driven modeling processes. In this perspective, the ontology development workflow is best framed as a human-in-the-loop (HITL) process, where AI systems provide automated suggestions and human experts iteratively refine and validate the resulting structures. From a methodological standpoint, this study also highlights limitations in current evaluation practices for generative ontology construction. In particular, node-level semantic matching provides only a partial proxy for quality and does not fully capture ontological correctness. Future work should therefore focus on: (i) extending the evaluation to relational structures; (ii) improving robustness with respect to prompt variation; (iii) validating the approach across multiple domains; and (iv) developing more principled and comprehensive evaluation metrics.

References

- Aracri R. M., Buonocore A., Concetti E., D'Alessandro G., A. F., Pantoni S., Barbieri B., Falcetta E., Fiorentini L., Radini R., and Puglisi R. (2025). Codice unico di progetto (cup). documento di analisi per la modellazione ontologica. Internal document, ISTAT - Istituto Nazionale di Statistica. general register: 0234414/26 del 20/01/2026.
- Aracri R. M., Frisardi D., Radini R., and Santarelli V. (2024). Empowering e-services through the semantic web. *ceur-ws.org*, 37-62:452–457.
- Bender E. M., Gebru T., McMillan-Major A., and Shmitchell M. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623.
- De Gasperis P. (2025). *Dialogare con i Large Language Models. Guida alla scrittura del prompt*. Edizioni Nuova CULtura.
- GoogleCloud (2024). Prompt engineering: Best practices for generative ai models. Official Documentation.
- Gruber T. (2009). Ontology. In Ling Liu M. T., editor, *Encyclopedia of Database Systems*. Springer-Verlag.
- ISTAT (2023). *Trasformazione digitale della Pubblica Amministrazione. Metodi per l'interoperabilità per lo sviluppo di e-service*, volume Letture statistiche – Metodi. Istituto nazionale di statistica.
- Kuhn H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.
- Lippolis A. S., Saeedizade M. J., Keskisarkka R., Zuppiroli S., Ceriani M., Gangemi A., Blomqvist E., and Nuzzolese A. G. (2025). Ontology generation using large language models. *Lecture Notes in Computer Science*, pp. 321–341.
- Reimers N. and Gurevych I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992.
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., and Polosukhin I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008.
- W3C (2012). Owl 2 web ontology language document overview (second edition). W3c recommendation, W3C.
- Wei J., Wang X., Schuurmans D., Maarten B., Ichter B., Xia F., Chi E., Le Q., and Zhou D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:24824–24837.
- You K. (2025). Semantics at an angle: When cosine similarity works until it doesn't. *arXiv*.

C08

Topic modeling, classification and interpretability

09-lug
09:00 - 10:30
Aula 12

Editorial Transitions and Thematic Change in the American Journal of Sociology: A Structural Topic Modeling Analysis

Stefano Sbalchiero¹, Valentina Rizzoli², Arjuna Tuzzi³

¹University of Padova – stefano.sbalchiero@unipd.it

²Sapienza University of Rome – valentina.rizzoli@uniroma1.it

³University of Padova – arjuna.tuzzi@unipd.it

Abstract

This study investigates whether editorial transitions are associated with changes in the thematic structure of the American Journal of Sociology. We analyze 4,249 article abstracts using a Structural Topic Model (STM), incorporating editor-in-chief and publication year as prevalence covariates and taking into account overlapping editorial periods. Texts were pre-processed using standard lexicometric procedures, including the consolidation of domain-specific multiword expressions. Models with 5 to 30 topics were estimated and evaluated using held-out likelihood, residuals, semantic coherence, and exclusivity. The results reflect both enduring core areas of sociological research such as race and inequality, urbanization, family, gender, and stratification as well as methodological, theoretical, and political themes. Analyses of topic prevalence across editors reveal systematic variations rather than abrupt thematic breaks, suggesting that editorial leadership subtly reshapes the journal's intellectual emphasis over time. Overall, the findings highlight the role of editorial governance in influencing the evolution of disciplinary agendas and demonstrate the usefulness of STM for studying institutional and epistemic change in academic publishing.

Keywords: Structural Topic Model, Editorial transitions, Academic journals, Sociology

1. Introduction

Sociology can be understood as a discipline continuously engaged in interpreting social change and therefore deeply shaped by the historical and institutional contexts in which it develops. Its intellectual trajectory has not been uniform across countries or periods (Calhoun, 2007; Clemens et al., 1995; Giordan et al., 2018). The sociology we know today is the outcome of a long process of theoretical debates, methodological innovations, and institutional transformations that unfolded across different historical and geographical settings (Collins, 1986; Connell, 1997). Throughout the twentieth century the discipline expanded in scope and complexity, producing an increasing diversification of perspectives and research areas. Although a number of studies have examined the historical development of sociology (Kalekin-Fishman and Denis, 2012; Turner, 1998; Scott and Desfor Edles, 2011), this article builds on the idea that exploring the discipline's history can be highly instructive. Understanding the conditions and trajectories of scientific work may be as important as theoretical frameworks and research methods in advancing sociological knowledge (Calhoun, 2007). Existing historical research has offered valuable paradigmatic reflections on the discipline; however, fewer studies have relied on empirical data and systematic evidence to analyze the evolution of sociological topics over time. While alternative bibliometric approaches (e.g., those implemented in the Bibliometrix framework; Aria and Cuccurullo, 2017) could be applied to similar research questions, the present study builds on a distinct tradition of text-based analysis of scientific

literature, aiming to capture the interpretive and contextual dimensions of topic evolution. From this perspective, this study focuses on the editorial transitions and thematic change topics within the *American Journal of Sociology* (AJS), with the aim of contributing to what can be described as a digital history of scientific disciplines (Cohen et al. 2008; Rizzoli et al., 2023). The history of a discipline is largely the history of ideas (Tuzzi, 2018), shaped by the contributions of prominent scholars and by the continuous boundary work between classical theories and emerging social problems. Such a history is traditionally reconstructed through milestone research works and the evolving dialogue between established paradigms and new intellectual developments. The purpose of this paper is grounded in the consideration that methods of research should include qualitative perspectives capable of interpreting the results obtained through quantitative and computational approaches. This implies focusing on the evaluation of contents to detect relevant features such as topic structures, their evolution over time, and the role played by the gatekeepers of science. Following some preliminary case studies on the history of scientific disciplines in the field of social sciences (Giordan et al., 2018; Sbalchiero & Tuzzi, 2018), which support the feasibility of the present research, this paper focuses on a large corpus of abstracts from a mainstream journal in the field of sociology. Despite the classical tradition represented by Durkheim, Simmel, and Weber (Connell, 1997; Calhoun, 2007), the literature on the history of sociology has paid limited attention to digital methods, and particularly to the statistical analysis of textual data aimed at detecting the temporal evolution of old, new, and emerging topics within the discipline. Such analyses can help identify chronological and comparative shifts in paradigms, tracing variations in the prominence of topics as they increase or decline over time, while also considering the institutional contexts of scientific journals and the influence of their editors-in-chief. This requirement stems from the awareness that scientific cultures, like other cultural systems, are characterized by porous boundaries (Shapin, 1995; Livingstone, 1995) which make them fluid rather than monolithic and homogeneous entities. Consequently, understanding the evolution of sociological knowledge requires engaging with the complexity of the contexts within which scientific ideas are produced and circulated. To interpret the evolution of topics, the present paper also considers the role played by editors-in-chief as “gatekeepers of science” (Crane, 1967). If we consider that the selection of ideas and the publication of certain topics in a scientific journal may inhibit the circulation of other ideas, it becomes possible to analyze which topics are effectively permitted to circulate within the journal. This implies examining topics from a chronological perspective and interpreting them within their contexts by connecting their evolution to changes in the editors-in-chief. Under this hypothesis, the trajectory of scientific topics may reflect both exogenous social factors and endogenous institutional dynamics. In the first case, topics may indicate how sociologists have reacted to emerging social problems and phenomena over time. In the second case, changes in the scientific production of editors-in-chief may influence the trajectories of the journal by shaping the selection of ideas that are published. If this is the case, journals may reposition their scientific agendas not only in response to broader social changes but also in relation to internal factors derived from the scientific profiles of their editors. As already noted, “the editors [of AJS], by their selection of papers for publication, have helped to delineate the field of sociology” (Shanas, 1945, p. 522). This possibility has long been suggested in the literature, but it has rarely been empirically tested.

2. Data and methods

The *American Journal of Sociology* (AJS), established in 1895 as the first scholarly journal in sociology in the United States, is widely regarded as one of the most prestigious journals in the

field and a leading venue for research in the social sciences. AJS is a multidisciplinary journal that aims to address a “general sociological reader” and the articles published in the journal are generally considered representative of the most significant contemporary theoretical and methodological debates in sociology. Our corpus includes all abstracts of papers published in AJS over more than a century (1921–2024), which were retrieved from major academic databases (Scopus and Web of Science) as well as from the journal’s official webpages. We chose to focus on abstracts because they provide concise and standardized summaries of the main contents of each article. At the same time, focusing on abstracts rather than full texts may favor the emergence of broader and more consolidated thematic structures, while potentially reducing the granularity of highly specialized or secondary themes that could appear in full-length articles. Nevertheless, abstracts were considered particularly suitable for the present longitudinal analysis because they provide relatively stable and comparable representations of article content across different historical periods (Tuzzi, 2018).

The selection criteria were based on the following consideration: abstracts that did not provide substantive information about the article’s content or did not correspond to relevant scientific contributions (e.g., editorials, mastheads, errata, acknowledgements, rejoinders, notes, announcements, corrections, lists of consultants, obituaries, etc.) were excluded from the analysis.

The final dataset includes 4,249 abstracts from the American Journal of Sociology. Each record corresponds to the abstract of a single article and contains six variables: publication year, volume, issue number, and an identifier for the editor-in-chief in charge at the time of publication. Because editorial transitions did not always occur at clear yearly boundaries, some years present overlapping editorships. In those cases, the dataset records the surnames of both editors - corresponding to the outgoing and incoming editor - allowing the analysis to account for transitional editorial periods. In the STM specification, overlapping editorships were treated as distinct editorial categories, whereas the graphical display reports only selected single-editor categories for readability.

Pre-processing was carried out using TALTaC². All texts were converted to lowercase to ensure consistency across the corpus. In addition, a set of domain-relevant multiword expressions (Pavone, 2010), such as *public good*, *white house*, and *panel data*, together with other validated polirematic expressions and compounds, was consolidated and treated as textual units (n=787). The corpus was subsequently processed using the *stm* R-package for Structural Topic Modeling (Roberts et al., 2019). In particular, the *textProcessor* function from the *stm* package was used to perform standard text-cleaning procedures prior to model estimation. In this step, punctuation marks and numbers were removed, as well as common grammatical words (e.g., articles, conjunctions, prepositions, and pronouns) using the default English-language stopword list. TALTaC² was employed in a preliminary phase to identify and consolidate domain-specific multiword expressions and validated polirematic compounds, while *textProcessor* was used for corpus standardization and tokenization within the STM workflow. Neither stemming nor lemmatization procedure was applied. This choice was motivated by the historical and interpretative nature of the analysis. Given the long temporal span of the corpus (1921–2024), preserving lexical forms was considered preferable in order to avoid merging historically or contextually distinct terms into overly generalized stems. Moreover, previous STM applications on scientific corpora have shown that aggressive lexical normalization may reduce interpretability by obscuring disciplinary nuances and theoretically meaningful distinctions between terms (Rizzoli et al., 2023 ; Sbalchiero and Eder, 2020).

Table 1. Basic lexical measures of the corpus of AJS abstracts

Token (N)	556,882
Type	21,817
Type/Token Ratio (%)	3.92%
Hapax legomena (%)	41.01%

The corpus is composed of 21,817 word-types and 556,882 word-tokens (tab. 1), and the measures (TTR and Hapax legomena %) indicate a sufficient level of lexical redundancy in the corpus, allowing the application of statistical methods for textual data analysis (Lebart et al., 1998; Bolasco, 2013; Tuzzi, 2024). To further contextualize the corpus and the prevalence covariates used in the STM specification, Table 2 summarizes the main descriptive characteristics of the dataset, including the temporal coverage of the corpus and the distribution of editorial configurations.

Table 2. Main descriptive characteristics of the dataset

Time span	1921–2024
Number of abstracts	4,249
Number of editors/editorial configurations	34
Mean abstracts per year	40.86
Years with overlapping editorships	18

3. Topic detection

To select an appropriate number of topics, we used the searchK R-function by estimating models with a range of solutions from 5 to 30 topics. The specification included a prevalence formula incorporating both the editor-in-chief and a smooth function of publication year. Building on recent developments in text analytics (Sciandra et al., 2025), this procedure generated the standard STM diagnostics—held-out likelihood, residuals, semantic coherence, and lower bound—to which we additionally considered exclusivity as an interpretative criterion. The diagnostic plots (Figure 1) suggest that the optimal number of topics lies within a moderate range rather than at the higher end of the tested solutions. Held-out likelihood increases up to approximately 10–12 topics and then declines, indicating a reduction in predictive performance when the number of topics becomes too large.

Semantic coherence also decreases progressively as K increases, suggesting that larger models produce more fragmented and less interpretable topics. Residuals show a marked improvement up to approximately $K = 15$, after which further gains become minimal. Exclusivity increases with higher values of K but tends to plateau around 18–20 topics, and this improvement occurs at the expense of semantic coherence. When considered jointly, these diagnostics point to a reasonable balance between coherence, exclusivity, and model fit in the range of 10 to 15 topics. After a qualitative inspection of all solutions between 10 and 15 topics, we retained a model with 15 topics, which lies within the optimal range suggested by the diagnostics (Sbalchiero and Eder, 2020; Rizzoli et al., 2023). Beyond approximately 20 topics, the model is likely to overfit the data, producing topics that are less meaningful and more difficult to interpret.

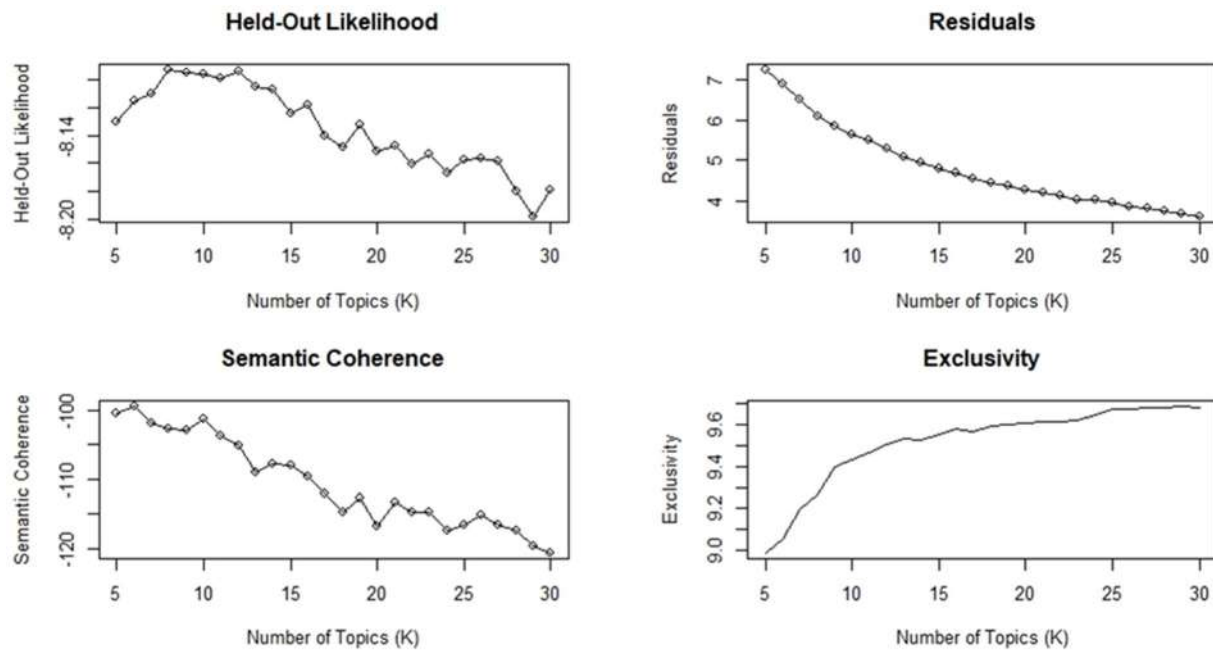


Fig. 1. Fitting the model: Diagnostic values by number of topics

3.1. Results

The fifteen topics cover the main domains of sociological inquiry, including inequality, institutions and organizations, family and gender, political and economic processes, theoretical reflection, and methodological development. Topic 1 concerns transnational activism and social movement networks (*transnational, activists, mobilization*); Topic 2 reflects discussions of war and collective morale (*war, soldier, morale*); Topic 3 addresses racial inequality and segregation (*black, segregation, inequality*); Topic 4 focuses on urban and community dynamics (*cities, urban, metropolitan*); Topic 5 refers to conceptual and social-psychological foundations of sociology (*social psychology, concept, social phenomena*); Topic 6 captures organizational structures and power relations (*organizations, hierarchy, firms*); Topic 7 concerns family, education, and life-course processes (*children, marriage, college*); Topic 8 addresses gender inequalities in work and careers (*women, gender, housework*); Topic 9 reflects economic crises and labor-market conditions (*unemployment, depression, relief*); Topic 10 focuses on social mobility and peer networks (*mobility, peer, homophily*); Topic 11 relates to the sociology of science and the development of the discipline (*science, psychoanalysis, research*); Topic 12 represents quantitative and statistical methods (*model, statistical, methods*); Topic 13 concerns classical sociological theory (*Weber, Durkheim, Parsons*); Topic 14 addresses immigration and ethnic relations (*immigrants, assimilation, intermarriage*); and Topic 15 captures political sociology (*political, party, ideology*).

Once the main thematic areas represented by the fifteen topics have been identified and interpreted, the next step is to examine how their prevalence changes over time. Topic prevalence was modeled as a function of editor and publication year to estimate both temporal trends and differences across editorial periods. Rather than presenting the trajectories of all topics, this discussion focuses only on those featuring clear-cut increasing or decreasing trends. We then compare these temporal trends with the estimated effects of the editor covariate to assess whether periods of thematic growth or decline correspond to specific editorial phases. This allows us to explore whether shifts in the thematic prominence correspond not only to broader social transformations but also to the influence of editors-in-chief as potential “gatekeepers of science” shaping the journal’s intellectual agenda.

Some topics display clear long-term decline (Figure 2, top). Topic 2, which captures discussions of war and collective morale, is highly prevalent in the early decades of the journal and decreases sharply after the 1930s, remaining at low levels in later periods. This pattern is consistent with the historical context of the interwar and World War II years and with the higher expected prevalence of the topic under early editors such as Faris and Burgess, followed by lower values in later periods of the journal, with only moderate differences across editors (Figure 3). A similar, though more gradual, decline can be observed for Topic 4 (urban and community dynamics), which is relatively more prevalent under editors associated with the Chicago tradition, including Burgess, Blumer, and Hughes, and becomes less prominent in later decades. Other topics show peaks that are more clearly associated with specific editorial periods (Figure 2, bottom).

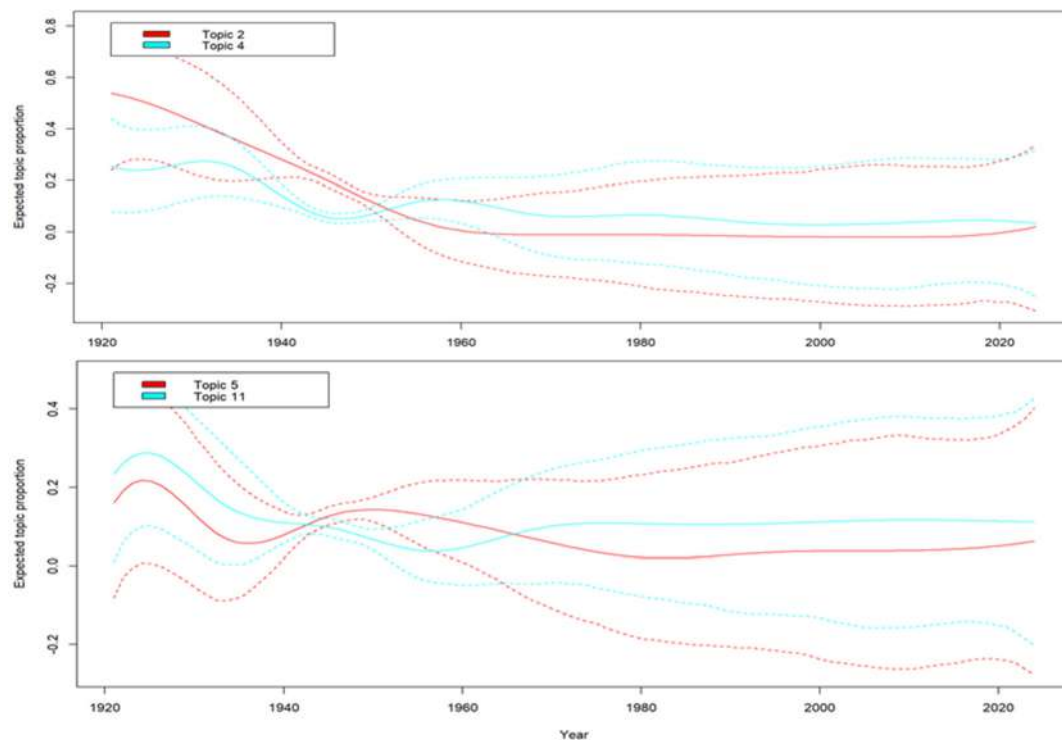


Fig. 2. Temporal trajectories of selected topics.

Topic 5 (conceptual and social-psychological foundations of sociology) displays its highest expected prevalence under Burgess, editor in the late 1930s, a period characterized by strong interest in social psychology, collective behavior, and the analysis of group processes. Likewise, Topic 11 (sociology of science and the development of the discipline) reaches higher levels under early editorial leadership, particularly under Small, and remains comparatively lower in subsequent periods, suggesting a stronger emphasis on disciplinary reflection in the formative phase of the journal. Not all changes, however, correspond closely to editorial transitions. Topic 3 (racial inequality and segregation) shows a marked increase from the 1970s onward, continuing to grow in recent decades (Figure 4), but differences across editorial periods are comparatively weak and characterized by wide confidence intervals. In this case, the temporal trajectory appears to reflect broader transformations in the discipline and in the social context rather than the influence of specific editors.

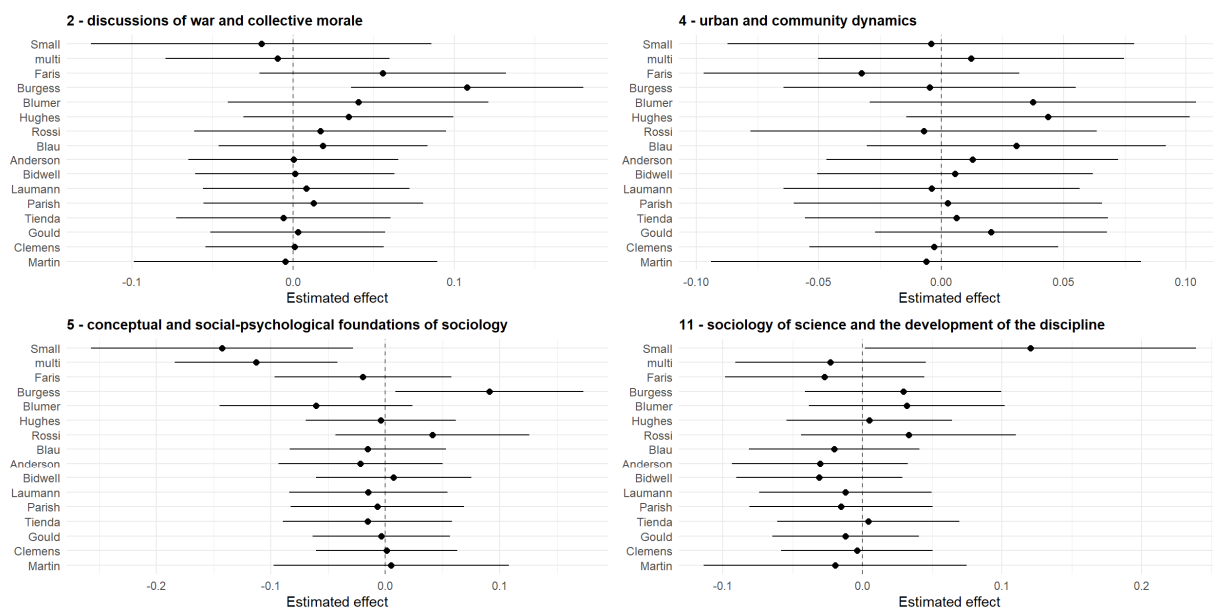


Fig. 3. Estimated effects of editors-in-chief on the expected prevalence of selected topics. Note. In some years editorial responsibilities overlapped (not represented in the figure).

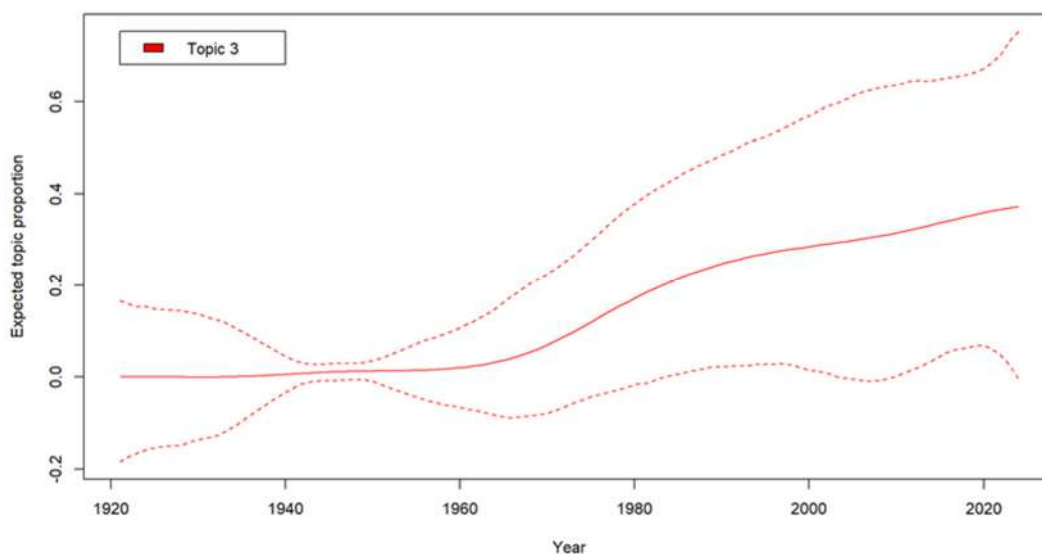


Fig. 4. Temporal trajectory of topic 3.

4. Discussion

The topic modeling analysis reveals a structured and evolving thematic landscape within the AJS over more than a century of publication. The results highlight both enduring core areas of sociological research (such as racial inequality and segregation, urban and community dynamics, organizational structures and power relations, family and life-course processes, gender and labor-market inequalities, immigration and ethnic relations, and political and ideological conflict) as well as methodological, theoretical, and epistemological themes related to quantitative methods, classical sociological theory, and the development of sociology as a scientific discipline. These patterns confirm that the journal has historically served as a venue where foundational sociological concerns coexist with emerging research agendas. The analysis

of topic prevalence over time further shows that these themes do not evolve uniformly. Some topics remain relatively stable across decades, while others display clear increasing or declining trajectories, reflecting the changing priorities of sociological research and the shifting social contexts that shape them. At the same time, the probabilistic nature of STM requires caution in the interpretation of topic prevalence and temporal trajectories. Topic assignments are inherently characterized by a degree of uncertainty, and some documents may display substantial overlap across multiple thematic dimensions. Similarly, estimated temporal trends should not be interpreted as deterministic patterns but rather as statistically modeled tendencies influenced by both corpus composition and model specification. For this reason, the interpretation focused primarily on topics displaying consistent and substantively meaningful trends across the selected model solutions and diagnostic evaluations.

Examining topic prevalence across editorial periods reveals systematic variations rather than abrupt thematic breaks. These patterns suggest that editorial leadership subtly reshapes the journal's intellectual emphasis over time. Editors-in-chief do not simply reproduce existing disciplinary agendas but participate in structuring the circulation of ideas by privileging certain themes over others. In this sense, they function as institutional "gatekeepers of science" (Crane, 1967), whose selection decisions contribute to defining which topics gain prominence within the discipline at specific historical moments. Overall, the results highlight the role of editorial governance in shaping the evolution of sociological agendas while also reflecting broader transformations in social reality and sociological inquiry. Moreover, this and previous studies (Tuzzi, 2018; Girodan et al., 2018; Sbalchiero and Eder, 2020; Rizzoli et al., 2023) demonstrate that the analysis of scientific literature through text data analysis methods represents a valuable approach for investigating the history of disciplines and the evolution of ideas. Within this methodological landscape, Structural Topic Modeling (STM) proves to be a particularly effective tool for capturing long-term intellectual dynamics in academic publishing and for tracing the institutional and epistemic evolution of scientific disciplines. More broadly, the analysis contributes to the sociology of scientific knowledge by showing how disciplinary agendas evolve through the interaction between broader social change and the institutional dynamics of academic publishing, with scientific journals acting as key arenas where intellectual priorities are selected, legitimized, and transformed over time.

Acknowledgements

This study was carried out within the framework of the project SBAL_BIRD24_01, University of Padova (PI Stefano Sbalchiero).

References

- Aria M., Cuccurullo C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4): 959–975.
- Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Carocci, Rome.
- Calhoun G. (2007). *Sociology in America: A history*. Chicago: University of Chicago Press.
- Clemens E. S., Powell W. W., McIlwaine K., Okamoto D. (1995). Careers in print: Books, journals, and scholarly reputations. *American Journal of Sociology*, 101(2): 433–494.
- Cohen D. J., Frisch M., Gallagher P., Mintz S., Sword K., Taylor A. M., Turkel W. J. (2008). Interchange: The promise of digital history. *The Journal of American History*, 95 (2): 452-491.
- Collins R. (1986). Is 1980s sociology in the doldrums? *American Journal of Sociology*, 91(6): 1336-1355.

- Connell R. W. (1997). Why is classical theory classical? *American Journal of Sociology*, 102(6): 1511-1557.
- Crane D. (1967). The Gatekeepers of Science: Some Factors Affecting the Selection of Articles for Scientific Journals. *The American Sociologist*, 2 (4): 195-201.
- Giordan G., Saint-Blancat C., Sbalchiero S. (2018), Exploring the History of American Sociology through Topic Modeling. In Tuzzi ,editor, A, *Tracing the Life-Course of Ideas in the Humanities and Social Sciences*. Cham, Springer, pp. 45-64.
- Kalekin-Fishman D., Denis A. (2012). *The shape of sociology for the 21st century: Tradition and renewal*. London: SAGE.
- Livingstone D. N. (1995). The spaces of knowledge: contributions towards a historical geography of science. *Environment and Planning D: Society and Space*, 13: 5-34.
- Lebart L., Salem A., Berry L. (1998). *Exploring textual data*. Kluwer Academic Publishers: Dordrecht.
- Pavone P. (2010). Sintagmazione del testo: una scelta per disambiguare la terminologia e ridurre le variabili di un'analisi del contenuto di un corpus. In Bolasco S., Chiari I, Giuliano L. (Eds.) *Statistical Analysis of Textual Data: Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, LED, pp. 131-140.
- Rizzoli V., Trevisani M. and Tuzzi A. (2023). Portraying the Life Cycle of Ideas in Social Psychology through Functional (Textual) Data Analysis: A Toolkit for Digital History. *Scientometrics*, 128: 5197–5226.
- Roberts M. E., Stewart B. M., Tingley D. (2019). Stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2): 1-40.
- Sbalchiero S., Eder M., (2020). Topic modeling, long texts and best number of topics: some problems and solutions. *Quality & Quantity*, 54 (4): 1095-1108.
- Sbalchiero S., Tuzzi A. (2018). What's Old and New? Discovering Topics in the American Journal of Sociology. In Iezzi D. F., Celdardo L., Misuraca M. (Eds.), *JADT '18 Proceedings of 14th International Conference on statistical analysis of textual data*, UniversItalia Editore, pp. 724-732.
- Sciandra A., Trevisani M., Tuzzi, A. (2025). Temporal trends and presidential traits in the Italian end-of-year addresses: comparing and contrasting kbs and stm results. *Quality & Quantity*, 59(1): 905–935.
- Scott A., Desfor Edles A. (2011). *Sociological theory in the contemporary era: Text and readings*. Thousand Oaks, Pine Forge Press.
- Shanas E. (1945). The American Journal of Sociology through fifty years. *American Journal of Sociology*, 50(6): 522–533.
- Shapin S. (1995). Here and everywhere: sociology of scientific knowledge. *Annual Review of Sociology*, 21: 289-321.
- Turner S. (1998). Who's afraid of the history of sociology? *Swiss Journal of Sociology*, 24: 3-10.
- Tuzzi A. (2018). *Tracing the Life-Course of Ideas in the Humanities and Social Sciences*. Cham, Springer.
- Tuzzi A. (2024). *Fondamenti di analisi dei dati testuali*. Carocci, Roma.

Titles, Abstracts, or Full-Texts? A Comparative Study of LDA and BERTopic Performance across Text Levels

Francis Lareau^{1,2,3}, Christophe Malaterre^{2,3}

¹Université de Sherbrooke – francis.lareau@USherbrooke.ca

²Université du Québec à Montréal – malaterre.christophe@uqam.ca

³Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST)

Abstract

Topic modeling has become an essential tool in the History, Philosophy, and Sociology of Science (HPSS), but researchers face a practical question: when exploring corpora of scientific articles, is it sufficient to analyze titles or abstracts, or is full-text analysis necessary? This question is critical given the significant resources required to obtain, preprocess, and analyze full-text corpora. We address this by comparing two prominent topic modeling approaches—Latent Dirichlet Allocation (LDA) and BERTopic—applied to titles, abstracts, and full texts from an astrobiology corpus of 3,869 scientific articles. Our analysis includes quantitative metrics such as Adjusted Rand Index, coherence CV, topic diversity, micro inner and micro joint recall, Jaccard top-words similarity. Results show that title-based models perform poorly across most metrics, while abstract-based models achieve a better balance. Lexical analysis reveals LDA models exhibit stability in their top-word vocabularies across models, while analysis of document assignment to topics reveals greater consistency among BERTopic models, reflecting different notions of topic similarity. BERTopic on abstracts emerged as the best overall performer, showing superior coherence and the strongest cross-method top-words alignment, though outperformed on document coverage by LDA on abstracts. Full-text models, while offering high recall, can suffer from overly broad (LDA) or unbalanced (BERTopic) topics. We recommend abstract-based topic modeling for most bird’s-eye HPSS applications, reserving full-text analysis for specialized investigations requiring high granularity.

Keywords: topic modeling, LDA, BERTopic, text mining, digital humanities, corpus analysis.

1. Introduction

Topic modeling has emerged as a crucial computational method for analyzing large volumes of research articles, particularly in the History, Philosophy, and Sociology of Science (HPSS). Researchers employ topic modeling for diverse purposes: identifying research trends (Griffiths & Steyvers, 2004), uncovering thematic substructures and interrelations (Blei & Lafferty, 2007), analyzing the evolution of scientific vocabulary (Chavalarias & Cointet, 2013), revealing hidden biases in scientific discourse (Sugimoto et al., 2013), studying the sociology of scientific communities (Gerow et al., 2018), examining interdisciplinary patterns (Hyeyoung et al., 2022), systematic reviewing (Galli et al., 2024) and enhancing historiography of science (Mimno, 2012) and art (Browman, 2023). Studies such as these apply topic modeling to different textual levels (titles, abstracts, or full-text) without comparative evaluation.

A critical methodological question remains largely unaddressed: should topic modeling be applied to article titles, abstracts, or full texts? This question has significant practical implications, as obtaining and preprocessing full-text corpora requires substantial computational resources, institutional access, and data management infrastructure. Many researchers face constraints in accessing full-text articles due to paywalls, copyright restrictions, or limited computational capacity.

While titles offer maximum accessibility and minimal preprocessing requirements, they are short and may lack sufficient information for robust topic identification. Longer than titles, and yet often available without paywall, abstracts provide structured summaries that capture the main contributions, offering more information. As for full texts, many of which are still behind paywalls, they contain the most comprehensive information, including methodological sections, extensive citations, and discussions, though this information may also introduce noise.

Some studies have begun to address this text-level question. Syed and Spruit (2017) compared LDA performance on abstracts versus full texts across scientific corpora, finding that abstracts contain approximately 30% less noise and produce more coherent topics, particularly in smaller collections. Riehl (2025) extended this comparison to include titles alongside abstracts and full texts in business and economics literature, concluding that abstracts and full texts can be used interchangeably for most purposes. However, these studies focused exclusively on LDA and did not examine modern neural topic modeling approaches. To our knowledge, no prior work has systematically compared classical statistical methods (e.g., LDA) and contemporary neural approaches (e.g., BERTopic) across all three text levels using a diverse set of metrics.

This study addresses this gap by systematically comparing two prominent topic modeling approaches—Latent Dirichlet Allocation (LDA) and BERTopic—across titles, abstracts, full texts levels using a well-characterized astrobiology corpus (Atanassova et al., 2025; Malaterre & Lareau, 2023). We employ quantitative metrics to evaluate model performance, providing evidence-based recommendations for researchers in HPSS and related fields.

2. Methodology

2.1. Topic modeling approaches

Both LDA and BERTopic share fundamental assumptions: (1) documents can be represented as numerical vectors, (2) topics are identifiable through linguistic regularities, and (3) machine learning can automatically detect these patterns. However, they differ substantially in implementation and underlying philosophy. LDA (Blei et al., 2003; Griffiths & Steyvers, 2004) is a classical statistical method where topics correspond to latent variables whose multinomial distribution follows Dirichlet’s law. The model requires as input a Bag-of-Words representation of the corpus—a term-document matrix—and to choose the number of topics, K . Training involves iteratively: (1) assigning topics to words, (2) updating topic probabilities in documents, and (3) updating word probabilities in topics. The process generates two probability matrices for document-topic and topic-word relationships. A topic is thus characterized by a probability distribution over words, and a document by a probability distribution over topics. LDA handles long texts effectively, making it suitable for full-text analysis.

BERTopic (Grootendorst, 2022) is a more recent, modular approach in which topics correspond to the topological document densities in the embedding space. Here, documents are represented by their embeddings obtained with neural methods. The approach comprises a first step during which documents are clustered, and a second which consists of identifying top words for document clusters so as to infer the topic of each cluster. The clustering step includes: (1) Transformer-based document representation, producing high-dimensionality embeddings as outputs, (2) embedding dimensionality reduction (UMAP), and (3) embedding clustering (HDBSCAN). For the second step, the initial documents are first represented with a classical bag-of-words approach, and top-words are extracted per cluster (with c-TF-IDF). Traditionally limited to short texts due to BERT’s limited context window (<512 tokens), the BERTopic pipeline can now benefit from recent embedding models that enable long-document processing.

2.2. Corpus and experimental protocol

We utilized an astrobiology corpus of 3,917 articles previously analyzed in depth (Malaterre & Lareau, 2023). From this corpus, 48 documents were removed for lack of abstract, resulting in a working corpus of 3,869 documents, each with a title, an abstract, and a full text. Every text level was lemmatized using TreeTagger, retaining only content-bearing parts of speech (nouns, verbs, adjectives, adverbs, modals). Lemmas shorter than three characters and tokens containing punctuation artifacts were removed. Each document was represented as a list of lemmas used as input for LDA and for BERTopic's c-TF-IDF top-word extraction, while BERTopic embeddings were computed on the raw text.

To carry out the embeddings needed for BERTopic, we employed `stella_en_1.5B_v5` on the raw text, which handles up to 131,072 tokens and achieved competitive performance on the MTEB benchmark (Muennighoff et al., 2022). This model, based on Alibaba-NLP (`gte-large-en-v1.5` and `gte-Qwen2-1.5B-instruct`), was fine-tuned on retrieval, semantic textual similarity, and Matryoshka Representation Learning tasks (Kusupati et al., 2022). To our knowledge, this is the first application of this embedding method to topic modeling in HPSS. BERTopic was then fitted to the corpus documents at each level of textual granularity (titles, abstracts, and full texts) in order to obtain topic solutions ranging from approximately 5 to 55 topics, with increments of roughly five. This range was achieved by adjusting the HDBSCAN's minimum cluster size parameter, which controls the granularity of the clustering and, consequently, the number of topics inferred. The exact number of topics K obtained for each structure was then used to fit corresponding LDA models, ensuring that both approaches were evaluated under identical topic dimensionality for each level of analysis. Outlier documents initially labeled as noise were reassigned using HDBSCAN soft clustering, based on their probabilistic membership.

We then implemented LDA in Python using the `lda` package with Gibbs sampling. We trained each model for 1000 iterations with $\alpha=5/K$, $\eta=0.01$, and a fixed random seed for replicability. In addition, for comparison with the other models (notably with BERTopic for which each document is assigned to only one topic), documents were crisp assigned to their dominant topic. In addition, the LDA full-text model with 25 topics was analyzed in depth so as to serve as comparative reference: each topic was manually analyzed using representative words and documents, then labeled with descriptive keywords; topic correlations were computed from document distributions, and community detection identified four larger thematic clusters.

2.3. Quantitative evaluation metrics

We comparatively evaluated the six models along two main dimensions. First, we investigated how the models performed in terms of top-word assignments to topics. This was done with three complementary metrics: a) Coherence CV (Röder et al., 2015) measures the semantic coherence of topics by computing the average cosine similarity between context vectors of top words, where context vectors are derived from word co-occurrence patterns within a sliding window (110 tokens usually) across the lemmatized corpus (range: 0 to 1, higher indicate that a topic's top words appear in similar local contexts); b) Topic Diversity measures the proportion of distinct words among top words across all topics (range: 0 to 1, higher indicates less redundancy); c) Jaccard similarities between each model's set of topics' top-10 words: provides a measure of lexical thematic convergence (range: 0 to 1, higher indicates more overlap).

Second, we examined how differently the models paired topics and documents. To this aim, we used another set of three measures that we applied to the $K = 25$ models: a) Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) measures similarity between clustering solutions,

corrected for chance (range: -1 to 1, where 0 indicates random clustering); b) Micro Inner Recall measures the average proportion of documents in each topic containing individual top words (range: 0 to 1, higher indicates more recall, see Lamirel et al., 2025); c) Micro Joint Recall measures the average proportion of documents in each topic containing any of the topic’s top words (range: 0 to 1, higher indicates more recall, see Lamirel et al., 2025).

3. Results

3.1. Evaluation of topic top-words

The objective of topic top-words evaluations is to assess how well the top-words of each model describe the topics, and do so in a way that clearly separates the topics from one another. Coherence measures provide a perspective on how the top-words fit together. Results show marked differences, especially across text levels (Fig. 1A). Title-based models consistently exhibit the worst coherence (0.38-0.48), indicating semantically weaker topic-word associations. Abstract-based models outperform full-text counterparts, across all K values, or are on par with them. BERTopic demonstrates superior coherence to LDA, though the advantage diminishes as K increases. BERTopic-Abstract achieves optimal coherence (0.67-0.73), suggesting it produces the most semantically meaningful topics.

Diversity measures reveal how distinct the top-words are across topics (Fig. 1B). In this respect, title-based models achieve higher diversity (0.72-0.93) compared to their abstract and full-text counterparts, suggesting topics inferred from titles use more distinct vocabulary—though this may reflect poor topic definition rather than strength. BERTopic outperforms LDA in diversity, except at low K values. BERTopic-Title and BERTopic-Full-text exhibit overall the highest diversity measures. BERT-Abstract lags behind.

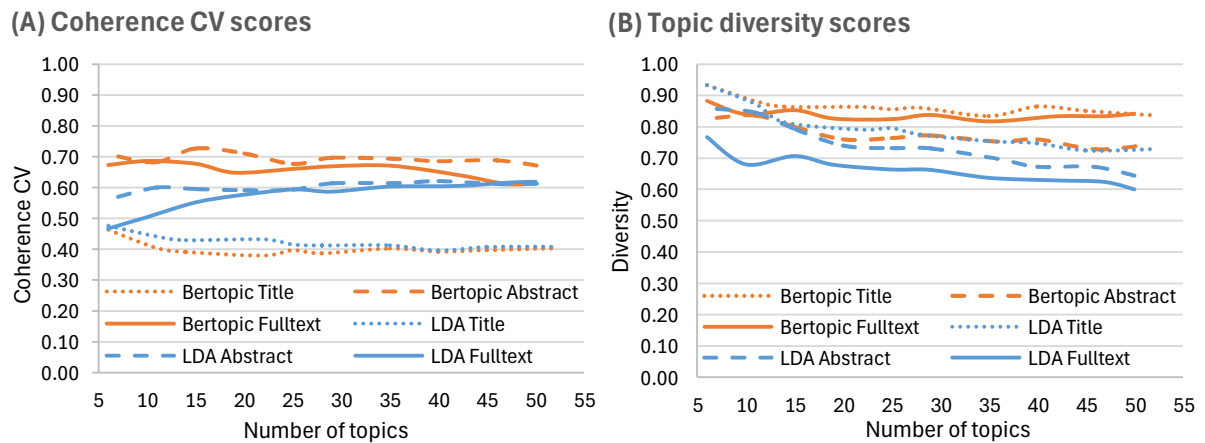


Figure 1: (A) Coherence CV scores by model as a function of the number of topics. Higher values indicate more semantically coherent topics (top-10 words). (B) Topic diversity scores by model as a function of the number of topics. Higher values indicate topics use more distinct terms (top-10 words).

To provide a perspective on top-word stability across models, we measured vocabulary overlap through pairwise Jaccard similarities between top-10 word sets across all six models at K=25 (Fig. 2A). The matrix reveals systematic patterns in lexical consistency. LDA models demonstrate higher internal coherence, with LDA-Full-text and LDA-Abstract sharing nearly half their vocabularies (0.48), while LDA-Abstract and LDA-Title maintain substantial overlap (0.44). In contrast, BERTopic models exhibit lower inter-model similarities (0.29-0.34), suggesting stronger variations in vocabulary selection. Cross-method comparisons show moderate overlap (0.20–0.39), with BERTopic-Abstract achieving the strongest alignment with

LDA models (0.36–0.39). This convergence is reflected in individual topics: for instance, the “Radiation-Spore” topic is highly stable across all six models (top words: radiation, spore, space, survival, exposure, experiment, condition, bacillus, simulated, cell), while the “Chemistry” topic fragments into narrower sub-topics in title-level models, illustrating how textual reduction can destabilize thematically broad topics.

(A) Jaccard similarities between sets of top-words

		LDA			BERTopic		
		Title	Abstract	Full-text	Title	Abstract	Full-text
LDA	Title		0,44	0,30	0,32	0,31	0,30
	Abstract	0,44		0,48	0,26	0,39	0,26
	Full-text	0,30	0,48		0,20	0,36	0,21
BERTopic	Title	0,32	0,26	0,20		0,29	0,34
	Abstract	0,31	0,39	0,36	0,29		0,31
	Full-text	0,30	0,26	0,21	0,34	0,31	

(B) ARI scores between clustering solutions

		LDA			BERTopic		
		Title	Abstract	Full-text	Title	Abstract	Full-text
LDA	Title		0,15	0,13	0,18	0,15	0,13
	Abstract	0,15		0,30	0,25	0,30	0,22
	Full-text	0,13	0,30		0,24	0,31	0,21
BERTopic	Title	0,18	0,25	0,24		0,38	0,38
	Abstract	0,15	0,30	0,31	0,38		0,36
	Full-text	0,13	0,22	0,21	0,38	0,36	

Figure 2: (A) Pairwise Jaccard similarities between the sets of top-10 words. Each cell represents the proportion of shared vocabulary relative to the union of the two corresponding top-word sets. (B) Adjusted Rand Index between all topic models. Values closer to 1 indicate higher similarity; values near 0 reflect chance-level agreement, with negative values indicating systematic disagreement.

3.2. Evaluation of document-topics pairing

The second set of measures we deployed aimed at assessing how the different models pair documents and topics together. A first simple statistic is to compare the spread of topic sizes in terms of documents (Fig. 3). In this respect, LDA models tend to exhibit a narrower distribution of topic sizes compared to BERTopic (for the targeted number of topics of 25). In particular, BERTopic-Full-Text appears to exhibit a surprisingly large topic (above 600), which is absent from the five other models. BERTopic-Abstract is more similar to LDA models in terms of document distribution, while BERTopic-Title displays the largest distribution.

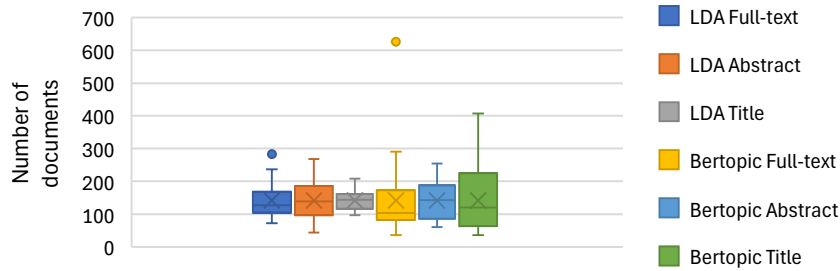
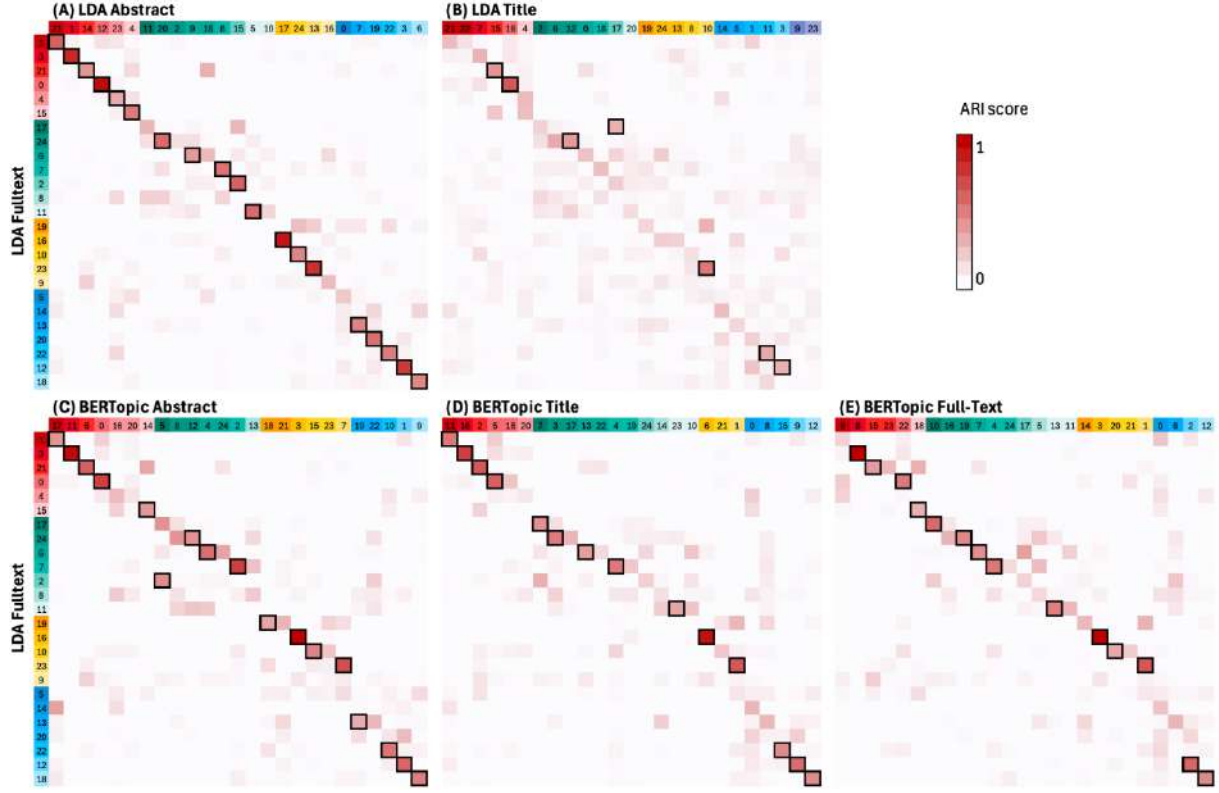
Model Comparison of Document Distribution Across Topics

Figure 3. Homogeneity of topic coverage across models at $K=25$: box-and-whisker summary.

To assess whether the models associated the same documents to the same topics, we calculated the Adjusted Rand Index between all six topic models at $K=25$, thus providing insights into inter-model similarity in terms of document clustering patterns (Fig. 2B). LDA-Title differs substantially from all other models ($ARI < 0.20$), suggesting poor clustering quality. Other models demonstrate moderate agreement ($ARI > 0.20$), with BERTopic models showing stronger inter-model consistency ($ARI > 0.35$). Notably, BERTopic-Abstract exhibits the highest average similarity to other models (excluding LDA-Title), suggesting it captures a central representation of document structure. A complementary analysis conducted at the topic level reveals which topics from one model best correspond to topics from another. This analysis relies on a one-against-all cluster evaluation, in which binary membership vectors (in the cluster

vs. all others) are compared for each cluster pair using the Adjusted Rand Index. Figure 4 shows the ARI values of five models at K=25 when compared to LDA-Full-text (used as reference).

Figure 4. Topic correspondence strength between LDA-Full-text and alternative models as measured by Adjusted Rand Index (ARI) scores: (A) LDA-Abstract, (B) LDA-Title, (C) BERTopic-Full-text, (D) BERTopic-Abstract, (E) BERTopic-Title. Shaded cells with borders indicate ARI scores >0.2 ; topic



colors based on LDA-Full-text model (purple when all ARI scores <0.05 , hence no clear pairing).

LDA-Abstract exhibits strong overall alignment with LDA-Full-text, with 19 topics exhibiting meaningful correspondence (ARI > 0.20), most of which maintain exclusive one-to-one relationships (Fig. 4A). However, 6 topic pairs represent splits or mergers between models (two or more ARI values > 0.15 per line or column) (e.g., LDA-Abstract topic 18 absorbing part of LDA-Full-text topics 6 and 21), while 6 other topics do not exhibit any strong pairing (all ARI values < 0.15) (e.g., LDA-Abstract topic 2, which is a small class with only 58 documents, suggesting potential over-fragmentation). The comparison of LDA-Title with LDA-Full-text (Fig. 4B) exhibits substantial reorganization with diminished alignment. Only 7 topic pairs reach the 0.20 ARI threshold. Numerous other topics appear to be splits and mergers. This fragmentation indicates titles provide insufficient information for stable topic identification.

BERTopic models show variable performance when compared to LDA-Full-text (Fig. 4C-E). BERTopic-Full-text achieves moderate alignment with 14 topics exhibiting meaningful correspondence (ARI > 0.20). Note that a severe class imbalance arises, with 1 large cluster and 4 sparse topics (< 50 documents). BERTopic-Abstract performs better, yielding 17 topics with meaningful correspondence, and exhibiting superior cluster balance. BERTopic-Title produces only 14 topics with meaningful correspondences, and also suffers from distributional problems, with 1 large cluster (>400 documents) and 3 undersized ones (<40 documents).

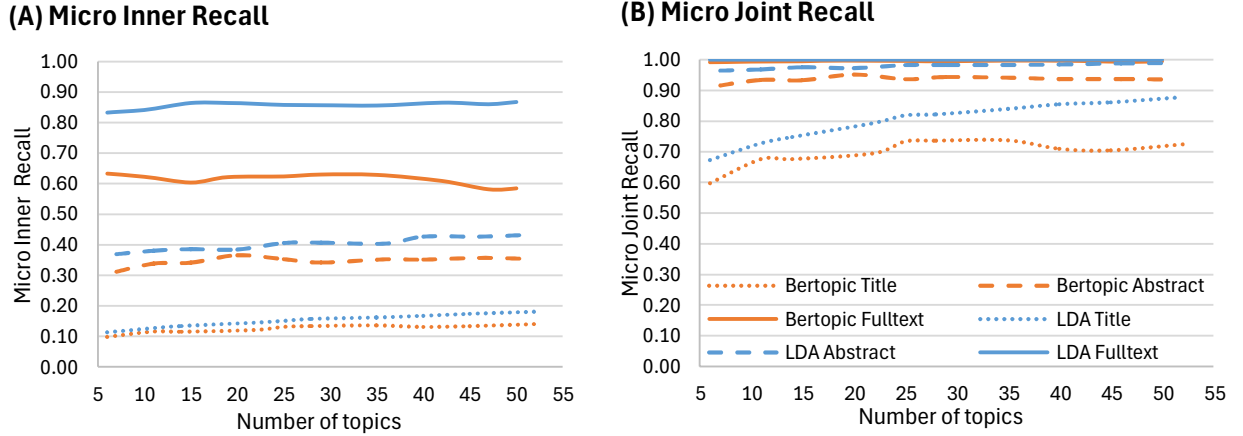


Figure 5: (A) Micro Inner Recall by model and number of topics. Measures the proportion of topic documents containing individual top words. (B) Micro Joint Recall by model and number of topics. Measures the proportion of topic documents containing any top word.

The last facet of our evaluation pipeline investigates how well the different models cover the entirety of the corpus. Micro inner recall measures how well individual top words appear in topic documents (Fig 5A), while micro joint recall assesses whether any top word appears in topic documents (Fig. 5B). Title models exhibit the worst recall performance independently of methods. This means that their topics only represent a small fraction of the corpus. Full-text models substantially outperform their abstract and title counterparts across both metrics. LDA demonstrates superior recall compared to BERTopic across all text levels. LDA Full-text achieves near-perfect micro joint recall (1.00) and excellent micro inner recall (0.83-0.87), indicating that its top words reliably appear in assigned documents.

4. Discussion

Our systematic comparison reveals strengths and weaknesses of different text levels and methods for topic modeling. A clear result is that title-based models consistently underperform across multiple dimensions. LDA-Title shows poor inter-model agreement ($ARI < 0.20$), low coherence, and minimal recall, suggesting titles lack sufficient information for robust topic identification. While BERTopic-Title achieves high diversity, the lowest coherence score indicates inconsistent rather than rich topic vocabulary. The poor performance likely stems from title brevity and optimization for impact rather than comprehensive content description.

Full-text models reveal different challenges for each algorithm. LDA-Full-text exhibits the highest micro inner recall (>0.8), thus indicating that its top words reliably appear in documents. Yet, it shows moderate coherence (0.59-0.62) and the lowest diversity of all models, suggesting these words may not form tight semantic clusters. BERTopic-Full-text achieves better coherence (0.61-0.69) and high diversity (>0.8) but suffers from class-size imbalance, with some topics capturing narrow concepts while leaving many documents in oversized residual clusters. Its micro inner recall performance is also lower compared to LDA-Full-text.

Abstract-based models demonstrate good balance and convergence across methods. BERTopic-Abstract achieves the highest coherence (0.67-0.73) alongside high diversity (>0.6), indicating semantically tight but diverse topics. LDA-Abstract performs similarly, though with slightly lower performance. Yet both models exhibit lower recall compared to full-text models (micro inner recall at about 0.4), but still higher than title-based models. LDA-Abstract and BERTopic-Abstract show strong consistency with each other and with LDA-Full-text ($ARI = 0.30$ -0.31).

The contrasting performance patterns suggest significant trade-offs between text levels. Notably, abstract-level models tend to display higher coherence (semantic tightness of topic words) while full-text-level models show higher recall (comprehensive document coverage). The relatively high vocabulary overlap between LDA-Full-text and LDA-Abstract (Jaccard: 0.48) indicates shared core terminology despite differences in text levels, this overlap decreasing with LDA-Title. On the other hand, vocabulary overlap between BERTopic models is generally lower. This suggests a higher sensitivity of BERTopic to text levels for extracting top-words (likely stemming from excessive information in full texts, including methodological sections, extensive citations, and peripheral discussions obscuring core thematic content), compared to LDA. The good alignment of BERTopic-Abstract to LDA-Full-text and LDA-Abstract (Jaccard: 0.36-0.39) suggests convergence between the three models.

Researchers must therefore select models based on research objectives and depending also on data access: exploratory studies prioritizing topic discovery may favor high-coherence models like BERTopic-Abstract (which is usually easy, abstracts being largely available), while comprehensive document classification tasks may require high-recall approaches like LDA-Full-text (provided full-text is accessible) despite looser topic definitions. Another consideration is computational time, which increases with textual amount: LDA processing time on our 3,869-article corpus was 6 sec, 1.5 min, and 59 min respectively for titles, abstracts and full-texts, while BERTopic took 5.5 min, 9 min, and 45 min (titles, abstracts, full-texts), in addition to generating embeddings via large transformer models.

Several limitations of this study point to avenues for future work. First, all results are derived from a single corpus—a moderately sized, English-dominated, interdisciplinary field—which constrains generalizability. Broader, multilingual, or corpora mixing genres (e.g., policy reports with journal articles), may yield different relative performances; in particular, larger collections may reduce the coherence gap between text levels (Syed & Spruit, 2017). Replicating the analyses across corpora would be essential to assess the robustness of the observed patterns, particularly the relative performance of text levels and modeling approaches. Second, our analyses of inter-model overlap—both in terms of top-word similarity and document assignments—were conducted for a single value of K ($K = 25$). While this choice was motivated by a balance between granularity and interpretability, extending these comparisons across a wider range of topic numbers would provide a more systematic understanding of model stability and convergence across granularities. Third, the evaluation framework remains primarily quantitative. Complementing it with qualitative analyses would significantly enrich interpretation, including close inspection of topic interpretability based on top-word sets, identification and characterization of documents most central or representative of each topic, and a more detailed examination of class imbalance—particularly in BERTopic models.

5. Conclusion

This study provides empirical guidance on a central methodological choice in topic modeling: the selection of text level. By systematically comparing LDA and BERTopic across titles, abstracts, and full texts, we identify clear trade-offs that bear directly on research design. Title-based modeling consistently underperforms—yielding unstable, low-coherence, low-recall topics—and should be avoided for substantive analyses. Full-text models offer the highest coverage and granularity, but at the cost of reduced coherence and, for BERTopic, pronounced class imbalance. Abstract-based models emerge as the most balanced option: they avoid the sparsity of titles and the noise of full texts, show greater cross-method consistency, and produce semantically tighter topics. In practice, we recommend: (1) avoiding titles except for preliminary exploration; (2) using abstracts as a default, with BERTopic-Abstract as the

preferred choice; and (3) reserving full texts for studies requiring high documentary coverage, preferring LDA over BERTopic at that level. More broadly, the marked variation in performance across text levels underscores the need for methodological transparency: researchers should explicitly justify their choice of textual input, and, when feasible, triangulate across levels or methods to strengthen confidence in the results.

Acknowledgements

C.M. acknowledges funding from Canada SSHRC (Grant 430-2018-00899) and Canada Research Chairs (CRC-950-230795). F.L. acknowledges funding from Canada SSHRC (Postdoctoral Fellowship 756-2024-0557). The authors thank the audience of the 2025 “LLMs in HPSS” workshop at TU Berlin where an earlier partial version of this study was presented.

References

- Atanassova, I., Ningrum, P. K., Gutehrlé, N., Lareau, F., & Malaterre, C. (2025). Is There Life on Mars? Studying the Context of Uncertainty in Astrobiology. *Proceedings of ISSI 2025*.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *Ann. Appl. Stat.*, 1(1), 17-35.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993-1022.
- Browman, D. (2023). Topic modeling art historical discourse. *Digit. Scholarsh. Humanit.*, 38(2), 567-584.
- Chavalarias, D., & Cointet, J. P. (2013). Phylomemetic patterns in science evolution—The rise and fall of scientific fields. *PLoS ONE*, 8(2), e54847.
- Galli, C., Cusano, C., Meleti, M., Donos, N., & Calciolari, E. (2024). Topic Modeling for Faster Literature Screening Using Transformer-Based Embeddings. *Metrics*, 1(1), 2.
- Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proc. Natl. Acad. Sci.*, 115(13), 3308-3313.
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci.*, 101(suppl 1), 5228-5235.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794*.
- Hyeyoung, K., Daehwan, A., & Meeyoung, L. (2022). Analyzing interdisciplinarity in biomedicine using topic modeling. *Scientometrics*, 127(5), 2655-2678.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *J. Classif.*, 2(1), 193-218.
- Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P. & Farhadi, A. (2022). Matryoshka representation learning. *Adv. Neural Inf. Process. Syst.*, 35, 30233-30249.
- Lamirel, J.-C., Lareau, F., & Malaterre, C. (2025). Old but Not Obsolete: Bag-of-Words vs. Embeddings in Topic Modeling. *Proceedings of ISSI 2025*.
- Malaterre, C., & Lareau, F. (2023). The Emergence of Astrobiology: A Topic-Modeling Perspective. *Astrobiology*, 23(5), 496-512.
- Mimno, D. (2012). Computational historiography: Data mining in a century of classics journals. *J. Comput. Cult. Herit.*, 5(1), 1-19.
- Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. *arXiv:2210.07316*.

- Riehl, K., Neumann, T., & Ahrweiler, P. (2025). On the scientometric value of full-text, beyond abstracts and titles: Evidence from the business and economics literature. *Manag. Rev. Q.*, 75, 2459-2513.
- Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2013). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. *J. Am. Soc. Inf. Sci. Technol.*, 62(1), 185-204.
- Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. *IEEE DSAA*, 165-174.

A Multi-Method Validation Framework for Large-Scale Multilingual Text Analytics

Stefano Blando^{1,2}, Domenica Fioredistella Iezzi³

¹Sant’Anna School of Advanced Studies — stefano.blando@santannapisa.it

²University of Pisa, Department of Computer Science — stefano.blando@phd.unipi.it

³University of Rome Tor Vergata — stella.iezzi@uniroma2.it

Abstract

When can researchers trust their text analytics results? Despite the proliferation of computational methods, from classical lexicometric approaches to transformer-based deep learning, the field lacks systematic criteria for distinguishing genuine findings from methodological artifacts. This paper addresses this gap by proposing a validation framework centered on the concept of *method-invariant patterns*: findings that emerge consistently across analytically independent approaches, operationalized as patterns where between-method variance remains below 5% of total variance.

The framework integrates two complementary validation levels. *Intra-method validation* quantifies reproducibility under data perturbations through K-fold cross-validation (stability coefficients: 0.991 to 0.997) and Procrustes analysis (configuration similarity: 0.969 to 0.991). *Inter-method validation* assesses convergence across eighteen techniques spanning four analytical dimensions: semantic structures (LDA, LSA, NMF, neural embeddings), sentiment polarity (lexicon-based and transformer approaches), document partitions (five clustering algorithms), and relational patterns (co-occurrence networks, Correspondence Analysis).

We validate this framework on 999,152 multilingual reviews across five languages, representing one of the largest systematic multi-method comparisons in text analytics. The central finding is methodologically consequential: robust patterns emerge consistently regardless of analytical approach. Variance decomposition reveals that substantive content (rating valence) accounts for 95.4% of total variance, while methodological choice explains less than 3%. Cross-platform replication (Python versus R; Pearson $r = 0.932$) further confirms that validated patterns transcend implementation differences.

The framework also provides guidance for method selection. While BERT achieves highest accuracy (91.3%), classical approaches such as SVM (89.1%) and Logistic Regression (88.7%) offer comparable performance at substantially lower computational cost (29-fold reduction in inference time). These findings demonstrate that multi-method validation offers researchers principled, operational criteria for identifying trustworthy results and making informed methodological decisions.

Keywords: multi-method validation, text analytics, convergence, reproducibility, multilingual NLP.

1. Introduction

The rapid expansion of computational text analytics has produced a rich methodological toolkit spanning classical lexicometric techniques (Benzécri, 1992; Lebart et al., 1998), probabilistic topic models (Blei et al., 2003), distributional semantics (Deerwester et al., 1990; Mikolov et al., 2013), and transformer-based deep learning (Devlin et al., 2019). Yet this very proliferation raises a fundamental question: when different methods yield different results, which findings should be trusted?

This problem is particularly acute in applied domains such as hospitality and tourism, where text analytics increasingly informs managerial decisions based on large volumes of user-generated

content amenable to computational analysis. A topic model may identify service-quality dimensions that a sentiment classifier fails to capture; a co-occurrence network may reveal relational structures invisible to dimensionality reduction. Without principled criteria for adjudicating between methods, researchers risk either cherry-picking favorable results or dismissing genuine findings as artifacts.

The existing literature offers limited guidance. Methodological comparisons typically focus on pairs of methods—LSA versus LDA (Blei et al., 2003; Deerwester et al., 1990), lexicon-based versus supervised sentiment (Hutto and Gilbert, 2014; Liu, 2012)—without systematic frameworks for assessing convergence across entire analytical pipelines. Cross-platform replication studies remain rare despite growing calls for computational reproducibility (Peng, 2011).

We propose a *multi-method validation framework* centered on the concept of *method-invariant patterns*—findings that emerge consistently across analytically independent approaches. The framework operationalizes robustness as patterns where between-method variance remains below 5% of total variance, and integrates two validation levels: *intra-method* stability (reproducibility under data perturbations) and *inter-method* convergence (agreement across different analytical families).

The framework is validated on 999,152 multilingual hotel reviews across five languages (2002–2024), systematically comparing eighteen techniques organized into four families. Section 2 describes the corpus and analytical pipeline. Section 3 presents results by method family. Section 4 reports the cross-method synthesis. Section 5 discusses implications.

2. Corpus, Methods, and Validation Protocol

2.1. Corpus

The corpus comprises 999,152 user-generated hotel reviews from a major online travel platform, covering 6,758 properties in Rome over 22 years (2002–2024). Five languages account for 91.6% of reviews (Table 1). The overall rating distribution is positively skewed (mean = 4.18, SD = 0.54): 77.2% of hotels are rated ≥ 4.0 . Service quality is the primary driver of overall satisfaction ($r = 0.924$), followed by Cleanliness ($r = 0.891$). A systematic data quality evaluation yields an overall score of 98.7% (Grade A), with 99.2% geospatial coverage.

Table 1: Corpus distribution by language.

Language	Reviews	%	Tokens
English	484,150	48.5	10,879,234
Italian	281,420	28.2	5,623,187
Spanish	64,857	6.5	1,298,456
French	61,388	6.1	1,024,321
German	22,729	2.3	487,234
Total	914,544	91.6	19,312,432

2.2. Preprocessing

Each language sub-corpus underwent independent preprocessing: tokenization and lemmatization via spaCy, stemming via Snowball (Porter, 2001), stopwords removal (NLTK standard lists plus 222 domain-specific and noise terms), and filtering of short tokens, numbers, and URLs.

The pipeline reduces the corpus from 95.6M to 20.7M tokens (78.4% reduction) while retaining 99.1% of information content. The frequency distribution follows Zipf’s law (Zipf, 1949) with $R^2 = 0.935$, confirming a well-formed corpus. Lexical diversity metrics (Covington and McFall, 2010) indicate moderate diversity (TTR = 0.0090), typical of review language. Cross-linguistic analysis reveals remarkable convergence: top-10 terms are semantically identical across languages (location/posizione/ubicación, room/camera/habitación).

2.3. Analytical Pipeline: 18 Methods, 4 Families

The framework integrates eighteen techniques organized into four analytically independent families:

Family 1: Semantic Representations (7 methods). TF-IDF weighted document-term matrices (Salton and Buckley, 1988) provide input for LSA via truncated SVD (Deerwester et al., 1990), LDA (Blei et al., 2003; Griffiths and Steyvers, 2004) with coherence optimization (Röder et al., 2015), and NMF. Neural embeddings include Word2Vec Skip-gram (Mikolov et al., 2013) (300d, window 5), GloVe (Pennington et al., 2014) (300d, window 10), and Sentence-BERT (all-MiniLM-L6-v2, 384d).

Family 2: Sentiment Analysis (4 methods). Lexicon-based: VADER (Hutto and Gilbert, 2014), SentiWordNet (Baccianella et al., 2010). Supervised ML: SVM, Logistic Regression, Random Forest on TF-IDF features (10K dimensions). Deep learning: fine-tuned multilingual BERT (Devlin et al., 2019) (110M parameters, 3 epochs).

Family 3: Document Partitions (5 algorithms). Hierarchical (Ward), spectral, GMM, DB-SCAN, and K-Means clustering applied to document representations from each semantic method. Cluster quality is assessed via silhouette score (Rousseeuw, 1987).

Family 4: Relational Structures (2 methods). Co-occurrence networks with PMI weighting (Church and Hanks, 1990) and Louvain community detection (Blondel et al., 2008), and Correspondence Analysis (CA; Benzécri, 1992; Greenacre, 2007, 1984) on term-rating contingency tables.

2.4. Validation Protocol

Intra-method validation assesses reproducibility via: (i) K-fold cross-validation ($K=10$, stratified), (ii) Procrustes analysis (Gower, 1975) of independent subsamples, (iii) permutation tests (1,000 iterations), and (iv) bootstrap confidence intervals (500 iterations).

Inter-method validation assesses convergence via: (i) variance decomposition into substantive, linguistic, and methodological components, (ii) cross-method Procrustes similarity and top-term overlap, and (iii) cross-platform replication in Python (scikit-learn, gensim, transformers) and R (quanteda (Benoit et al., 2018), FactoMineR (Lê et al., 2008), igraph (Csárdi and Nepusz, 2006), tidytext (Silge and Robinson, 2016)).

Commensurability across method families. The eighteen methods produce heterogeneous outputs—probability distributions (LDA), dense vectors (embeddings), class labels (sentiment), contingency tables (CA), graphs (networks). Comparability is achieved through two strategies. For geometric methods (LSA, LDA, CA, embeddings), outputs are projected into low-dimensional term-level representations and compared via Procrustes analysis (Gower, 1975), complemented by top- k term overlap. For classification and clustering methods, agreement is assessed via shared metrics (accuracy, F1) on identical held-out sets and cross-tabulation of predicted labels. Variance decomposition operates at the corpus level: for each method, we extract a summary

score per document (dominant topic probability, sentiment polarity, cluster assignment) and decompose total variance via nested ANOVA with rating, language, and method as factors.

The 5% robustness threshold. Method-invariant patterns are operationalized as those where between-method variance remains below 5% of total variance. This threshold reflects both empirical and methodological reasoning. Empirically, it was calibrated on the observed variance distribution: the maximum between-method contribution was 3.7%, so the 5% bound provides a conservative margin. Methodologically, it aligns with conventions in multivariate analysis, where variance components below 5% are regarded as negligible (Hair et al., 2019), and with effect-size benchmarks where $\eta^2 < 0.06$ indicates a small effect (Cohen, 1988). The specific threshold may require recalibration for corpora with different characteristics, but the underlying principle—that method-invariant findings are those where analytical choices contribute negligibly to total variance—is general.

3. Results

3.1. Lexicometric Analysis

The top terms reveal a clear semantic structure organized around seven categories: positive evaluation (27.0%), geography/location (20.1%), service quality (16.7%), facilities (15.4%), amenities (9.9%), cleanliness (7.2%), and negative evaluation (3.0%). Distinctive terms per rating class, identified via the Log-Likelihood Ratio (Dunning, 1993), reveal marked asymmetry in lexical intensity: negative reviews deploy more extreme vocabulary (LLR 3,000–3,700) than positive reviews (LLR 2,200–2,900). Sentiment-contrastive analysis shows that positive terms are 37–47 $\times$ over-represented in 5-star reviews, while negative terms are 84–92% under-represented, confirming high discriminative power. Temporal evolution (2015–2024) reveals COVID-related terms (+1,900–9,000%), technology vocabulary (+100–2,800%), and sustainability terms (+1,200–1,600%).

3.2. Latent Semantic Analysis

LSA applies truncated SVD to TF-IDF matrices, extracting latent dimensions that capture co-occurrence patterns and implicit synonymy (Deerwester et al., 1990; Landauer et al., 1998). The analysis reveals remarkably consistent dimensional structures across languages. The first component captures 31–34% of total variance ($CV < 3\%$), representing a sentiment/quality axis (excellent $\leftrightarrow$ terrible). Components 1–4 capture aspectual dimensions: location vs. infrastructure, service vs. amenities, value vs. luxury, and size vs. experience. The parallelism of multilingual variance profiles validates universal semantic organization. Document clustering in LSA space identifies 5–8 coherent clusters per language, strongly correlated with ratings ($r = 0.91$).

3.3. LDA Topic Modeling

LDA models documents as mixtures of latent topics (Blei et al., 2003; Griffiths and Steyvers, 2004). Coherence optimization (Röder et al., 2015) yields $K = 8$ –10 optimal topics across languages (mean $C_v = 0.445$). Seven thematic categories emerge consistently—a finding aligned with polylingual topic modeling (Mimno et al., 2009): location, room quality, service, breakfast, value, attractions, and overall evaluation. Topic-rating correlations are strong ($r = 0.74$ – 0.85). LDA and LSA are complementary: LDA’s silhouette (0.062) exceeds LSA’s (0.034), while LSA provides richer geometric interpretability. Their convergence (Procrustes 0.82) validates the shared underlying structure.

3.4. Word Embeddings

Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) embeddings provide distributed vector representations capturing semantic proximity. Domain-specific analogy tasks validate the learned representations (Top-1 accuracy 78–85%, Top-3: 92–96%). Dimensionality reduction via t-SNE (van der Maaten and Hinton, 2008) reveals four primary clusters converging with LDA topics and network communities: Location, Service, Positive sentiment, and Negative sentiment. GloVe is marginally superior on analogy tasks (+2.4pp, $p < 0.05$), while document clustering quality is comparable. The Procrustes alignment between Word2Vec and GloVe spaces (0.91) confirms both methods converge on the same latent structure.

3.5. Correspondence Analysis

CA operates on the term-rating contingency table, projecting both terms and ratings into the same factorial space via SVD of standardized residuals (Benzécri, 1992; Greenacre, 2007). Applied to 500 terms, it yields three significant dimensions explaining 97.3% of inertia (all $p < 0.01$): Dimension 1 (76.8%) captures sentiment polarity, Dimension 2 (15.1%) contrasts extreme and intermediate ratings, and Dimension 3 (5.4%) captures residual variation. The biplot structure is isomorphic across all five languages (Procrustes 0.84–0.87). Pairwise inertia analysis reveals that the 1 vs. 5 star contrast alone contributes 34% of total inertia (Cohen’s $w = 0.87$).

3.6. Sentiment Classification

Table 2 presents the comparative performance across six methods, spanning lexicon-based (Hutto and Gilbert, 2014; Taboada et al., 2011), supervised ML (Liu, 2012), and deep learning approaches (Devlin et al., 2019). The largest gap lies between lexicon-based and supervised approaches: +10.9pp accuracy from VADER to SVM, driven by improvement on neutral (+48.6pp F1) and negative (+31.4pp F1) classes. BERT achieves the highest accuracy (91.3%), but the gain over SVM is modest (+2.2pp) relative to the computational cost (29× inference time, 23× memory). Logistic Regression offers the optimal Pareto balance of accuracy (88.7%), speed (2.8ms), and interpretability. The method hierarchy is universal across all four languages tested.

Table 2: Sentiment classification performance. The largest gain occurs at the lexicon → supervised ML boundary (+10.9pp). BERT’s marginal advantage over SVM (+2.2pp) comes at 29× inference cost.

Method	Acc.	F1 Mac.	F1 Pos	F1 Neu	Inf. (ms)
BERT fine-tuned	91.3%	0.834	0.948	0.741	87
SVM (linear)	89.1%	0.798	0.935	0.701	3.1
Logistic Reg.	88.7%	0.791	0.932	0.687	2.8
Random Forest	87.3%	0.769	0.924	0.658	12.4
VADER	78.2%	0.512	0.876	0.215	1.2
SWN Ensemble	76.3%	0.521	0.857	0.215	8.7

3.7. Co-occurrence Network Analysis

The English network (500 terms, PMI-weighted (Church and Hanks, 1990)) exhibits small-world properties (Watts and Strogatz, 1998): density 0.099, clustering 0.342, diameter 4, power-law degree distribution ($\gamma \approx 2.3$) characteristic of scale-free networks (Barabási and Albert,

1999). Centrality analysis (Freeman, 1978) identifies evaluative terms as lexical hubs. Louvain community detection (Blondel et al., 2008) identifies 7 communities ($Q = 0.487$) corresponding to Location, Service, Room, Positive evaluation, Breakfast, Value, and Negative evaluation. Intra-community density (0.432) is $4.4\times$ higher than global density, with low conductance (0.152). The community structure is stable over time (mean ARI = 0.827 across 2015–2024 windows) and convergent across languages (modularity 0.46–0.52).

4. Cross-Method Synthesis

4.1. Intra-Method Stability

K-fold cross-validation and Procrustes analysis confirm high reproducibility across all languages (Table 3). Permutation tests validate non-trivial structure ($p < 0.001$): for English, real inertia (0.1855) substantially exceeds the null distribution (0.1021 ± 0.0007).

Table 3: Intra-method stability (CA on term-rating matrices). All languages exhibit high reproducibility: stability > 0.99 , Procrustes > 0.96 .

Language	K-fold CV	Procrustes	Stability
English	0.006	0.991	0.994
Italian	0.003	0.991	0.997
Spanish	0.007	0.988	0.993
French	0.009	0.969	0.991
German	0.007	0.988	0.993
Mean	0.006	0.985	0.994

4.2. Cross-Method Alignment

The alignment across extraction methods confirms convergent validity (Table 4). The strong CA–LSA alignment (Procrustes 0.892) is expected given shared SVD-based decomposition, but the substantial CA–LDA and LSA–LDA alignment (0.763–0.820) is more noteworthy, as these methods rest on fundamentally different mathematical foundations. Network communities align with LDA topics at 68–82% term overlap, and Word2Vec–GloVe Procrustes similarity reaches 0.91.

Table 4: Cross-method alignment: Procrustes similarity and top-50 term overlap confirm convergent validity across mathematically distinct decompositions.

Comparison	Procrustes	Spearman	Overlap (%)
CA Dim1 $\leftrightarrow$ LSA Comp0	0.892	0.912	84
CA Dim1 $\leftrightarrow$ LDA Pos/Neg	0.763	0.789	72
LSA Comp0 $\leftrightarrow$ LDA Pos/Neg	0.820	0.845	78
Network $\leftrightarrow$ LDA topics	—	—	79.6
Word2Vec $\leftrightarrow$ GloVe	0.910	—	—

4.3. Variance Decomposition

The central finding emerges from variance decomposition (Table 5): substantive content (rating valence) accounts for the overwhelming majority of observed variance, while methodological

and linguistic factors contribute minimally. Between-method variance remains well below the 5% threshold that operationalizes our definition of robustness, confirming that method-invariant patterns capture genuine structure.

Table 5: Variance decomposition: rating (content) explains 95.4% of total variance. Methodological choice accounts for less than 3%.

Source	Variance (%)
Rating (substantive content)	95.4
Language	0.9
Residual (incl. method)	3.7

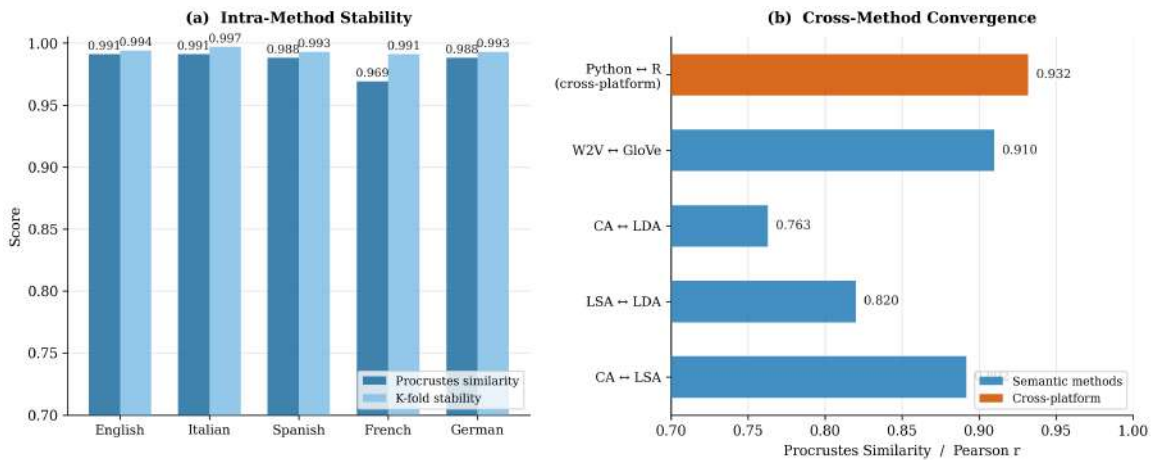


Figure 1: Convergence evidence across two validation levels. (a) Intra-method stability: Procrustes similarity and K-fold stability coefficients are consistently high across all five languages (range: 0.969–0.997). (b) Cross-method convergence: Procrustes similarity between analytically distinct method pairs and cross-platform Pearson r (Python vs. R) confirm that validated patterns transcend both methodological and implementation choices.

4.4. Cross-Platform Replication

Independent implementation in Python and R confirms that findings transcend software-specific choices (Table 6). All methods achieve Pearson $r > 0.89$, with mean $r = 0.932$. Deterministic methods show near-perfect replication ($r > 0.99$); stochastic methods exhibit slightly lower convergence due to algorithmic variants (VEM vs. Gibbs for LDA, SVD solver differences for LSA).

5. Discussion and Conclusion

This paper has introduced a multi-method validation framework for text analytics, centered on the identification of *method-invariant patterns*: findings that emerge consistently across analytically independent approaches. Applied to 999,152 multilingual reviews through eighteen techniques, the results converge on a methodologically consequential conclusion.

The central finding is that robust patterns are method-invariant. Variance decomposition reveals that substantive content accounts for 95.4% of total variance, while methodological choice contributes less than 3%—well below the 5% threshold operationalizing our robustness criterion.

Table 6: Cross-platform validation (Python vs. R): mean Pearson $r = 0.932$, confirming implementation independence.

Method	Pearson r	Overlap (%)	Δ time
Lexicometric	0.998	94	<1%
LSA	0.892	84	R 33% faster
LDA	0.913	87	Py 57% faster
CA	0.900	76	<9%
Sentiment	0.892	89	R 39% faster
Network	0.995	93	R 65% faster
Mean	0.932	87.2	—

This invariance manifests across the full range of analytical approaches, from distributional semantics to transformer-based classification, and is reproduced independently in both Python and R implementations (mean Pearson $r = 0.932$). Intra-method stability further reinforces this picture: K-fold coefficients of 0.991–0.997 and Procrustes similarities of 0.969–0.991 confirm reproducibility across data perturbations and all five languages.

The framework also yields practical guidance for method selection. The most consequential methodological divide lies between lexicon-based and supervised approaches, where the transition to supervised learning produces a gain of 10.9 percentage points in accuracy. Within the supervised tier, however, differences are modest: Logistic Regression (88.7%) and SVM (89.1%) approach BERT’s accuracy (91.3%) at a fraction of the computational cost—2.8ms versus 87ms inference time, 0.16GB versus 4.20GB memory. For most large-scale applications, classical supervised methods offer the better balance of accuracy, speed, and interpretability.

The framework’s scope is necessarily bounded, and generalizability deserves explicit consideration. User-generated reviews exhibit particular properties—short texts, evaluative language, star-rating anchors—that may favor convergence. However, several features support broader applicability: the validation protocol is domain-agnostic (Procrustes alignment, variance decomposition, and cross-platform replication require no corpus-specific adaptation), the four method families are standard across text analytics, and the observed convergence spans methods with fundamentally different mathematical foundations (generative vs. discriminative, geometric vs. probabilistic). The 5% variance threshold should therefore be understood as an operational criterion calibrated on the present corpus rather than as a universal constant; for more heterogeneous domains (e.g., scientific literature, political discourse, social media), threshold recalibration remains an open empirical question. Similarly, the five Indo-European languages do not address typologically distant families. A further practical limitation is that exhaustive multi-method validation can be computationally demanding, although our results also suggest that lighter supervised approaches may offer an efficient baseline when a full eighteen-method comparison is not feasible. Extending the protocol to diverse domains, genres, and languages is a natural direction for future research.

The principal contribution is the operationalization of robustness as method-invariance, giving researchers principled, replicable criteria for identifying trustworthy text analytics results. By systematically comparing eighteen techniques across four analytical families and validating findings across two independent software platforms, the framework addresses a methodological gap that has received limited attention despite the proliferation of computational text analysis tools.

References

- Baccianella S., Esuli A., and Sebastiani F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC'10*, pp. 2200–2204. European Language Resources Association.
- Barabási A.-L. and Albert R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- Benoit K., Watanabe K., Wang H., Nulty P., Obeng A., Müller S., and Matsuo A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.
- Benzécri J.-P. (1992). *Correspondence Analysis Handbook*. Marcel Dekker, New York, NY.
- Blei D. M., Ng A. Y., and Jordan M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Blondel V. D., Guillaume J.-L., Lambiotte R., and Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Church K. W. and Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition.
- Covington M. A. and McFall J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.
- Csárdi G. and Nepusz T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695:1–9.
- Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- Devlin J., Chang M.-W., Lee K., and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, pp. 4171–4186. Association for Computational Linguistics.
- Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- Freeman L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.
- Gower J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51.
- Greenacre M. (2007). *Correspondence Analysis in Practice*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition.
- Greenacre M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.
- Griffiths T. L. and Steyvers M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.
- Hair J. F., Black W. C., Babin B. J., and Anderson R. E. (2019). *Multivariate Data Analysis*. Cengage Learning, Andover, 8th edition.
- Hutto C. J. and Gilbert E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pp. 216–225.
- Landauer T. K., Foltz P. W., and Laham D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.
- Lê S., Josse J., and Husson F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

- Lebart L., Salem A., and Berry L. (1998). *Exploring Textual Data*. Springer Science & Business Media, Dordrecht.
- Liu B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Mikolov T., Chen K., Corrado G., and Dean J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, pp. 1–12.
- Mimno D., Wallach H. M., Naradowsky J., Smith D. A., and McCallum A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 880–889. Association for Computational Linguistics.
- Peng R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.
- Pennington J., Socher R., and Manning C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pp. 1532–1543.
- Porter M. F. (2001). Snowball: A language for stemming algorithms.
- Röder M., Both A., and Hinneburg A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399–408. ACM.
- Rousseeuw P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.
- Salton G. and Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.
- Silge J. and Robinson D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1(3):37.
- Taboada M., Brooke J., Tofiloski M., Voll K., and Stede M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.
- van der Maaten L. and Hinton G. (2008). Visualizing data using t-SNE. In *Journal of Machine Learning Research*, volume 9, pp. 2579–2605.
- Watts D. J. and Strogatz S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442.
- Zipf G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

Transforming UN Internal Audit Reports into Oversight Knowledge: Lexicometric Perspective

Li Chen¹, Domenica Fioredistella Iezzi²

¹University of Rome “Tor Vergata” – li.chen@students.uniroma2.eu

²University of Rome “Tor Vergata” – stella.iezzi@uniroma2.it

Abstract

Internal audit reports are central to organizational oversight and accountability in the United Nations (UN), yet their narrative complexity and volume impede systematic, longitudinal analysis. This study analyzes 1,309 internal audit reports issued by the UN Office of Internal Oversight Services from 2013 to 2025 through a multi-paradigm framework that compares lexicometric clustering (Reinert/Descending Hierarchical Classification, DHC), probabilistic modeling (Latent Dirichlet Allocation, LDA), and a modified embedding-based approach (BERTopic coupled with K-Means clustering), with comparative characterization based on intrinsic interpretability proxies through semantic coherence and structural geometric properties derived from silhouette separation.

The study shows that five stable oversight domains structure the corpus at the macro level. Longitudinal patterns indicate a shift from material asset accountability and field-level compliance in earlier years toward strategic governance, digital systems, and performance management in recent years. Methodologically, the three modeling paradigms operate at complementary resolutions: the Reinert method establishes the macro-structural backbone of audit discourse, LDA represents mixed-membership thematic patterns, and BERTopic isolates finer-grained and action-oriented sub-themes. The findings demonstrate that combining classical lexicometry with probabilistic and embedding-based models yields a more faithful and scalable representation of internal audit discourse, providing a data-driven foundation for risk-based internal audit practices in international organizations.

Keywords: United Nations, Internal Audit, Topic Modeling, Reinert Method, LDA, BERTopic, Institutional Accountability.

1. Introduction

Internal audit reports constitute an essential mechanism for organizational accountability and for evaluating risk management and internal controls (Power, 1997). Within the United Nations (UN), these documents are critical given the high-risk, multifaceted nature of its peacekeeping and humanitarian mandates. However, the narrative complexity and sheer volume of these archives hinder systematic oversight, as manual reviews may overlook systemic risk patterns or longitudinal trends across the Organization (UN OIOS, 2025).

This paper analyzes 1,309 reports by the UN Office of Internal Oversight Services (OIOS) from 2013 to 2025, focusing on the standardized “Audit Results” sections. The public availability and continuous release of these reports make longitudinal analysis possible. At the same time, extracting thematic structures from such a corpus presents a significant methodological challenge, as topic modeling comprises diverse statistical paradigms that are not inherently commensurable. To establish a robust analytical framework, we apply and compare three distinct approaches: lexicometric clustering (the Reinert method), probabilistic modeling (Latent Dirichlet Allocation, LDA), and the embedding-based approach (BERTopic). Acknowledging their distinct statistical foundations, these models are integrated into a

comparative analytical framework to reveal a gradual shift in audit attention from material asset control toward strategic governance and digital systems, while capturing the structure of internal audit discourse at different analytical resolutions.

This study contributes to reframing internal audit reports, moving beyond their conventional view as mere compliance instruments to highlight their role as a cumulative body of oversight knowledge within international organizations. To our knowledge, this is the first systematic, comparative textual analysis of UN internal audit discourse at scale, treating audit findings as a longitudinal corpus rather than isolated documents. Methodologically, we demonstrate how a multi-paradigm topic modeling strategy enables a multi-resolution reading of audit discourse, extracting both stable structural domains and granular themes to complement further analysis.

We structure the remainder of the paper as follows. In Section 2, we outline the theoretical and methodological background on oversight discourse and textual analysis. Section 3 details our corpus, preprocessing procedures, and the comparative topic modeling framework. We then discuss the empirical findings in Section 4, focusing on the structural organization and dynamic evolution of UN internal audit discourse. Finally, Sections 5 and 6 conclude the paper by addressing limitations and avenues for future research.

2. Background

2.1. The lexical construction of oversight in international organizations

Internal audit reports are not merely neutral recordings of facts but also “rituals of verification” (Power, 1997) that construct organizational reality through specific linguistic choices. For international organizations, the linguistic structure of oversight is dictated by multidimensional mandates and standardized reporting protocols.

Despite the proliferation of research on leveraging textual analysis for public-facing UN texts, such as the UN General Debate Corpus (Baturu et al., 2017), technical oversight archives have remained under-researched in this field. This study addresses this gap by treating the OIOS “Audit Results” from 2013 to 2025 as a longitudinal corpus, enabling a transition from manual case-based review to a comprehensive analysis of institutional governance evolution.

2.2. Textual analysis in the domain of internal audit

The transition from manual qualitative review to the systematic analysis of large-scale audit archives necessitates the integration of formal textual analysis. Current research in this domain primarily follows two trajectories. One involves classification and anomaly detection, where models are deployed for task-specific risk identification. For instance, Beltran (2023) utilized paragraph classification to identify monetary discrepancies in sub-national audits, where text is treated as a secondary feature for predictions rather than the primary object of thematic inquiry.

The other trajectory is situated within corporate disclosure analysis where topic modeling is applied to 10-K filings to correlate risk disclosure with financial variables such as audit fees (Dyer et al., 2017). However, these findings are not directly transferable to international organizations as their reporting follows a non-promotional and evidence-oriented reporting logic that lacks the strategic “impression management” found in corporate financial disclosures.

To address the gaps above, this study proposes a multi-paradigm comparative framework to decode thematic analysis of UN’s internal audit focus. The extraction of thematic structures from complex institutional corpora involves a fundamental tension between different statistical

traditions: the lexicometric stability of Reinert’s hierarchical clustering, the probabilistic mixtures of LDA, and the embedding-based approach (BERTopic). Rather than selecting a single approach or assuming cross-model performance equivalence, this study leverages the complementarity of these distinct statistical paradigms. This comparative design allows for a multi-resolution analysis of oversight discourse spanning from the macro-structural domains captured by classical lexicometry to granular functional patterns identified by neural models, thereby providing a more robust basis for risk-based internal audit practices.

3. Methodology

3.1. *Corpus and preprocessing*

We compiled a text corpus of OIOS internal audit reports published between 2013 and 2025. From an initial download of 1,520 publicly available PDF documents, 1,309 monolingual English reports were retained for analysis, while 211 documents were systematically excluded for containing extensive confidentiality redactions and institutional blackouts that caused severe character fragmentation and prevented stable parsing, or for consisting of short administrative summaries or financial tables lacking narrative findings.

To minimize repetitive administrative boilerplate and ensure textual homogeneity, we restrict the extraction to only Section III (“Audit Results”), which contains the substantive audit findings and follows a relatively stable structure across reports, using the PyMuPDF library. Text normalization was executed via a rule-based heuristic pipeline that: (i) detects the “Audit Results” section boundaries using rule-based structural markers; (ii) removes formatting artifacts; and (iii) reconstructs coherent paragraphs while preserving headings and numbered paragraph markers. Each report is represented as a single document consisting only of the reconstructed “Audit Results” section. Report-level metadata detailing year and audited entity were subsequently appended to each standardized document for downstream analyses.

3.2. *Topic extraction methods*

To compare thematic structures obtained under different statistical paradigms, we apply three complementary approaches: a lexicometric clustering baseline (Reinert/DHC), a probabilistic topic model (LDA), and an embedding-based clustering approach (BERTopic).

3.2.1. *Chunking strategy for LDA and BERTopic*

To standardize the structural input for methods requiring discrete document boundaries, both LDA and BERTopic were trained on a shared chunk-level representation. Each “Audit Results” section is segmented using a recursive splitting strategy into chunks of approximately 1,000 characters, with a 150-character overlap. Segmentation prioritizes paragraph and sentence boundaries before falling back to finer-grained splits, preserving local semantic coherence while avoiding arbitrary truncation. This strategy generates a total of 34,106 text chunks.

This design reduces topic dilution when long, multi-thematic documents are modeled as single units, a common characteristic of internal audit reports. Overlapping windows further mitigate boundary effects. These chunked units constitute the input for both LDA and BERTopic models, while the Reinert method follows its native segmentation logic as implemented in IRaMuTeQ.

3.2.2. *Reinert method (Descending Hierarchical Classification, DHC)*

As a lexicometric baseline, the Reinert method uses a top-down hierarchical procedure that partitions text segments into lexical classes based on systematic patterns of word co-occurrence

(Reinert, 1983). We utilize IRaMuTeQ to perform DHC, maximizing between-class separation via χ^2 -based specificity (Camargo & Justo, 2013). Unlike subsequent models, this approach automatically determines the number of classes based on the hierarchical stability of the DHC process, providing an objective lexicometric reference point and serving as a baseline for comparison for probabilistic and embedding-based topic models.

3.2.3. Latent Dirichlet Allocation (LDA)

To provide a probabilistic topic modeling benchmark, we fit LDA, a generative Bayesian model in which each document is represented as a mixture of latent topics and each topic is a distribution over words (Blei et al., 2003). We estimate LDA using Gensim’s LdaModel, whose core inference follows an online variational Bayes procedure (Hoffman et al., 2010). Prior to modeling, preprocessing was conducted including tokenization, stopword removal, and lemmatization; terms appearing in fewer than five documents or more than 50% of the corpus are pruned. As noted, LDA is estimated on the same chunked units described above to maintain granularity consistency with BERTopic for downstream evaluation.

3.2.4. BERTopic with external K-Means clustering

We implement an embedding-based pipeline in which clustering and topic representation are handled as separate steps. Each chunk is encoded into a 768-dimensional dense semantic vector using the all-mpnet-base-v2 Sentence-BERT model, which is robustly pre-trained for formal institutional English (Reimers & Gurevych, 2019). Dimensionality reduction was performed using UMAP (McInnes et al., 2018) configured with 5 components, 15 nearest neighbors, and cosine distance. To circumvent the unconstrained cluster counts inherent to standard HDBSCAN and enable controlled cross-model comparison, a modified BERTopic pipeline was constructed by coupling the UMAP reduced embedding space with an external K-Means clustering layer using an automated initiation setting, and thematic groups are obtained for a grid of k from 5 to 50 at a step of 5. Finally, human-interpretable topic descriptors were generated via class-based TF-IDF (c-TF-IDF) representation (Grootendorst, 2022).

3.3. Evaluation framework

Given the different statistical paradigms of the three approaches, we evaluate the extracted thematic structures using two complementary internal criteria: topic coherence for semantic interpretability and Silhouette index for structural quality. This strategy is motivated by the gap between statistical fit and human interpretability in topic modeling, where models optimizing likelihood-based criteria do not necessarily yield more meaningful topics (Chang et al., 2009).

Rather than identifying a single optimal number of topics for LDA and BERTopic, we compare models across a grid of topic numbers $k \in \{5, 10, \dots, 50\}$, focusing on the stability and trends of structural and semantic indicators as model complexity increases.

3.3.1. Interpretability and proxy topic coherence

To evaluate semantic interpretability, we compute topic coherence using the C_v measure, widely adopted for intrinsic topic evaluation. It is defined as the degree to which a topic’s most representative words co-occur in the reference corpus under a sliding-window probability estimator, using normalized pointwise mutual information (NPMI) and cosine similarity, aggregated across the topic’s top- N words (Röder, Both, & Hinneburg, 2015).

For two words w_i and w_j , NPMI is defined as below in the formula (1):

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \quad (1)$$

where probabilities are estimated from co-occurrences within a fixed sliding window over the reference corpus. In the C_v configuration, these pairwise associations are combined into word vectors and summarized via cosine similarity, with the final topic coherence obtained by averaging the resulting similarities across the top- N words.

In practice, each topic is represented by its top 10 words and the C_v is computed with respect to the chunk-level reference corpus rather than full reports, as report-level co-occurrence statistics may obscure fine-grained thematic structure in long and multi-thematic audit reports. The same chunk-level corpus is used consistently across all methods for fair comparisons.

For the Reinert method, which yields lexical classes rather than probabilistic topics, we compute a proxy coherence by treating each class as a theme and using the top χ^2 -specific words as descriptors. This value is explicitly reported as an interpretability proxy for the descriptor set rather than a mathematical equivalent to probabilistic topic coherence.

3.3.2. Document-level pooling and structural quality

For clustering-based approaches, we assess structural separation using the Silhouette index (Rousseeuw, 1987), an internal validation measure that contrasts within-cluster compactness with between-cluster separation. For an observation i , let $a(i)$ denote the mean distance from i to all other points in its assigned cluster, and let $b(i)$ denote the minimum of the mean distance from i to points in other clusters. The Silhouette of i is defined as in the formula (2):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1] \quad (2)$$

and the overall Silhouette score is obtained by averaging $s(i)$ over all observations.

As BERTopic assigns topics at the chunk level, we derive a document-level representation to enable meaningful structural evaluation. Specifically, chunk embeddings belonging to the same report are mean-pooled to obtain a single document embedding, while the corresponding document label is assigned via majority voting over chunk-level topic assignments. The Silhouette index is then computed on these document-level embeddings and labels.

The Reinert method is excluded from this evaluation due to its non-metric, lexicometric partitioning logic, and LDA is excluded as it produces mixed-membership topic distributions rather than hard document assignments.

4. Results

4.1. Lexicometric structure of audit discourse

This initial lexicometric analysis provides a macro-level overview of the corpus, identifying a limited set of stable lexical classes corresponding to the main oversight domains in UN internal audit reports. Applying DHC to the reconstructed corpus results in a high classification rate of 93.14%, indicating that the preprocessing strategy effectively reduced noise while preserving the semantic coherence of the texts. This level of classification stability is consistent with the highly standardized and technical nature of internal audit reporting.

Five stable lexical classes are identified, which can be interpreted as distinct “lexical worlds” corresponding to core oversight domains. Three classes are closely aligned with the UN’s operational mandates: Governance and Project Management (34.4%), Peacekeeping Operations (13.6%), and Humanitarian Field Operations (13.0%). Two additional classes represent cross-cutting administrative pillars, namely Operational Activities (such as procurement and financial management, 23.8%) and Inventory and Asset Management (15.2%). The results show that UN internal audit discourse is organized around a limited number of recurring oversight domains. The balance between mandate-specific and cross-cutting administrative classes indicates that UN internal audit addresses both operational risks and system-wide governance issues.

4.2. Longitudinal evolution of audit discourse

Beyond the static mapping of domains, a Correspondence Analysis (CA) on the lexical-metadata contingency table models the temporal dynamics of the corpus. The geometric topology reveals a robust chronological trajectory, with Dimension 1 accounting for 27.9% of the total inertia and Dimension 2 explaining 12.1%. The distribution displays a horseshoe effect, reflecting a continuous, systemic repositioning of structural audit focus over time across three distinct institutional phases.

The earliest block (2013–2016) clusters exclusively on the extreme negative pole of Dimension 1, signifying tangible asset accountability and mission-specific operational compliance, driven by the high specificity of terms referencing physical assets, institutional accounting standards (e.g., ‘ipsas’), overhead costs, and field-level identifiers. This pattern corresponds to a baseline period in which OIOS oversight prioritized material asset control and operational compliance.

The second block (2017–2019) occupies the lower apex of the parabolic horseshoe curve. The discourse in this transitional phase gravitated toward structural field mission reorganizations and more integrated oversight frameworks, empirically marked by the rising prominence of institutional concepts such as ‘jmac’ (Joint Mission Analysis Centre) and ‘sdgs’ (Sustainable Development Goals). This indicates the structural pivot of the audit focus, bridging localized field compliance with macro-level review on organizational strategies.

In contrast, audit reports from later period (2020–2025) migrate entirely into the positive quadrant of Dimension 1, showing a significant concentration of terms related to strategic governance and digital transformation. This late-stage shift is evidenced by the growing prominence of terms such as ‘esg’ (Environmental, Social, and Governance), cloud, and ‘kpi’ (key performance indicator), which signal an increasing institutional concern with performance management, sustainability, and system-based oversight. The temporal gradient captured by the CA thus suggests a paradigmatic transformation of UN internal audit from a predominantly operational compliance function toward a forward-looking strategic governance instrument.

4.3. Comparative topic modeling results: LDA and BERTopic

Using the Reinert results as a lexicometric baseline, we examine how probabilistic and embedding-based topic models structure the audit corpus across a grid of topic numbers $k \in \{5, 10, \dots, 50\}$ using topic coherence C_v as a common indicator of semantic interpretability. For the embedding-based approach, the Silhouette index was computed to assess the structural separation of document clusters in the semantic embedding space.

4.3.1. Semantic coherence trends across model complexity

Table 1 reports the evolution of topic coherence C_v for all three methods and the Silhouette index for BERTopic as the number of topics increases.

For LDA, topic coherence increases markedly from $k = 5$ to $k = 15$, reaching a local maximum around $C_v \approx 0.61$. Beyond this point, coherence stabilizes and fluctuates within a narrow band. This plateau indicates that increasing the number of topics mainly redistributes existing semantic content rather than introducing new themes. Such behavior is consistent with the mixed-membership nature of LDA, where increasing k primarily refines probabilistic topic allocations within documents rather than inducing sharper semantic distinctions.

In contrast, BERTopic exhibits a nearly monotonic increase in topic coherence across the entire range of k , indicating that embedding-based clustering continues to extract semantically cohesive topic descriptors as model granularity increases. Unlike LDA, the BERTopic approach appears less constrained by topic overlap at higher k , allowing for finer-grained thematic distinctions without an immediate loss of interpretability.

Table 1. Benchmarking model performance: coherence (C_v) and Silhouette index

Topics (k)	Coherence (C_v)			Silhouette Index
	LDA	BERTopic	Reinert*	BERTopic
5	0.5073	0.5163	0.6700	0.0374
10	0.5659	0.5788	-	0.0232
15	0.6056	0.6013	-	0.0301
20	0.5658	0.6457	-	0.0335
25	0.5877	0.6562	-	0.0218
30	0.5740	0.6636	-	0.0210
35	0.6007	0.6920	-	0.0265
40	0.5958	0.6988	-	0.0329
45	0.5995	0.7054	-	0.0373
50	0.5994	0.7136	-	0.0350

* Reinert results are reported for the optimal hierarchical partition ($k=5$) as determined by the DHC.

4.3.2. Structural separation in the embedding space

Figure 1 summarizes the trends in C_v and Silhouette index across k , complementing the numerical values reported in Table 1. Topic coherence is shown for both LDA and BERTopic on the left axis, while the Silhouette index is reported for BERTopic on the right axis.

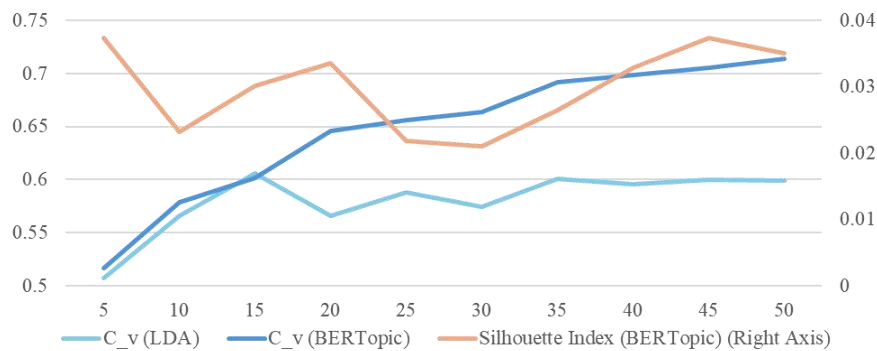


Figure 1. Evolution of Coherence (C_v) and Silhouette Index across different numbers of topics (k)

As illustrated in Figure 1, BERTopic exhibits a nearly monotonic increase in semantic coherence (C_v) as the number of topics scales, whereas the corresponding Silhouette scores remain consistently low across all tested configurations of k . Rather than indicating a methodological limitation, these divergent trends offer a cohesive insight into the structural nature of the corpus.

The steady trajectory of C_v demonstrates that higher granularities empower the embedding-based model to isolate highly cohesive, specialized lexical domains. Conversely, the suppressed Silhouette values mathematically capture the multi-thematic membership inherent to internal audit documentation. Because individual reports sequentially interweave disparate risk vectors into a continuous narrative, aggregating chunk-level embeddings via mean-pooling back to the document level inevitably compresses the report coordinates toward the barycenter of the dense embedding space. Consequently, these low Silhouette values reflect textual continuity and cross-cutting thematic overlap rather than cluster invalidity; the resulting topics remain highly interpretable despite their geometric proximity.

4.4. Qualitative thematic discovery and methodological synthesis

To assess substantive utility, we perform a comparative interpretation of top words in topics generated at two representative points as shown in Figure 1: LDA at $k = 15$ (where coherence reaches its local maximum) and BERTopic at $k = 20$ (where a significant coherence leap coincides with structural stability). The topics from LDA ($k = 15$) are largely defined by standardized audit language, with many themes focusing on procedural follow-ups (e.g., ‘accepted’, ‘action’, ‘pending’) rather than specific risk domains. While LDA identifies broad operational sectors, its mixed-membership structure tends to aggregate diverse risks into single clusters. For instance, LDA Topic 9 combines environmental site inspections, medical personnel, and waste management. While this provides a general overview of certain field operations, it lacks the resolution needed for targeted risk assessment.

In contrast, BERTopic ($k = 20$) effectively isolates these tangled themes into more actionable, localized functional regularities, mapping back to the Reinert baseline. For example, within the Inventory and Asset Management domain (Reinert Class 5), BERTopic differentiates generic administrative oversight into two mutually exclusive risks: stationary warehouse management (‘inventory’, ‘stock’, ‘warehouse’ in Topic 11) and dynamic fleet logistics (‘fuel’, ‘generators’, ‘vehicles’ in Topic 18). This operational granularity is absent in LDA and Reinert results.

Similarly, within the Operational Activities domain (Reinert Class 4), BERTopic isolates ERP system governance (‘umoja’, ‘leave’, ‘travel’ in Topic 13) and even captures the specific interplay between system roles and access controls within UN’s Umoja system, enabling the targeted oversight of “segregation of duties”, a core internal control principle.

Furthermore, BERTopic also separates the field risks aggregated by LDA, splitting them into autonomous themes like tactical peacekeeping support (‘aviation’, ‘medical’, ‘safety’ in Topic 7) and environmental compliance (‘waste’, ‘environmental’ in Topic 17), and captures gender parity as a standalone theme (‘gender’, ‘mainstreaming’ in Topic 19), reflecting its sensitivity to strategic policy shifts that traditional methods categorize as general administrative overhead.

4.5. Synthesis of the results

The joint analysis shows that the three methods operate at complementary resolutions. The Reinert method defines the macro-structural organization of audit discourse; LDA captures the

multi-thematic mixtures and standardized characteristics of audit narration; and BERTopic identifies the specific functional patterns required for risk-based internal audit practices.

5. Conclusions

This study evaluates the thematic organization of OIOS internal audit reports from 2013 to 2025 by comparing lexicometric, probabilistic, and embedding-based modeling approaches. By analyzing “Audit Results” of UN Secretariat for over a decade, we show that internal audit discourse reflects both stable operational mandates and evolving strategic priorities.

Consistent core thematic domains emerge across modeling paradigms, such as governance and project management, peacekeeping and field operations, and asset accountability. The Reinert method offers a lexicometric map of these domains, revealing stable lexical classes structuring audit discourse. This baseline is complemented by probabilistic and embedding-based topic models, which recover similar macro-level themes while providing varied semantic granularity.

The longitudinal analysis shows a shift in audit focus over time. Between 2013 and 2016, internal audit prioritized asset accountability and field operational compliance. The subsequent 2017–2019 triennium functioned as an institutional pivot, witnessing a transitional shift toward structural reorganizations and macro-level strategic planning. From 2020 onward, the discourse significantly migrated toward strategic governance, digital transformation, and performance management. This shift indicates an increased internal audit emphasis on strategic governance and system-level risks, alongside traditional concerns with risks and controls in operations.

Methodologically, our comparative framework indicates that no single modeling approach performs best across all evaluation criteria, as each provides a different level of detail. While the Reinert method establishes the structural backbone of the corpus, the probabilistic LDA model captures the multi-thematic mixtures and standardized characteristics of audit narration. In contrast, the embedding-based BERTopic approach provides more fine-grained thematic distinctions and identifies risk patterns that are frequently diluted in traditional models.

For oversight bodies like OIOS, these results support the use of complementary methods that combine structural consistency with finer-grained thematic analysis. The ability of embedding-based models to isolate granular themes, such as segregation of duties or mission-specific environmental risks, illustrates how textual analysis can support more systematic and scalable audit analysis than keyword-based approaches. These methods provide a data-driven basis for risk-based audit planning, allowing oversight functions to identify systemic vulnerabilities across vast historical data unreviewable manually. In conclusion, as international organizations face increasingly complex environments, integrating these automated frameworks supports a shift from reactive compliance monitoring toward more proactive, data-driven risk analysis.

6. Limitations and Future Research

This study has several limitations that suggest directions for future research. First, the analysis focuses only on the “Audit Results” sections, and future work could extend the framework to other report components to compare oversight narratives across functions. Second, the evaluation relies on intrinsic metrics of semantic coherence and structural separation; future studies could integrate expert-driven or task-based validation linked to audit practice and policy analysis. Third, although embedding-based models identify granular themes, their interpretation still depends partly on the researcher's expertise, which motivates the use of external validation strategies, such as expert labelling or LLM-assisted semantic assessment. Finally, the study is limited to the UN Secretariat, and future comparative analyses across

international organizations could help identify how institutional contexts shape internal audit discourse and support the development of more standardized data-driven oversight frameworks.

References

- Baturo A., Dasandi N., and Mikhaylov S.J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2):2053168017712821.
- Beltran A. (2023). Fiscal data in text: Information extraction from audit reports using Natural Language Processing. *Data & Policy*, 5, e7.
- Blei D., Ng A. and Jordan M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993-1022.
- Camargo B. and Justo A. (2013). IRAMUTEQ: um software gratuito para análise de dados textuais. *Temas em Psicologia*, 21(2): 513-518.
- Chang J., Boyd-Graber J., Gerrish S., Wang C. and Blei D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22.
- Dyer T, Lang M, Stice-Lawrence L (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2-3):221-45.
- Grootendorst M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Hoffman M., Bach F., and Blei D. (2010). Online learning for Latent Dirichlet Allocation. *advances in neural information processing systems*, 23.
- McInnes L., Healy J., and Melville J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Power M. (1997). *The Audit Society: Rituals of Verification*. Oxford University Press.
- Reimers N., and Gurevych I. (2019, November). Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP* (pp. 3982-3992).
- Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2), 187-198.
- Röder M., Both A., and Hinneburg A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).
- Rousseeuw P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53-65.
- United Nations Office of Internal Oversight Services (2025). *Activities of the Office of Internal Oversight Services for the period from 1 July 2024 to 30 June 2025 (A/80/332 Part I)*.

Comparing textometric approaches: Reinert-style Descending Hierarchical Classification versus probabilistic and neural topic modeling on a web-sourced corpus of French rap lyrics.

Robin Quillivic¹, Jules Dilé-Toustou², Pierre Ratinaud³

¹CESSP, Institut des Systèmes Complexes, Paris Île-de-France, CNRS – quillivicrobin@hotmail.fr

²Université de Toulouse - LERASS – dile.jules@gmail.com

³Université de Toulouse - LERASS – pierre.ratinaud@univ-tlse2.fr

Abstract

A genre in its own right, rap has dominated the French music industry since the 2000s and now enjoys unprecedented popularity. Caught between legitimization and stigmatization, French-language rap is becoming increasingly institutionalized, partly through public policy initiatives (Hammou, 2022). It is also gaining autonomy in economic and media terms, through growing market share, the emergence of independent labels, specialized journalists, and dedicated media platforms. Yet while this musical genre was originally rooted in social and political expression, it may now be undergoing a process of "mainstreaming" (variétisation), characterized by a thematic diversification driven by commercial interests (de Courson, 2024). Drawing on a corpus of rap lyrics from the Genius website, Benoît de Courson conducted lexicometric and topic modeling analyses on over 37,000 French-language rap tracks released between 1992 and 2024. He identified seven topics and opened up several avenues of research that we aim to extend. First, the corpus is particularly complex, featuring extensive use of slang, verlan (French backslang), and various forms of textual noise, which call for dedicated linguistic preprocessing. Second, the seven identified topics may obscure certain thematic nuances, as the author himself acknowledges.

Our proposal involves conducting a systematic comparative analysis of this same corpus using three complementary approaches: Descending Hierarchical Classification (DHC) via Iramuteq (Ratinaud, 2014), probabilistic topic modeling with LDA, and neural topic modeling with BERTopic. The comparison draws on multiple dimensions, including inter-model agreement, thematic distribution across artists, temporal dynamics, vocabulary overlap, and intra-topic coherence. Through this work, we aim not only to identify finer-grained themes and better characterize the evolution of French-language rap but also to contribute methodologically to the growing literature on comparing topic modeling approaches.

Keywords: Topic modeling; French rap; Textometry; Comparative analysis; BERTopic; LDA; Iramuteq

Résumé

Genre à part, le rap submerge l'industrie musicale française depuis les années 2000 et connaît désormais une notoriété sans précédent. Entre légitimation et stigmatisation, le rap francophone s'institutionnalise, notamment à travers l'action des pouvoirs publics (Hammou, 2022). Il tend à s'autonomiser sur le plan économique et médiatique (en parts de marché, par la création de labels indépendants, de journalistes spécialisés et de médias dédiés). Mais si ce genre musical est initialement empreint de revendications sociales et politiques, il serait désormais sujet à un phénomène de variétisation, soit une diversification des thématiques à vocation commerciale (de Courson, 2024). Grâce à la constitution d'un corpus de textes de rap issus du site Genius, Benoît de Courson propose des analyses lexicométriques et de topic modeling sur plus de 37 000 morceaux de rap francophone parus entre 1992 et 2024. Il identifie ainsi 7 topics et ouvre plusieurs pistes de recherche que nous souhaitons prolonger. En premier lieu, il s'agit d'un corpus complexe, largement constitué de mots d'argot, de verlan et de bruits divers, ce qui motive un traitement linguistique dédié. Ensuite, les 7 topics identifiés masquent potentiellement certaines spécificités thématiques, comme le souligne d'ailleurs l'auteur.

Notre proposition consiste à mener une analyse comparative systématique de ce même corpus selon trois approches complémentaires : une Classification Hiérarchique Descendante (CHD) via Iramuteq (Ratinaud, 2014), un topic modeling probabiliste par LDA et un topic modeling neuronal par BERTopic. La comparaison s'appuie sur plusieurs dimensions, notamment l'accord inter-modèles, la distribution thématique selon les artistes,

les dynamiques temporelles, le recouvrement lexical et la cohérence intra-topic. À travers ce travail, nous souhaitons non seulement identifier des thématiques plus fines et mieux caractériser l'évolution du rap francophone, mais également contribuer méthodologiquement à la littérature croissante portant sur la comparaison des approches de topic modeling.

Mots-clés : Rap français ; Analyse comparative ; BERTopic ; LDA ; Iramuteq ; Distance de Labbé

1. Introduction

Originally a protest genre that emerged from the United States within the hip-hop culture, rap has gradually gained popularity in France since the 1990s. It is now recognized for both its cultural and economic significance, in terms of market share as well as by public institutions (Hammou et Sonnette-Manouguian 2022). However, its spread is also the result of contributions from a wide range of artists from diverse backgrounds. This lends it a heterogeneity that seems to be constantly growing, as evidenced by the emergence of new subgenres (emodril, cloud rap, frapcore, plugg music, jersey, hyperpop, etc). To explore this diversity and identify its main trends, Benoît de Courson compiled a collection of 37,307 rap lyrics from the Genius website (De Courson 2024), a collaborative platform that catalogs song lyrics from around the world, similarly to Wikipedia. By offering a fresh perspective on French rap, his work identifies seven topics, using Bunka (Dampierre et al. 2024). This suggests a shift toward "mainstreaming" (variétisation) following a period marked by a more aggressive, vulgar style. The complexity of this analysis stems from the fact that rap lyrics constitute a rich but difficult-to-analyze corpus, composed largely of slang, French back slang (*verlan*), neologisms, and "ad-libs". Our goal here is to continue this work by refining these results, based on the hypothesis that rap is more subject to a process of diversification than one *variétisation*, and to determine which textometric method best captures the lexical richness of this corpus.

Comparing topic modeling approaches has a long history. Within textometry, Lebart (in Iezzi et al. 2020) applied six methods to Shakespeare's 154 Sonnets and judged the overall configuration "not that good", concluding that exploratory tools remain essential. Alboni et al. (2023) extended cross-paradigm comparison to five methods, including Reinert's ALCESTE on an e-mobility newsletter, finding moderate agreement (Cramér's $V = 0.52-0.68$) with Reinert's method showing the lowest overlap. Lelu and Cadot (2021) demonstrated through 600 runs across three corpora that optimal clustering-dataspace combinations are corpus-dependent. Within NLP, Egger and Yu (2022) compared LDA, NMF, Top2Vec, and BERTopic on Twitter. Toolkits such as OCTIS (Terragni et al. 2021) standardized the evaluation pipeline but kept the metrics narrow (coherence, diversity). Hoyle et al. (2021) showed that automated coherence metrics do not reliably predict human judgments for neural topic models, and Pereira et al. (2025) argued that word-level metrics miss whether topics induce meaningful document grouping.

Yet several fundamental challenges persist. First, a critical validation gap undermines the dominant evaluation paradigm: Hoyle et al. (2021) demonstrated that automated coherence metrics, developed for classical models, do not reliably predict human judgments when applied to neural topic models; automated evaluations declared a winning model when human evaluations did not. This calls into question the widespread practice of ranking models

solely on coherence scores, which characterizes most BERTopic-vs-LDA studies (Egger and Yu 2022; Gan et al. 2024). Second, existing comparisons, from Lebart's literary sonnets to Alboni et al.'s structured news items and to Cadot and Lelu's benchmark corpora, focus exclusively on well-structured, English-language text. Highly informal, non-standard registers such as French rap lyrics pose challenges that remain entirely unaddressed. Third, while Alboni et al. (2023) demonstrated the value of including Reinert's method alongside factorial and probabilistic approaches, no comparison framework has yet brought together the three major paradigms now available to researchers: Reinert's lexical classification, probabilistic modeling (LDA), and neural embedding-based clustering (BERTopic). Fourth, existing frameworks evaluate topic models through intrinsic quality metrics (coherence, diversity) and pairwise agreement (Cramér's V, as in Alboni et al.), but do not test whether topics capture known external structure (such as artist identity or temporal evolution) nor whether partitioning structures the full lexical space beyond the top-K representative words.

Our work addresses these gaps by proposing a multi-dimensional comparison framework applied to a large corpus of French rap lyrics (115,805 verses, 605 artists, 1992–2023), systematically comparing Reinert's DHC via Iramuteq, LDA, and BERTopic across four dimensions. All code that can be reused in other contexts is available here: https://github.com/psycholinguistics2125/JADT_rap_fr and more results can be found here <https://www.compare-topic-model-jadt2026.fr/>.

2. Corpus and Methods

2.1. Corpus

The corpus is a subpart of de Courson's (2024) collection, consisting of lyrics scraped from the Genius platform. It comprises 115,805 individual verses (couplets) from 605 unique artists, spanning 1992 to 2023 (120,662 before preprocessing). The unit of analysis is the verse rather than the full song, preserving intra-song thematic variation. The corpus is temporally skewed: the median year is 2017, with 52.6% of documents from the 2010s and 26.6% from the 2020s. The most prolific artist is JuL (2,878 verses). Preprocessing was performed with spaCy's French model for tokenization and lemmatization, with specific handling of slang and verlan using stop words and ruled based approach (not applicable for Iramuteq).

2.2. Models

We fix $K=20$ across all paradigms, required for partition-based metrics (ARI, NMI, Cramér's V), refining de Courson's 7 topics. This may disadvantage BERTopic, which typically prefers higher K , a point we return to in §4.4.

The Reinert method in IRaMuTeQ. The results labeled "Iramuteq" in the remainder of this article were obtained by applying the Reinert method (Ratinaud and Marchand 2012; Reinert 1983) to the corpus indexed in IRaMuTeQ (Ratinaud, 2024, « IRaMuTeQ »). The analysis, therefore, focuses on the lemmas of content words (nouns, adjectives, adverbs, verbs, and forms not present in the dictionary) with a frequency of 3 or greater ($n=54,256$). To facilitate comparison with other methods, the corpus is not segmented and uses the 120,662 couplets

resulting from pre-processing as the units of analysis. It should be recalled that the classification is produced through a series of Correspondence Analyses (CA) performed on presence/absence matrices, where the units (in this case, song couplets) are crossed in rows with the selected content word lemmas in columns. The first factor of the CA is used to identify the bipartition of the matrix rows that maximizes the extracted information. At each step, the algorithm proceeds with the largest generated matrix and stops once it has reached the pre-set number of classes. For this analysis, we requested 400 classes to obtain 20 terminal classes that meet the default minimum-size criterion (i.e., the number of analyzed units divided by the number of requested classes). These 20 terminal classes retain 115,805 units (95.97% of the corpus). These specific texts are then used as input for the other methods.

LDA. Latent Dirichlet Allocation (Blei et al. 2003) is a generative probabilistic model that represents documents as mixtures of topics, each topic being a distribution over words. We used Gensim's implementation with bigram preprocessing, symmetric α , automatic η , 15 passes, and 400 iterations, minimum document frequency of 5, and maximum ratio of 0.3. Multiple parameter configurations were tested; we report only the best-performing one here.

BERTopic. BERTopic (Grootendorst 2022) uses transformer embeddings, UMAP (15 neighbors, 5 components, cosine), and KMeans. We test three embeddings: all-mpnet-base-v2, multilingual-e5-large, and a CamemBERT-based encoder, for robustness on French informal text (§3.4). KMeans is preferred over BERTopic's default HDBSCAN because comparisons via ARI/NMI/Cramér's V require a complete partition at $K=20$.

2.3. Evaluation metrics

We evaluate models across four dimensions, going beyond the coherence-and-diversity paradigm dominant in the NLP literature (Terragni et al. 2021; Hoyle et al. 2021). Dimensions 3 and 4 are, to our knowledge, absent from standard topic model comparison frameworks.

D1 - Inter-model agreement. We measure whether the three models produce similar partitions using two complementary metrics: the Adjusted Rand Index (Hubert and Arabie 1985), which counts pairwise agreement between clusterings corrected for chance (range $[-1, 1]$, where 0 is random), and Normalized Mutual Information (Strehl and Ghosh 2002), which quantifies shared information between partitions normalized by entropy (range $[0, 1]$, where 0 is independence).

D2 - Metadata sensitivity. We test whether topics capture known external structure, specifically artist identity, using Cramér's V (Cramér 1946) on the artist $\times$ topic contingency table.

D3 - Intra-topic homogeneity and inter-topic separation. We compute pairwise distances between documents within the same topic (intra) and between each topic and the rest of the corpus (inter), using two complementary metrics: the Labbé distance (Labbé 2007), a standard in French textometry and stylometry, and the Jensen-Shannon divergence (Lin 1991). Because the Labbé distance is sensitive to text length (Cortelazzo et al. 2013), and individual verses average only ~ 68 lemmatized tokens, we compute distances on aggregated pseudo-documents: for each topic, n verses are randomly concatenated into larger textual

units. We report results for three aggregation sizes ($n = 8, 20$, and 72 verses) to show that the ranking between models is stable across scales. We derive the Separation Ratio ($SR = \text{mean inter-topic distance} / \text{mean intra-topic distance}$); values above 1 indicate that topics are more distinct from each other than they are internally heterogeneous. We note that SR serves a similar function to the silhouette score commonly used in cluster evaluation, but operates on raw word distributions shared by all models rather than in a model-specific embedding space, making it a fairer basis for cross-paradigm comparison.

D4 - Global lexical structuration. We compute χ^2/n on the full word $\times$ topic contingency table built from the presence/absence matrix, where N is the total number of word-form occurrences. Unlike coherence (which evaluates only the top- K words per topic) or Labbé/JS (which evaluate document pairs), χ^2/n captures how strongly the entire partition structures the full lexical space. Note that χ^2/n is aligned by construction with Reinert's optimization criterion and is therefore expected to favor Iramuteq; we report it as a transparency measure, not as an arbiter of model quality.

3. Results

3.1. Inter-model agreement (D1)

All pairwise agreements between models are very low, indicating that each paradigm captures genuinely distinct structure in the data (see Table 1).

Table 1. Intermodel agreement metrics

Pair	ARI	NMI	Interpretation
BERTopic vs LDA	0.008	0.015	Near-random
BERTopic vs Iramuteq	0.044	0.102	Weak agreement
LDA vs Iramuteq	0.005	0.015	Near-random

The best, though still modest, agreement is between BERTopic and Iramuteq ($NMI = 0.102$). BERTopic–LDA and LDA–Iramuteq are both near-random ($NMI \approx 0.015$). This contrasts with the moderate agreement found by Alboni et al. (2023) on structured news (Cramér's $V = 0.52$ – 0.68) and the partial convergence reported by Lebart (2020) on literary text, suggesting that inter-model divergence increases with corpus informality and complexity.

3.2. Artist topic association

Table 2. Artist–topic association (Cramér's V).

Model	Cramér's V
Iramuteq	0.385
BERTopic	0.208
LDA	0.114

Iramuteq produces the strongest association between artists and topics ($V = 0.385$), with 12.2% of artists classified as specialists ($>50\%$ of their verses in a single topic). This is consistent with Reinert's design: the method identifies distinct lexical worlds, which, in this corpus, correspond partly to artist-specific vocabularies. BERTopic shows moderate

association ($V = 0.208$), while LDA distributes artists more evenly across topics ($V = 0.114$), suggesting it captures broader co-occurrence patterns rather than artist-level signatures (see Table 2).

3.3. Cluster quality (D3)

We report intra-topic and inter-topic Labbé distances for three aggregation sizes. As pseudo-documents grow longer, absolute Labbé distances decrease monotonically (short texts have higher sampling variance), but the Separation Ratio remains stable or increases, confirming that the ranking between models is robust and structural.

Table 3. Labbé distances by aggregation size and SR ratio.

Model	N aggregation = 8			N aggregation = 20			N aggregation = 72		
	Intra	Inter	SR	Intra	Inter	SR	Intra	Inter	SR
Iramuteq	0.854	0.891	1.043	0.754	0.813	1.078	0.588	0.688	1.170
BERTopic	0.862	0.881	1.022	0.766	0.794	1.037	0.603	0.653	1.083
LDA	0.863	0.875	1.014	0.766	0.784	1.024	0.601	0.630	1.048

Three observations emerge: first, Iramuteq achieves the highest Separation Ratio across all aggregation sizes. Second, the ordering $\text{Iramuteq} > \text{BERTopic} > \text{LDA}$ on the Separation Ratio is stable across all three scales, though absolute differences are modest (1.04 to 1.17). Third, the SR increases with aggregation size, from 1.043 to 1.170 for Iramuteq, meaning that the lexical distinctiveness of Iramuteq’s classes becomes more apparent as sampling noise is reduced. This growing gap confirms that the differences are structural, not artifactual. The Jensen-Shannon divergence yields the same ranking (see the code report).

Figure 1. Labbé distance stabilization by aggregation size. Left: intra-topic (homogeneity). Right: inter-topic (separation). Shaded areas show standard deviation across topics.

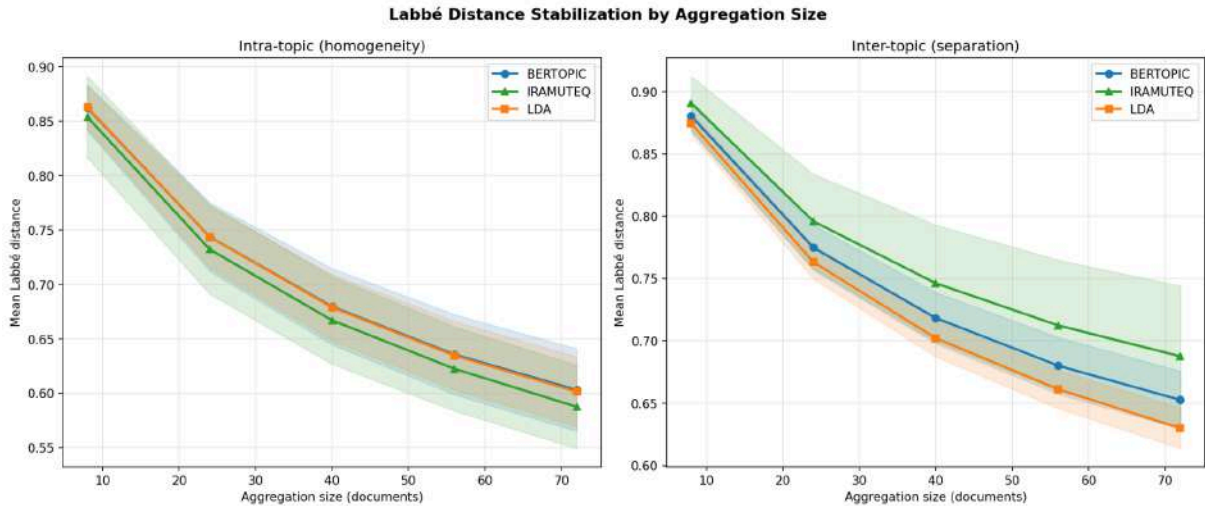


Figure 1 illustrates the smooth, monotonic decrease of Labbé distances with aggregation size for all models. Crucially, the vertical gap between inter- and intra-topic curves (i.e., the numerator of SR minus its denominator) widens with aggregation for Iramuteq, remains moderate for BERTopic, and is smallest for LDA. Iramuteq also exhibits the largest standard

deviation across topics, reflecting the heterogeneity of its classes: some capture tight, artist-specific lexicons, while others capture broader thematic registers.

3.4. Global lexical structuration (D4)

The χ^2/n metric measures how strongly each model's partition structures the full vocabulary. We compute it on the word-form $\times$ topic contingency table built from Iramuteq's presence/absence matrix (5,257,462 total occurrences).

Table 4. Global lexical structuration (χ^2/n) across models

Model	χ^2	N	χ^2/n
Iramuteq	1,143,624	5,257,462	0.218
BERTopic (mpnet)	1,075,395	5,257,462	0.205
LDA (bigram)	1,020,161	5,257,462	0.194

Iramuteq's partition produces a χ^2/n of 0.218, ahead of BERTopic (0.205) and LDA (0.194). We re-emphasize (§2.3) that χ^2/n is aligned with Reinert's criterion by construction, so this ordering reflects lexical-contrast structuration rather than overall model quality. The ordering nonetheless holds across all BERTopic embeddings (mpnet 0.205, CamemBERT 0.203, multilingual-e5 0.210) and LDA n-gram variants (bigram 0.194, trigram 0.201, both 0.197), indicating robustness to configuration.

4. Discussions

4.1. The evolution of French Rap

Based on Iramuteq analysis, we observe a radical shift since 2014, as de Courson (2024) already pointed out. However, this method allows us to identify a more complex phenomenon. The 20 Iramuteq classes cluster into 4 macro-groups corresponding to distinct stylistic registers. The first part provides a clearer definition of the so-called "vulgar" category and of certain artists associated with it (such as the Neochrome collective and Sexion d'Assault). This category accounted for a proportionally significant share of the content until around 2018. A second distinct group of songs emerged in 2014. It is characterized by the influence of the prolific rapper JuL, known for his atypical phrasing, and the emergence of ad-libs typical of "trap music" and later "drill". At the same time, four other categories emerge: slang terms and neologisms associated with illegality, trafficking, and marginal lifestyles. These, too, are markers specific to the emergence of the "drill" style. Another distinct category consists of names and acronyms associated with, among others, the rappers of the 667 collective. A third category includes poetic, romantic, and even philosophical lexical forms. It is the only one that has persisted over the years. A fourth group has declined significantly. It consists of a category associated with the old-school style of the 1990s and 2000s, which has almost completely disappeared. The share of "conscious rap" remains minimally significant, as does a category focused more on sinister and dark vocabulary. Even less significant, a category consisting of English tool-words persists over time despite its small part. Despite the proportional decline in explicitly politicized rap, these analyses offer a broader interpretation

of the categories of French-language rap. The gradual emergence of new subgenres suggests a diversification of styles, some of which echo hip-hop's origins as a marginal culture grappling with issues of violence and drugs, yet one that manages or is trying to break free through music.

4.2. Methodological implications

The near-zero inter-model agreement ($NMI < 0.11$) is itself an informative finding rather than a failure. It means the three paradigms are not interchangeable: Iramuteq captures lexical worlds organized by distinctive vocabulary, BERTopic groups documents by semantic proximity in embedding space, and LDA models word co-occurrence distributions. This extends the gradient observed in the literature: from partial convergence on small literary corpora (Lebart 2020), to moderate agreement on structured news (Alboni et al. 2023), to near-independence on a large informal corpus, suggesting that divergence might increase with corpus complexity and register heterogeneity.

Iramuteq's stronger score on all lexical metrics is partly by design: Reinert's method explicitly maximizes χ^2 between classes, directly optimizing lexical structuration. But beyond this structural advantage, Iramuteq offers several practical strengths that the current literature tends to overlook. It is fully transparent: the classification relies on Correspondence Analysis and χ^2 tests, both well-understood statistical tools, and the resulting classes are directly interpretable through their characteristic vocabulary ranked by χ^2 test-values. It is computationally efficient: the entire 115,805-document analysis can run on a standard laptop, with no GPU requirement. And it naturally produces a hierarchical structure (the dendrogram), enabling researchers to explore topics at varying levels of granularity without re-running the analysis.

By contrast, claims of BERTopic's superiority in recent literature (Egger and Yu 2022; Gan et al. 2024) rest on narrow empirical grounds: coherence scores — which Hoyle et al. (2021) showed do not reliably predict human judgments for neural topic models — applied to well-structured English corpora where pre-trained embeddings are in their comfort zone. On our French informal corpus, our primary embedding (all-mpnet-base-v2) is English-optimized, and even the French-language alternatives we tested (§3.4) showed only marginal differences in χ^2/n . Its topics, characterized by c-TF-IDF weights over embedding clusters, are also less directly interpretable than Iramuteq's χ^2 -ranked profiles. This does not make BERTopic valueless: it captures cross-cutting semantic groupings that Iramuteq's sharp lexical boundaries miss by design, providing a complementary perspective for exploratory work. LDA, conversely, produces the smoothest but least informative partition. The choice between methods should be guided by the research question and corpus characteristics, not by claims of universal superiority based on a single metric on a single type of corpus.

We propose χ^2/n on the word $\times$ topic contingency table as a practical corpus-level metric for comparing how strongly different models structure a given corpus. Its interpretation is straightforward: higher values mean stronger lexical contrast between topics. The absolute value depends on vocabulary size and K , so it is meaningful only for comparisons on the same corpus with the same K .

4.3. Future work

Several avenues remain open. Methodologically, the framework should be applied to less heterogeneous corpora (e.g., news articles) to test whether inter-model divergence indeed scales with register informality; the released code is designed for this. Substantively, the analysis of rap-specific subgenres and of sexism markers across artists and periods is left for future work.

4.4. Limitation

Four limitations should be noted. (i) BERTopic and LDA were applied to the 115,805 verses retained by Iramuteq (4% reduction), tailoring the document base partly to the Reinert pipeline; we accept this trade-off because partition-based metrics require strictly identical units, but it is a structural asymmetry. For LDA, which assigns each text to a single highest-probability topic, this likely represents a disadvantage. (ii) The choice not to segment forces Iramuteq onto units averaging ~160 tokens rather than its default ~40, reducing analytic precision. (iii) $K=20$ simplifies comparison but conceals heterogeneity, particularly for BERTopic which typically prefers higher K . (iv) Three of our four evaluation dimensions are grounded in lexical-distribution geometry — the same geometry Reinert optimizes — and our primary BERTopic embedding (all-mpnet-base-v2) is English-optimized; we mitigated the latter with multilingual-e5 and CamemBERT variants (§3.4) but cannot exclude that a French-trained encoder would change BERTopic's qualitative output. We also acknowledge that one of Iramuteq's developers is among the authors of this study.

References

- Alboni, Fabrizio, Pasquale Pavone, et Margherita Russo. 2023. « The Search for Topics Related to Electric Mobility: A Comparative Analysis of Some of the Most Widely Used Methods in the Literature ». *METRON* 81 (3): 367-91.
- Blei, David M., Andrew Y. Ng, et Michael I. Jordan. 2003. « Latent Dirichlet Allocation ». *Journal of Machine Learning Research* 3 (Jan): 993-1022.
- Cortelazzo, Michele, Paolo Nadalutti, et Arjuna Tuzzi. 2013. « Improving Labbé's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature ». *Journal of Quantitative Linguistics* 20 (mai). <https://doi.org/10.1080/09296174.2013.773138>.
- Dampierre, Charles de, Andrei Mogoutov, et Nicolas Baumard. 2024. « Towards Transparency: Exploring LLM Trainings Datasets through Visual Topic Modeling and Semantic Frame ». arXiv:2406.06574. Prépublication, arXiv, juin 3. <https://doi.org/10.48550/arXiv.2406.06574>.
- De Courson, Benoît. 2024. « LRFAF : une exploration numérique du rap français depuis les années 1990 ». Prépublication, SocArXiv, mars 13. <https://doi.org/10.31235/osf.io/d96tr>.
- Egger, Roman, et Joanne Yu. 2022. « A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts ». *Frontiers in Sociology* 7 (mai). <https://doi.org/10.3389/fsoc.2022.886498>.
- Gan, Lin, Tao Yang, Yifan Huang, et al. 2024. *Experimental Comparison of Three Topic Modeling Methods with LDA, Top2Vec and BERTopic*. https://doi.org/10.1007/978-981-99-9109-9_37.
- Grootendorst, Maarten. 2022. « BERTopic: Neural topic modeling with a class-based TF-IDF procedure ». arXiv:2203.05794. Prépublication, arXiv, mars 11. <https://doi.org/10.48550/arXiv.2203.05794>.
- Hammou, Karim, et Marie Sonnette-Manouguian. 2022. « Chapitre VI. 40 ans de musiques hip-hop en France : (il)légitimation, institutionnalisation et patrimonialisation ». In *40 ans de musiques hip-hop en France*. Ministère de la Culture - DEPS. <https://doi.org/10.3917/deps.hammo.2022.01.0173>.
- Hoyle, Alexander, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, et Philip Resnik. 2021. « Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence ». arXiv:2107.02173. Prépublication, arXiv, octobre 27. <https://doi.org/10.48550/arXiv.2107.02173>.

- Hubert, Lawrence, et Phipps Arabie. 1985. « Comparing Partitions ». *Journal of Classification* 2 (1): 193-218. <https://doi.org/10.1007/BF01908075>.
- Iezzi, Domenica Fioredistella, Damon Mayaffre, et Michelangelo Misuraca, éd. 2020. *Text Analytics: Advances and Challenges*. Studies in Classification, Data Analysis, and Knowledge Organization. Springer International Publishing. <https://doi.org/10.1007/978-3-030-52680-1>.
- Labbé, Dominique. 2007. « Experiments on Authorship Attribution by Intertextual Distance in English* ». *Journal of Quantitative Linguistics* 14 (1): 33-80. <https://doi.org/10.1080/09296170600850601>.
- Lebart L. (2020). Looking for Topics: A Brief Review. In: Text Analytics, Springer, pp. 215–223.
- Lelu, Alain, et Martine Cadot. 2021. « Evaluation of Text Clustering Methods and Their Dataspace Embeddings: An Exploration ». In *Data Analysis and Rationality in a Complex World*, édité par Theodore Chadjipadelis, Berthold Lausen, Angelos Markos, Tae Rim Lee, Angela Montanari, et Rebecca Nugent. Studies in Classification, Data Analysis, and Knowledge Organization. Springer International Publishing. https://doi.org/10.1007/978-3-030-60104-1_15.
- Pereira, Antônio, Felipe Viegas, Diego Roberto Colombo Dias, et al. 2025. « ``Are the Current Topic Modeling Evaluation Metrics Enough?`` Mitigating The Limitations of Topic Modeling Evaluation Metrics Using a Multi-Perspective Game Theoretic Approach ». SSRN Scholarly Paper N° 5202376. Social Science Research Network, avril 2. <https://doi.org/10.2139/ssrn.5202376>.
- Ratinaud, Pierre, et Pascal Marchand. 2012. « Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec IRaMuTeQ ». *11èmes Journées internationales d'Analyse statistique des Données Textuelles* (Liège, Belgium), 835-844 (en ligne : <http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Ratinaud,%20Pierre%20et%20al.%20-%20Application%20de%20la%20methode%20Alceste.pdf>). <https://hal.science/hal-03695856>.
- Ratinaud, P. (2024). IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (Version 0.8 alpha 7) [Windows, GNU/Linux, Mac OS X]. <http://www.iramuteq.org>
- Reinert, Max. 1983. « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte ». *Les cahiers de l'analyse des données* 8 (2): 187-98.
- Strehl, Alexander, et Joydeep Ghosh. 2002. *Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions*.
- Terragni, Silvia, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, et Antonio Candelieri. 2021. « OCTIS: Comparing and Optimizing Topic models is Simple! » In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, édité par Dimitra Gkatzia et Djamé Seddah. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.eacl-demos.31>.

C17

Translation, stylistics and literary variation

09-lug
09:00 - 10:30
Aula Multimediale C

Demonstratives in human and machine translation between English and French

Daniel Henkel¹

¹Université Paris 8 – daniel.henkel@univ-paris8.fr

Abstract

Demonstrative determiners were inventoried in a 20m-word English/French “bidirectional” (comparable-parallel) corpus containing both human and machine-translated target texts. The frequency of demonstratives is higher in French than in English, leading to overuse in English-translated-from-French. Translators also exhibit a preference for *This* at the expense of *That*. The Wilcoxon-Mann-Whitney test confirms that English-translated-from-French is distinct from Original English, while correlation tests and manual analysis of random samples reveal that source language imitation is the main factor. Overall, machine translation amplifies the tendencies observed in human translation.

Keywords: Demonstrative determiners, Corpus linguistics, Translation, English, French

1. Introduction

Since the early 21st century, many corpus-based translation studies have demonstrated differences between corpora of target texts and untranslated reference corpora containing original texts in the same languages. The hybrid nature of translated texts is now widely recognized, although opinions often differ as to whether these differences should be interpreted as evidence of the influence of the source language (“shining-through”) or as consequences of the translation process (i.e., interlinguistic interference or “translationese”). While providing valuable insights into the languages involved and the translation process itself, these comparisons have most often focused on specific linguistic features in terms of overall frequency, without taking into account the variation that exists between texts within the same corpus. This study therefore implements several methodological innovations aimed at overcoming these limitations, such as calculating frequency data on a text-by-text basis rather than across the entire corpus, and analyzing individual examples through the random extraction of pairs of sentence-aligned segments.

This study, one in a series that aims to systematically identify major lexical and syntactic differences between English and French in translation compared with original English and French, seeks to provide a comprehensive account of the use of demonstrative determiners in English, French, and in translations between these two languages, building and expanding upon Johansson's (2007) “bi-directional” model to include both machine translation and human translation, with sentence alignment enabling the analysis of individual source/target segment pairs in addition to macroscopic frequency analysis. The objectives are to determine the level of convergence/divergence between target texts (translations) and the norms of the target language, the degree of correlation between source and target texts, where and when disparities arise, and how demonstrative determiners are translated.

2. Methods

An inventory of public domain literary works and corresponding translations from the late 19th century onward (median=1900) was made from Project Gutenberg and Noslivres.net. All 20th-c. authors whose translations from the same period were also available in the public domain were included, as well as an equivalent number of 19th-c. authors and translators, into a ± 13.5 -million-word corpus consisting of 35 works by 35 different authors in each language and the same number of translations (4×35 authors/translators = 140) in four ± 3.5 m-word subcorpora. Source and target-texts were aligned at sentence level with *LF Aligner* and then double-proofread. A second translation for each source text was produced with DeepL Pro in July 2024, bringing the total to ± 20 m-words with the addition of two more ± 3.5 m-word subcorpora. Finally, all texts were tagged for POS and lemma in *TreeTagger*.

Table 1. Summary of characteristics by subcorpus

	Tokens total	Tokens/text Med. (Min.-Max.)	MAATR ₅₀ ¹ Med. (Min.-Max.)
En0	3488993	82994 (34101-323577)	0.81 (0.78-0.83)
EtrFh	3286812	76053 (23950-239482)	0.81 (0.78-0.85)
EtrFm	3159394	83420 (21711-213519)	0.82 (0.77-0.83)
Fr0	3380712	88865 (23249-231121)	0.85 (0.81-0.87)
FtrEh	3612514	82285 (37963-332440)	0.85 (0.83-0.86)
FtrEm	3799062	89473 (37369-353865)	0.84 (0.81-0.85)

This combination of comparable and parallel subcorpora into a “bidirectional” super-corpus provides a basis for several types of comparisons:

- between Original English (En0) and Original French (Fr0), in order to establish benchmark data for each language;
- between English-translated-from-French (EtrF-h for human translation, EtrF-m for machine translation) and En0, so as to ascertain whether the linguistic indicator under investigation, i.e. demonstrative determiners, has a similar distribution in EtrF compared to En0, and likewise for French-translated-from-English (FtrE) in comparison with Fr0;
- between source- and target-texts, to determine whether correlations exist between the parallel subcorpora (i.e. EtrF-h/m~Fr0 and FtrE-h/m~En0) which could be taken as evidence of interlinguistic interference;
- between human and machine translation.

Noun phrases (NPs) containing demonstrative determiners in English (sing. *This/That*, plur. *These/Those*) and French (sing. *Ce(t)/Cette*, plur. *Ces*), with or without Adjective Phrases (AdjP), were inventoried using regular expressions targeting part of speech and lemma.

¹Moving Average Type Token Ratio was calculated in LancsBox for raw text files prior to lemmatization, hence the higher values in French due to greater morphological diversity of verb forms in particular.

Results for each text were converted to normalized frequencies per 10,000 words (f/10k). Comparisons between subcorpora of target-texts and original texts representing the target-language were evaluated using the Wilcoxon rank-sum test, with Cohen's d as the effect-size measurement. The correlation between source/target texts was assessed using Spearman's correlation coefficient (Spearman's rho). All statistical analyses were performed in R.

3. Results

3.1. Original English/Original French

3.1.1. *This/That/These/Those* vs. *Ce/Ces* + *N* (no Adj)

Demonstrative determiners are more common in French than in English (Fig. 1). The median frequency for all forms combined is 42.64/10k in En0 vs. 53.26/10k in Fr0, an overall difference of +24.9%². Further differences can be observed with respect to singular and plural NPs. The median frequencies for *This* and *That* in English are roughly equal (16.16/10k and 17.66/10k respectively), whereas in plural NPs *These* is over twice as frequent as *Those* (4.2/10k vs. 1.84/10k). In English, the combined median frequency (*This+That*) in singular NPs is 34.91/10k vs. 42.05/10k in French (+20.46%), while the combined median frequency (*These+Those*) in plural NPs is 6.61/10k vs. 10.69/10k in French (+61.63%). The difference between English and French in the use of demonstratives is thus greater in plural NPs.

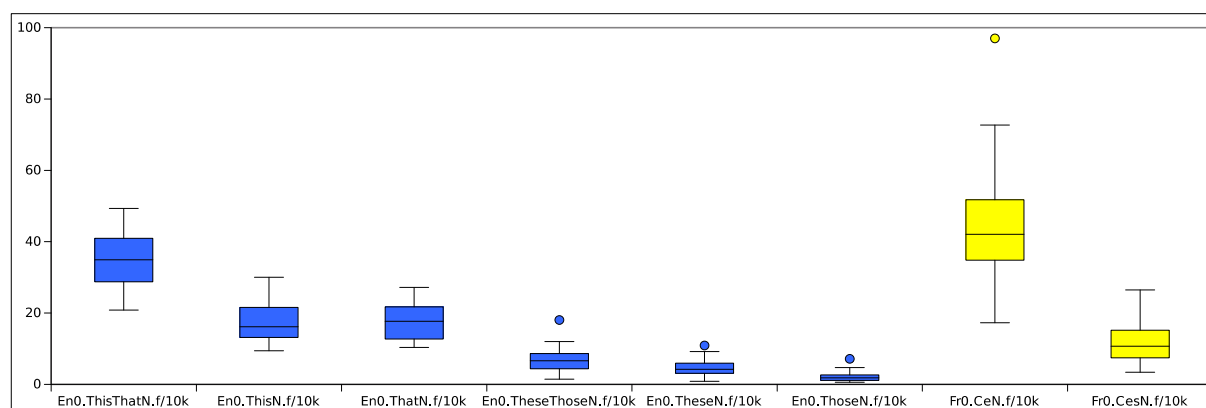


Fig. 1. En0 vs Fr0. Range of frequencies per 10k words in each subcorpus for demonstrative determiners in NPs without Adjs

3.1.2. *This/That/These/Those* vs. *Ce/Ces* + *Adj* + *N*

Similar tendencies hold for demonstratives in NPs containing adjectives (Fig. 2). For all forms combined, the median frequency in English is 12.93/10k vs. 18.26/10k in French (+41.2%). In English, the combined median frequency (*This+That*) in singular NPs is 9.95/k vs. 11.95/k for *Ce* in French (+20.1%), while the combined median frequency (*These+Those*) in plural NPs is 2.38/k vs. 5.06/k for *Ces* in French (+112.6%). In English, the median frequencies of *This* and *That* in singular NPs are again roughly equal (4.44/10k and 4.99/10k), while in plural NPs *These* is somewhat more frequent than *Those* (med. 1.36/10k vs. 1.03/10). In French, the frequency of *Ce* in NPs with adjectives is roughly the same regardless of position (med. 5.86/10k for postnominal Adjs vs. 5.21/10k for prenominal Adjs), whereas *Ces* occurs more often in NPs with postnominal than prenominal Adjs (med. 2.55/10k vs.

²Percentage differences in this article are calculated based on medians, which are taken as the most reliable indicator of central tendency.

1.69/10k). The tendency to use demonstrative determiners more often in French is thus stronger again for plural NPs.

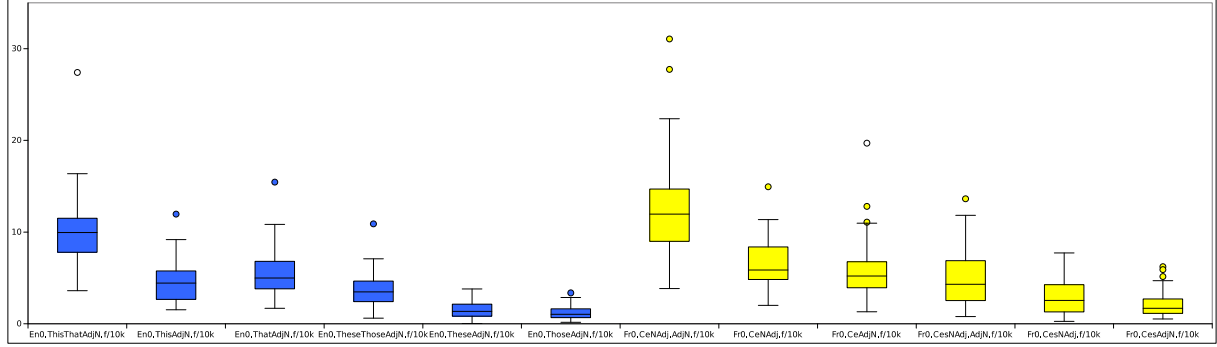


Fig. 2. *En0* vs. *Fr0*. Range of frequencies per 10k words in each subcorpus for demonstrative determiners in NPs with Adjs

3.2. Target texts vs. Target language norms

3.2.1. Demonstratives + N (no Adj)

3.2.1.1 This/That/These/Those

Target texts in English-translated-from-French (EtrF) show a clear tendency to overuse *This/These* and to underuse *That* (Fig. 3). In human-translated texts (EtrF-h), *This* occurs around 47.1% more often than in *En0* (med. 23.78/10k vs. 16.16/10k, +47.1%, $W=922$, $p<0.001$, $d=0.89$ large), *These* occurs about 50.2% more often (med. 6.32/10k vs. 4.2/10k, +50.2%, $W=829$, $p=0.01$, $d=0.69$ medium), while *That* occurs around 28.9% less often than in *En0* (med. 12.55/10k vs. 17.66/10k, -28.9%, $W=378$, $p=0.005$, $d=-0.695$ medium). In machine-translated texts, *This* occurs around 81.1% more often (med. 29.27/10k vs. 16.16/10k, +81.1%, $W=1027$, $p<0.001$, $d=1.32$ large), *These* occurs around 90.7% more often (med. 8.02/10k vs. 4.2/10k, +90.7%, $W=923$, $p<0.001$, $d=0.99$ large), whereas *That* is around 42.9% less frequent than in *En0* (med. 10.13/10k vs. 17.66/10k, -42.6%, $W=131$, $p<0.001$, $d=-1.67$ large). For *Those*, the differences with respect to *En0* are not statistically significant.

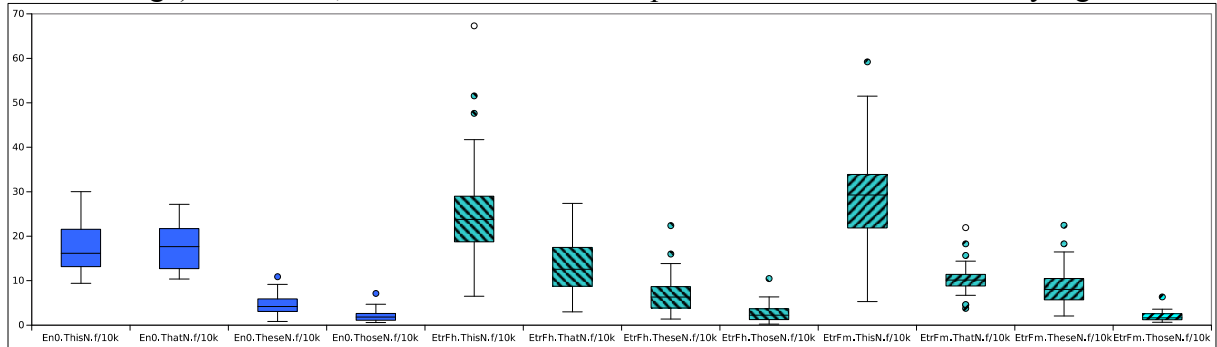


Fig. 3. *En0* vs. *EtrF*. Range of frequencies per 10k words in each subcorpus for demonstrative determiners in NPs without Adjs

3.2.1.2 Ce/Ces

In French-translated-from-English (Fig. 4), the use of demonstratives in human-translated target texts (FtrE-h) is indistinguishable from Original French (*Fr0*). The median frequencies for *Ce* and *Ces* are practically identical (*Ce*: 41.35/10k vs. 42.05/10k, $W=612$, $p=1$, $d=-0.07$ negligible; *Ces*: 9.94/10k vs. 10.69/10k, $W=537$, $p=0.38$, $d=-0.35$ small). Machine-translated target texts (FtrE-m), however, show statistically significant deficits (*Ce*: med. 35.67/10k vs.

42.05/10k, -15.2%, $W=417$, $p=0.02$, $d=-0.65$ medium; *Ces*: med. 7.32/10k vs. 10.69/10k, -31.5%, $W=342$, $p=0.001$, $d=-0.87$ large).

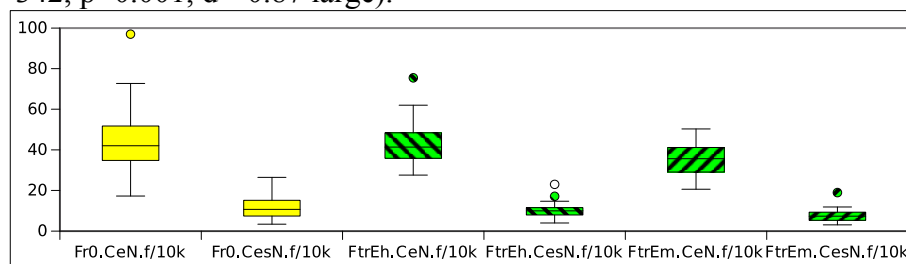


Fig. 4. *Fr0* vs. *FtrE*. Range of frequencies per 10k words in each subcorpus for demonstrative determiners in NPs without Adjs

3.2.2. Demonstratives + Adj + N

3.2.2.1 *This/That/These/Those* + Adj + N

Disparities between English-translated-from-French (*EtrF*) and Original English (*En0*) are greater in NPs with adjectives (Fig. 5). Human translators show a strong tendency to overuse *This* (med. 7.81/10k vs. 4.44/10k, +75.7%, $W=967$, $p<0.001$, $d=1.07$ large), a relatively strong tendency to overuse *These* (med. 2.38/10k vs. 1.36/10k, +75%, $W=844$, $p=0.006$, $d=0.72$ medium), and a small but statistically significant tendency to underuse *That* (med. 3.52/10k vs. 4.99/10k, -29.3%, $W=419$, $p=0.02$, $d=-0.47$ small). These same tendencies are amplified in machine translation: *This* occurs over twice as often as in *En0* (med. 9.52/10k vs. 4.44/10k, +114%, $W=1092$, $p<0.001$, $d=1.52$ large), likewise the frequency of *These* is more than double (med. 3.12/10k vs. 1.36/10k, +129.7%, $W=931$, $p<0.001$, $d=0.94$ large), and underuse of *That* is again quite prominent (med. 2.65/10k vs. 4.99/10k, -46.8%, $W=272$, $p<0.001$, $d=-0.95$ large). The small tendency to overuse *Those* is not statistically significant.

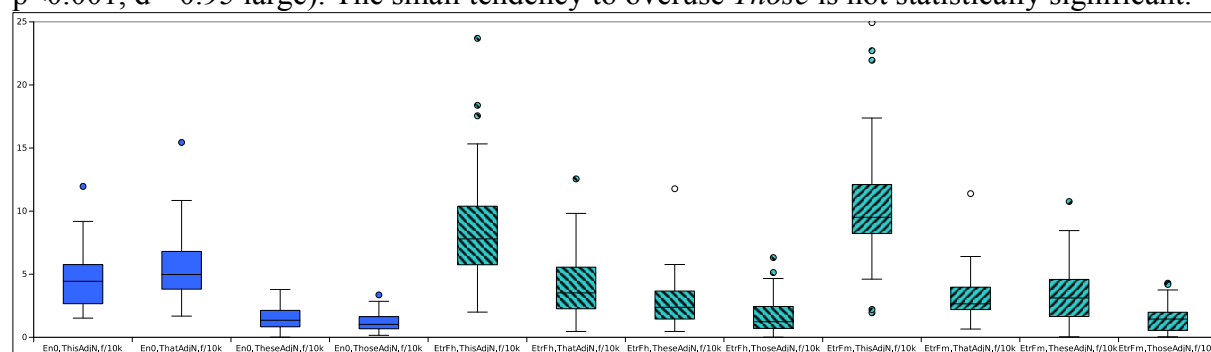


Fig. 5. *En0* vs. *EtrF*. Range of frequencies per 10k words in each subcorpus for demonstrative determiners in NPs with Adjs

3.2.2.2 *Ce/Ces* + N-Adj / Adj-N

In French-translated-from-English (Fig. 6), both human and machine translation diverge significantly with respect to Original French (*Fr0*). Human translators (*FtrE-h*) tend to underuse *Ce* with postnominal adjectives (med. 4.52/10k vs. 5.86/10k, -23%, $W=350$, $p=0.002$, $d=-0.72$ medium), whereas with prenominal adjectives the difference is not significant ($p=0.58$). The same tendency is marginally significant for *Ces* with postnominal adjectives (med. 1.59/10k vs. 2.55/10k, -37.8%, $W=435$, $p=0.037$, $d=-0.75$ medium), but not with prenominal adjectives ($p=0.46$). Machine-translated texts (*FtrE-m*) show a strong tendency to underuse *Ce* both with postnominal adjectives (med. 3.88/10k vs. 5.86/10k, -33.9%, $W=266$, $p<0.001$, $d=-1.04$ large) and prenominal adjectives (med. 4.12/10k vs. 5.21/10k, -20.8%, $W=354$, $p=0.002$, $d=-0.69$ medium). In plural NPs, underuse of *Ces* occurs

in similar proportions with both postnominal and prenominal adjectives (*Ces*+N+Adj: med. 1.34/10k vs. 2.55/10k, -47.3%, $W=376$, $p=0.005$, $d=-0.81$ large; *Ces*+Adj+N: med. 1.19/10k vs. 1.69, -29.2%, $W=385$, $p=0.007$, $d=-0.76$ large).

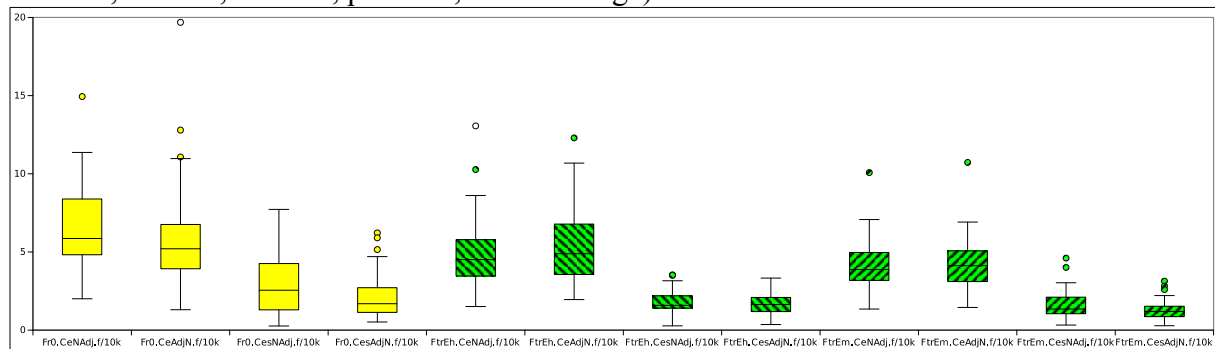


Fig. 6. *Fr0* vs. *FtrE*. Range of frequencies per 10k words in each subcorpus for demonstrative determiners in NPs with Adjs

3.3. Source text influence

Significant correlations can be observed between the frequencies of demonstratives in source and target texts. In English-translated-from-French (EtrF), the correlation between the frequencies of *This* and *That* in human-translated target texts (EtrF-h) and *Ce* in the corresponding Fr0 source texts is strong for *This* (Fig. 7a, $\rho=0.74$, $p<0.001$), moderate for *That* (Fig. 7b, $\rho=0.45$, $p=0.006$) and strongest when the frequencies of *This* and *That* are combined (Fig. 7c, $\rho=0.82$, $p<0.001$). In machine-translated target texts (EtrF-m), the correlation with *Ce* in the Fr0 source texts is nearly perfect for *This* (Fig. 7d, $\rho=0.92$, $p<0.001$), again only moderate for *That* (Fig. 7e, $\rho=0.48$, $p=0.003$) and even closer to perfect for the combined frequencies of *This* and *That* (Fig. 7f, $\rho=0.94$, $p<0.001$).

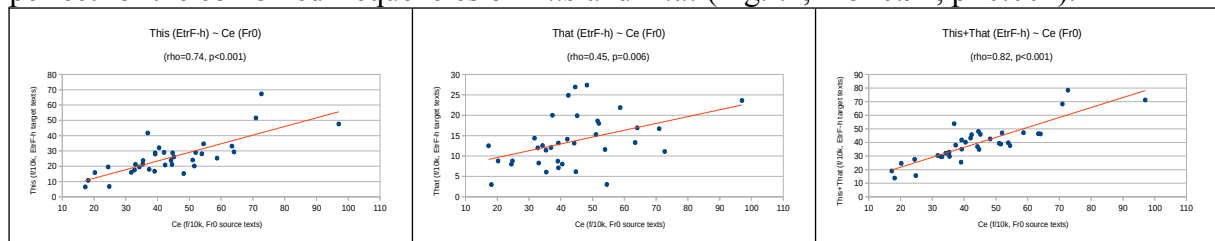


Fig. 7a. *This* in EtrF-h target vs. *Ce* in Fr0 source.

Fig. 7b. *That* in EtrF-h target vs. *Ce* in Fr0 source.

Fig. 7c. *This+That* in EtrF-h target vs. *Ce* in Fr0 source.

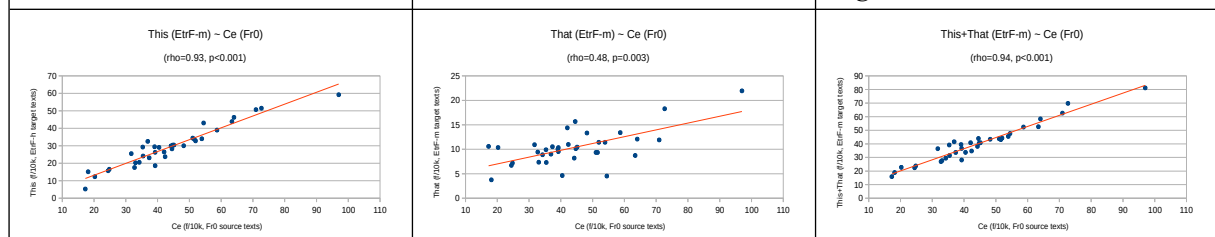


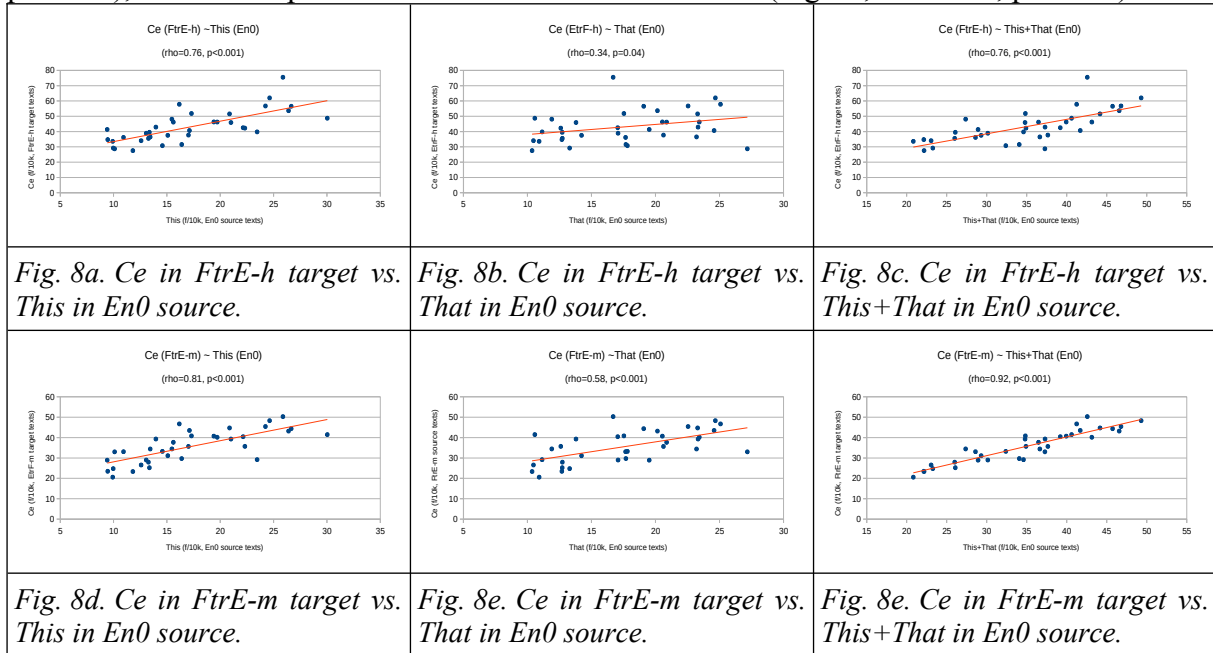
Fig. 7d. *This* in EtrF-m target vs. *Ce* in Fr0 source.

Fig. 7e. *That* in EtrF-m target vs. *Ce* in Fr0 source.

Fig. 7f. *This+That* in EtrF-m target vs. *Ce* in Fr0 source.

In the opposite direction, the frequency of *Ce* in human-translated target texts (FtrE-h) is strongly correlated with the frequency of *This* in the corresponding En0 source texts (Fig. 8a, $\rho=0.76$, $p<0.001$), weakly correlated with *That* (Fig. 8b, $\rho=0.34$, $p=0.04$) and strongly correlated with the combined frequencies of *This* and *That* (Fig. 8c., $\rho=0.76$, $p<0.001$). For machine-translated texts, the correlation between *Ce* and *This* is even stronger than for human

translation (Fig. 8d, $\rho=0.81$, $p<0.001$), moderately strong with *That* (Fig. 8e, $\rho=0.58$, $p<0.001$), and almost perfect with *This* and *That* combined (Fig. 8f, $\rho=0.92$, $p<0.001$).



These tendencies remain essentially the same in plural NPs. Correlations in both directions are strong between *These* and *Ces*, weak or moderate for *Those*, and strongest when the frequencies of *These* and *Those* are combined, with the exception of *Ces* in machine-translated target texts which is more strongly correlated with *These* alone.

3.4. Translation counterparts

Given that correlations between frequencies suggest source text imitation, but do not actually show what happens at the level of individual occurrences, random samples of paired source/target segments (5 per text $\times$ 35 texts = 175 per series/per corpus, total=2450) were taken to determine the proportions of the most common translation counterparts.

3.4.1. *This/That/These/Those*, *Ce/Ces* + *N* (no *Adj*)

When no adjectives are present, human translators (*FtrE-h*) use *Ce* and *Ces* as target counterparts for English demonstratives in 54.3-66.9% of occurrences (Table 2). In machine-translated texts, the correspondence with *Ce* and *Ces* is considerably higher at 76.3-91.4%. The second most common counterpart is the French definite article *Le/La/Les*.

Table 2. Most common target counterparts in *FtrE* for *En0* demonstratives without adjectives.

	Human		Machine	
This	ce:	66.9%[59.9-73.8] ³	ce:	91.4%[87.3-95.6]
	def.art:	7.4%[3.5-11.3]	def.art:	3.4%[0.7-6.1]
	other:	6.9%[3.1-10.6]	adv.:	2.3%[0.1-4.5]
That	ce:	56%[48.6-63.4]	ce:	84.6%[79.2-89.9]
	def.art:	17.1%[11.6-22.7]	def.art:	9.1%[4.9-13.4]

³Percentages are followed by their 95 % confidence intervals.

	omission:	7.4%[3.5-11.3]	omission:	1.7%[0-3.6]
These	ces:	59.4%[52.2-66.7]	ces:	90.9%[86.6-95.1]
	def.art:	10.3%[5.8-14.8]	pos.:	4.6%[1.5-7.7]
	omission:	6.3%[2.7-9.9]	pro.dem/rel:	1.7%[0-3.6]
Those	ces:	54.3%[46.9-61.8]	ces:	76.3%[70-82.6]
	def.art:	14.5%[9.2-19.7]	def.art:	12.1%[7.3-17]
	other:	7.5%[3.6-11.4]	ce:	8.1%[4-12.2]

In the other direction, *Ce* and *Ces* most commonly correspond to *This* and *These*. In human-translated texts, the proportion is around 42% and between 54.6-57.7% for machine translated texts (Table 3). The next most common counterparts are the English definite article *The* and *That/Those*, which both occur in similar proportions, between 10-20%.

Table 3. Most common target counterparts in *EtrF* for *Fr0* demonstratives without adjectives.

	Human		Machine	
Ce	this:	42%[34.6-49.3]	this:	54.6%[47.2-62]
	def.art:	19%[13.1-24.8]	def.art:	15.5%[10.1-20.9]
	that:	15.5%[10.1-20.9]	that:	14.9%[9.6-20.2]
Ces	these:	42.9%[35.5-50.2]	these:	57.7%[50.4-65]
	def.art:	14.9%[9.6-20.1]	those:	18.9%[13.1-24.7]
	those:	14.9%[9.6-20.1]	def.art:	10.3%[5.8-14.8]

3.4.2. *This/That/These/Those*, *Ce/Ces* + *Adj* + *N*

When English demonstratives are used with adjectives, the correspondence with French *Ce/Ces* increases to 61.5-69.7% in human-translated texts and reaches 83.4-95.4% in machine-translated texts (Table 4). Again, the French definite article *Le/La/Les* is regularly the next most common alternative.

Table 4. Most common target counterparts in *FtrE* for *En0* demonstratives with adjectives.

	Human		Machine	
This	ce:	69.7%[62.9-76.5]	ce:	95.4%[92.3-98.5]
	def.art:	10.9%[6.2-15.5]	def.art:	1.7%[0-3.6]
	omission:	5.1%[1.9-8.4]	adv.:	0.6%[0-1.7]
That	ce:	62.3%[55.1-69.5]	ce:	83.4%[77.9-88.9]
	def.art:	22.9%[16.6-29.1]	def.art:	13.7%[8.6-18.8]
	indef.art:	4.6%[1.5-7.7]	other:	1.7%[0-3.6]
These	ces:	61.5%[54-69]	ces:	91.9%[87.7-96.1]
	def.art:	15.5%[9.9-21.1]	pos:	2.5%[0.1-4.9]
	pos.:	3.1%[0.4-5.8]	ce:	1.2%[0-3]
Those	ces:	66.5%[59.1-73.8]	ces:	86.1%[80.7-91.5]

def.art:	16.5%[10.7-22.2]	def.art:	10.1%[5.4-14.8]
omission:	3.8%[0.8-6.8]	pos.:	1.3%[0-3]

In NPs with adjectives, source occurrences of French demonstratives again correspond most commonly to *This/These* (Table 5), with a slight increase in the overall level of correspondence. The position of the adjective, before or after the noun, however, has no discernible effect.

Table 5. Most common target counterparts in EtrF for Fr0 demonstratives with adjectives.

	Human		Machine	
Ce N+Adj	this:	47.7%[40.3-55.1]	this:	63.2%[56.1-70.4]
	that:	21.8%[15.7-28]	that:	19%[13.1-24.8]
	def.art:	17.8%[12.1-23.5]	def.art:	10.3%[5.8-14.9]
Ce Adj+N	this:	50.9%[43.4-58.4]	this:	74%[67.3-80.6]
	def.art:	18.3%[12.5-24.2]	that:	14.8%[9.4-20.1]
	that:	16%[10.5-21.5]	def.art:	8.3%[4.1-12.4]
Ces N+Adj	these:	44.5%[36.9-52.1]	these:	57.9%[50.4-65.5]
	those:	23.2%[16.7-29.6]	those:	28%[21.2-34.9]
	def.art:	13.4%[8.2-18.6]	def.art:	6.1%[2.4-9.8]
Ces Adj+N	these:	40.1%[32.7-47.6]	these:	60.5%[53.1-67.9]
	those:	26.9%[20.2-33.7]	those:	24%[17.5-30.4]
	def.art:	15.6%[10.1-21.1]	def.art:	7.8%[3.7-11.8]

4. Discussion

Demonstrative determiners in Original French are more common than their counterparts in Original English. English-translated-from-French shows a tendency to overuse demonstratives, especially *This* at the expense of *That*, which is amplified in machine translation in comparison with human-translated texts. Conversely, French-translated-from-English would be expected to show a deficit, as indeed occurs in machine-translated texts, but not in human-translated texts. Overuse of *This/These* in EtrF is stronger in NPs containing adjectives. Likewise, the deficit of demonstratives in FtrE is greater in NPs with adjectives and reaches statistical significance in both machine and human-translated texts, albeit with a smaller effect size for human translation and only with postnominal adjectives.

Correlation tests between the frequency of demonstratives in pairs of source and target texts reveal statistically significant correlations in both directions. The combined frequency of *This* and *That* is strongly correlated to the frequency of *Ce* in the source texts (nearly perfect in machine-translated texts), while the frequency of *Ce* in target texts is also strongly correlated with the combined frequencies of *This* and *That* in the source texts. Individually, the correlations between *Ce* and *This* are stronger than between *Ce* and *That*, suggesting a higher degree of correspondence, as is confirmed in the random samples of source/target segment pairs. These findings confirm that source language imitation or “shining-through” (Teich, 2003) is a major factor.

The greatest divergence with respect to target language norms is in English-translated-from-French. Bringing EtrF into closer conformity with the norms observed in Original English (En0) would require decreasing the overall frequency of demonstratives and restoring the balance between *This* and *That*. In practical terms, this would mean translating certain occurrences of *Ce* as *That* instead of *This* or by other alternatives such as the definite article. Conversely, increasing the frequency of demonstratives in French-translated-from-English would involve finding other source counterparts than the English demonstratives.

However, which occurrences would be most eligible for alternative translation cannot be determined from the results of this study alone. A complete analysis would include samples of target occurrences to determine how many can be traced back to demonstratives or other counterparts in the source texts, given that correspondence between source and target counterparts may be asymmetrical (Henkel, 2023). Although overuse of demonstratives in EtrF is clearly due to source language imitation, it remains unclear why translators prefer *This* to *That*. Developing a predictive model using logistic regression, for example, will require tagging individual occurrences for syntactic and semantic variables. Qualitative interpretation of source/target segment pairs could help elucidate the reasons as well, but for lack of space such considerations are far beyond the scope of this paper and must therefore be reserved for future investigation.

5. Conclusion

Previous corpus-based research in translation studies has demonstrated differences between translated and original texts. This is the first study to demonstrate such differences in the translation of demonstratives between English and French. Moreover, these findings reveal differences between human and machine translation, introducing a new dimension in bi-directional corpus analysis. Given that demonstrative determiners are more common in French than in English, human translators exhibit a predictable tendency to overuse demonstratives in English-translated-from-French, as well as an imbalance between *This* and *That*, which are amplified in machine translation. Correlation tests and manual analysis of random samples confirm that these tendencies are largely due to source-language imitation. Both overuse and underuse of demonstratives are more pronounced with NPs containing adjectives. Further research will thus be needed to understand how the semantics of NPs and other contextual parameters influence the translation of demonstratives and help explain the differences observed between translated and original texts.

References

- Dixon, R. M. (2003). Demonstratives: A cross-linguistic typology. *Studies in Language*. International Journal sponsored by the Foundation “Foundations of Language”, 27(1), 61-112.
- Henkel, D. (2023). Verbs of perception and evidentiality in English/French translation. *Languages in Contrast*, 23(2), 161-198.
- Johansson, S. (2007). *Seeing through multilingual corpora*. In *Corpus Linguistics 25 Years on*, R. Facchinetti (ed), 51–72. Amsterdam: Brill Rodopi.
- Kayne, R. S., & Pollock, J. Y. (2010). *Notes on French and English demonstratives*. In *Structure preserved: Studies in syntax for Jan Koster* (pp. 215-228). John Benjamins Publishing Company.
- Teich, E. (2003). *Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts*. Berlin: Mouton de Gruyter.
- R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <<https://www.R-project.org/>>

Beyond Questions: A Corpus-Based Analysis of Interrogatives and Their Responses in Spoken French

Yeo Jun Yun¹, Iris Eshkol-Taravella², Marianne Desmets³

¹ Laboratoire MoDyCo, Paris Nanterre University – yeojun.yun@parisnanterre.fr

² Laboratoire MoDyCo, Paris Nanterre University – ieshkolt@parisnanterre.fr

³ Laboratoire MoDyCo, Paris Nanterre University –desmets@parisnanterre.fr

Abstract

This study aims to examine the functional distribution of interrogatives in spontaneous spoken French, drawing on data from the ESLO 2 corpus, and incorporates into the analysis their relationship to responses. We identify canonical questions, which seek information unknown to the speaker, as well as "non-questions", in which interrogative utterances fulfill other pragmatic aims. The results confirm that many interrogatives do not serve an information-seeking purpose but rather function as resources for interaction management. These findings highlight the importance of dialogic context in the interpretation of interrogative utterances and propose a data-driven functional typology for corpus-based discourse analysis.

Keywords: interrogatives, responses, spoken French, corpus, dialogue, functional typology

Résumé

Cette étude vise à examiner la distribution fonctionnelle des interrogatives en français parlé spontané, à partir des données du corpus ESLO 2, en intégrant dans l'analyse leur relation avec les réponses. Nous identifions les questions canoniques, qui cherchent une information présumée inconnue du locuteur, ainsi que les « non-questions », dans lesquelles les énoncés interrogatifs remplissent d'autres fonctions pragmatiques.

Les résultats confirment que de nombreuses interrogatives ne servent pas à demander de l'information mais fonctionnent plutôt comme des ressources pour la gestion de l'interaction. Ces résultats soulignent l'importance du contexte dialogique dans l'interprétation des énoncés interrogatifs et proposent une typologie fonctionnelle basée sur les données pour l'analyse du discours sur corpus.

Mots clés : interrogatives, réponse, français oral, corpus, dialogue, typologie fonctionnelle

1. Introduction

Interrogative constructions have traditionally been associated with “canonical” questions, that is, information-seeking speech acts in which a speaker requests information presumed unknown to them (Riegel *et al.* 2018).¹ However, research in dialogue theory and interactional linguistics has shown that interrogatives do not systematically correspond to information-seeking acts (Beyssade and Marandin 2009; Eshkol-Taravella *et al.* 2022). Instead, they perform a wide range of interactional functions, including self-addressed questions (1) (Desmets and Gautier 2009), rhetorical uses (2) (Abeillé *et al.* 2014):

¹ In this article, the term *interrogative* is used to refer to utterances displaying interrogative form in the transcription (e.g. utterances ending with a question mark), while the term *question* is reserved for interrogatives whose primary function is to request information.

(1) #spk1 : en Corse bah ça dépend là là cet été **on était où cet été ?**

#spk1 : cet été on fait vers le le Cap Corse vers le haut (ID 98)

#spk1: in Corsica well it depends this summer where were we this summer?

#spk1: this summer we went to Cap Corse up north

(2) #spk2 : et bah ils veulent ça

#spk2 : comment voulez-vous que ça marche ?

#spk2 : c'est pas possible (ID 347)

#spk2: well they want that

#spk2: how do you expect that to work?

#spk2: it's impossible

While this functional diversity is well documented at a theoretical level, quantitative corpus-based studies of interrogatives in spontaneous spoken corpora remain relatively limited (see Stenström 1984 for English, and Ginzburg *et al.* 2022 for English and Polish). For spoken French, such analyses are particularly scarce, and existing work has primarily focused on syntactic variation or discourse description but rarely provides a statistically grounded account of the interaction between interrogatives and their responses (reviewed in Section 2.2). To address this question, we propose a corpus-based statistical analysis of interrogative utterances, based on the hypothesis that interrogative form and intention influence response behavior. Specifically, we address three research questions: (1) Does the interactional intention of an interrogative influence the response type? (2) Which formal features are associated with interactional intention? (3) What factors are associated with semantic alignment between interrogatives and responses?

This study draws on data from the ESLO2 interview corpus, with a manually annotated subset of 491 interrogative utterances (see Section 3). By combining corpus annotation with statistical analysis, this study provides structured empirical data for dialogue modeling, with resources scalable for larger lexicometric investigations.

2. Literature Review

2.1 From sentence type to interactional modeling

Early formal accounts of interrogatives were primarily grounded in sentence-type distinctions. Within classical semantics, interrogative clauses were treated as a distinct grammatical category corresponding to questions understood as sets of possible answers (Hamblin 1973). However, classical speech act theory (Austin 1962; Searle 1969), and later dialogical approaches (Gazdar 1981; Gunlogson 2002; Beyssade and Marandin 2009) established that grammatical form does not always predict illocutionary force. A speaker may use an interrogative to perform requests, suggestions, or indirect assertions. Hence, utterances cannot be fully understood independently of the commitments they create or modify within a conversational context, as it has been modeled in the Dialogue Game Board framework (Ginzburg 2012), and canonical vs. non-canonical questions can be distinguished based on bias and epistemic stance, as proposed by Farkas (2024).

On NLP grounds, we have two annotation frameworks for dialogue analysis based on Searle's model: DAMSL (Core and Allen 1997), that introduces a multi-layered approach, distinguishing forward and backward-looking functions, and explicitly recognizing that interrogative forms may serve functions beyond information-seeking and DIT++ taxonomy

(Bunt, 2009), that provides structured categories for feedback acts and discourse management. We adapted their categories to spoken French interactions.

2.2 Interrogatives in Spoken French

Research on interrogatives in spoken French has primarily focused on structural variation and their discourse usage including subject–verb inversion, *est-ce que* constructions, and *in-situ* interrogatives (Coveney 2011; Druetta 2018; Larrivé and Guryev 2021). Fragment interrogatives in spoken interaction have also received particular attention. Lefeuvre (2021) and other recent corpus-based studies have further highlighted functional diversity : Barbedette and Eshkol-Taravella (2020) proposed an intention typology distinguishing explicit categories (request for information, request for agreement) from implicit ones (opinion, will, doubt), demonstrating that speaker intentions can be automatically classified, and Eshkol-Taravella *et al.* (2022) showed that information-seeking and non-information-seeking questions can be distinguished computationally. However, no study provides a quantitative account of how these functions correlate with observable response behavior in spontaneous dialogue, and the statistical modeling of interrogative–response patterns remains underdeveloped. We address this issue by linking interrogative form and interactional function to empirically observed response types within a unified annotation and statistical framework. Following Farkas (2024) and Trotzke (2023), we adopt the canonical/non-canonical distinction as an empirically testable annotation variable.

3. Data

The data are drawn from the ESLO2 corpus. ESLO corpus (*Enquêtes Sociolinguistiques à Orléans*) is a large-scale spoken French corpus collected in the Orléans region (Baude and Dugua, 2017). ESLO2, the second collection phase (2010s), comprises approximately 400 hours of recordings and 7.5 million words across various speech situations such as guided interview, asking directions, meal conversation, etc. In this study, we focus on the guided interviews. They provide a high density of interrogative-response exchanges, as the interviewer regularly elicits information from the interviewee, particularly suitable for studying the relationship between interrogative form, function, and response behavior. The annotated subset comprises 491 interrogative-response pairs. Questions total 4095 tokens (mean length: 8.3 tokens) with 762 unique forms, of which 452 are hapaxes. Responses, excluding 66 absent cases (without response), total 3231 tokens (mean length: 7.5 tokens) with 707 unique forms and 427 hapaxes.

The corpus was transcribed by human transcribers using Transcriber software, following orthographic conventions without punctuation marks, except for question marks. These are inserted by the transcriber whenever an utterance is interpreted as interrogative, whether the interrogative force is expressed syntactically or conveyed primarily through intonation ([ESLO annotation guide](http://eslo.huma-num.fr/images/TAL/Guide_Annotation-cor1.pdf)² 2013). The ESLO2 transcriptions are distributed as XML files (.trs) containing metadata such as timestamps. In the present study, interrogatives are therefore extracted by selecting turns ending with a question mark by python script. Prior to analysis, the XML files are cleaned to remove elements not relevant to the linguistic analysis, such as timestamps and transcription metadata. This procedure yielded 5847 interrogative utterances. To preserve the interactional context of each interrogative, the dataset includes the ten

² http://eslo.huma-num.fr/images/TAL/Guide_Annotation-cor1.pdf

preceding and ten following turns in the dialogue. The present study focuses on a subset of **491 interrogative utterances**, selected randomly from the extracted material. These interrogatives constitute the basis for the annotation procedure described in the following section.

4. Annotation

The annotation scheme is as follows. Each interrogative was examined within its local interactional context, consisting of the interrogative turn and the surrounding dialogue. Since responses lack formal markers in the transcription, the annotation unit was defined as a single turn: the first turn providing a relevant reaction to the interrogative, identified within the ten subsequent turns. Interrogatives were annotated for **syntactic type** (polar, wh-, alternative), **formal markers** (subject inversion, *est-ce que*, interrogative words, or unmarked), and **fragment status**. Fragmentary interrogatives correspond to utterances that lack canonical sentence structure but convey a full interrogative meaning in context, such as ‘*les films ?*’ in reaction to a preceding turn. Interrogatives were also annotated for the **semantic category** of the requested information (*information demandée*), following the ESLO questionnaire in which questions are organized by thematic domains, and then extended using the named entity categories from QUAERO project (Grouin *et al.*, 2011). The semantic categories include *proposition(oui/non)*, *person*, *location*, *activity*, *function*, *organisation*, *product*, *time*, *quantity*, *language*, *cause*, *manner*, *description*, *repetition*, as well as *others* (a category for other thematic information). Finally, interrogatives were annotated for their **interactional intention**, corresponding to the communicative function of the interrogative within the dialogue. Table 1 presents our typology with illustrative examples.

Intention	Description	Exemple
Information-seeking question (<i>demande d'information</i>)	Canonical interrogative function introducing an information gap and requesting new information (Hamblin 1973)	#spk1: <i>pa- vous faites vous partez où ?</i> (ID 95) #spk1: <i>lea- where did you leave ?</i>
Clarification request (<i>demande de clarification</i>)	Interrogatives used to resolve ambiguity or incomprehension concerning a preceding utterance (Ginzburg 2012)	#spk1: <i>vous avez parlé de mai soixante-huit euh il s'est passé des des choses ici</i> #spk2: <i>comment ?</i> (ID 334) #spk1: <i>you talked about 68 may eh something happened here</i> #spk2: <i>what ?</i>
Confirmation request (<i>demande de confirmation</i>)	Interrogatives seeking validation of information already introduced in the discourse (‘check question’ from Bunt 2009)	#spk1: <i>c'est anonyme</i> #spk2: <i>anonyme ?</i> (ID399) #spk1: <i>it's anonymous</i> #spk2: <i>anonymous?</i>
Request for action (<i>demande d'action</i>)	Directive use of interrogatives where the speaker requests the interlocutor to perform an action (Eshkol-Taravella <i>et al.</i> 2022)	#spk1: <i>Julie , tu peux me passer le sel?</i> #spk2: <i>Tiens</i> #spk2: <i>Julie, can you pass me the salt?</i> spk3: <i>here</i>
Rhetorical question (<i>question rhétorique</i>)	Interrogatives that do not expect an answer but express an assertion (Desmets and Gautier 2009)	#spk2: <i>comment voulez-vous que ça marche ?</i> #spk2: <i>c'est pas possible</i> #spk1: <i>hm hm</i> (ID347) #spk2 : <i>how do you want it to work ?</i> #spk1: <i>it's impossible</i> spk1: <i>hm hm</i>

Phatic question (<i>question phatique</i>)	Questions used primarily to maintain or regulate interaction rather than request information (Boritchev 2021)	#spk2: <i>pour moi c'est un outil</i> #spk2: <i>tu vois ?</i> (ID 411) #spk2: <i>for me it's a tool</i> #spk2: <i>you see ?</i>
Self-addressed question (<i>question à soi-même</i>)	Interrogatives produced by the speaker and directed to themselves, often as part of a thinking-aloud process (Tian et al. 2017)	#spk1: <i>en Corse bah ça dépend là là cet été on était où cet été ?</i> #spk1: <i>cet été on fait vers le le Cap Corse vers le haut</i> (ID 98) #spk1: <i>In Corsica well it depends there there this summer where were we this summer ?</i> #spk1: <i>this summer we went around Cap Corse up north</i>

Table 1. Interaction intention categories

Responses were annotated for their type (see Figure 1), distinguishing between **direct** (directly providing the requested information by the interrogative), **indirect** (engaging without directly providing the requested information), and **absent** (no identifiable response). They were also annotated for the semantic category of information provided (*information fournie*), using the same categories as requested information (*information demandée*). Two additional categories were introduced: *question* when the response consists of a new interrogative, and *ignorance* when the interlocutor explicitly signals lack of knowledge.

id	question	type question	marqueur formelle	fragmentaire	information demandée	Intention	réponse	type de réponse	information fournie
5	#spk1: c'est ça ?	totale	marqueur lexical	non	OUI/NON	demande de confirmation	#spk2: ouais	réponse directe	OUI/NON
24	#spk1: et y a des anciens dans le truc ?	totale	non-marqué	non	OUI/NON	demande d'information	#spk2: on est j- on va dire que c'est une boîte de jeunes parce que	réponse indirecte	organisation
66	#spk2: ouais c'est moins cher hein ?	totale	marqueur lexical	non	OUI/NON	demande de confirmation		absente	

Figure 1. Excerpt from the annotated dataset

Figure 1, excerpt from the annotated dataset, illustrates how the different annotation layers are combined for each interrogative–response pair. Each row includes the interrogative, the identified response, and the associated annotation variables. ID 5 shows a confirmation request followed by a direct and matching response, as the interlocutor explicitly confirms the propositional content of the question ‘*c'est ça ?*’ ID 24 is an information-seeking question with an indirect response, since the interlocutor engaged with the topic does not provide the requested information (yes/no), but introduces new information instead. ID 66 is a confirmation request with no identifiable response (absent), as no subsequent turn can be interpreted as addressing the interrogative

4.1 Annotation reliability

To assess the reliability of the annotation scheme, intra-annotator agreement was measured on a random sample after a one-year interval, and Cohen's Kappa (κ) was calculated for each annotation variable. While inter-annotator agreement is often preferred, intra-annotator

consistency provides a complementary measure of annotation stability (Hovy and Lavid, 2010). The one-year interval between annotations minimizes potential memory effects.

Variable	Agreement	κ	Interpretation
Interrogative Type	93.0%	0.854	Almost perfect
Formal Marker	93.0%	0.888	Almost perfect
Fragment	94.0%	0.831	Almost perfect
Intention	84.0%	0.712	Substantial
Response Type	92.0%	0.850	Almost perfect
Requested Information	89.0%	0.798	Substantial
Provided Information	74.2%	0.667	Substantial

Table 2. Intra-annotator reliability (N = 100, 1-year interval)

Following Landis and Koch (1977), all variables show substantial to almost perfect agreement. Formal properties (interrogative type, formal marker, fragment status) show the highest reliability ($\kappa > 0.83$). Pragmatic and semantic variables show somewhat lower but still substantial agreement ($\kappa = 0.667$ – 0.798), consistent with the greater interpretive complexity of these categories. Disagreements in Intention annotation primarily involved distinctions between confirmation and clarification requests, reflecting the theoretical difficulty of categorizing context-dependent forms such as echo questions (Ginzburg, 2012).

5. Statistical Results

The annotated corpus allows us to conduct a quantitative analysis in order to test our hypothesis that the formulation of interrogatives-both in the form and their interactional intention-influences the type of response they receive in dialogue. To assess the relationships between categorical variables (e.g. intention, response type, formal features), we use chi-square tests of independence. This method allows us to determine whether the observed distributions differ significantly from what would be expected if the variables were independent. The analysis is structured around three dimensions: (i) the relationship between interactional intention and response type, (ii) the association between formal features and interactional intention, and (iii) the factors underlying semantic alignment between interrogatives and responses.

5.1 Interrogative intentions and response types

We first examine whether interrogative intention influences the type of response observed in interaction. Figure 2 presents the distribution of interrogative intentions and response types.

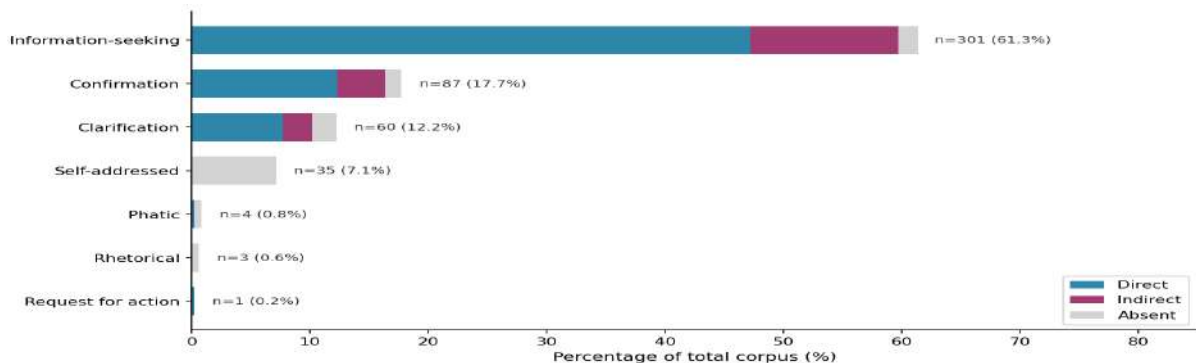


Figure 2. Interrogative intention and response distribution (N = 491)

Canonical information-seeking questions account for 61.3% ($n = 301$), while non-canonical interrogatives represent 38.7%. The response patterns vary systematically across intentions. Information-seeking questions receive direct responses in 77.1% of cases, reflecting their canonical question–answer structure. Confirmation and clarification requests show lower direct response rates (69.0% and 63.3%), with higher proportions of absent responses. This is expected, as these requests do not primarily introduce new information but operate on prior discourse, so their call for direct response should be weaker than that of canonical questions. Self-addressed and rhetorical questions show 100% absent responses, consistent with their non-addressee-directed nature. Due to low frequencies in several categories, the analysis was restricted to the three main intention types ($N = 448$). A chi-square test confirmed a significant association between intention and response type ($\chi^2(4, N = 448) = 20.64, p < .001$). These results support the hypothesis that interactional intention influences response behavior in dialogue.

5.2 Formal features and interactional intention

Given that intention influences response type, we examined which formal features of interrogatives are associated with interactional intention, to understand how form indirectly shapes response behavior. Three formal properties were tested: fragment status, interrogative type, and formal markedness³.

Formal feature	χ^2	df	p	Min expected
Fragment	66.74	2	< .001	18.35
Interrogative type	48.27	4	< .001	2.54
Formal markedness	0.52	2	.773	27.59

Table 3. Formal features and intention: Chi-square results

Fragment status shows a significant association with intention ($\chi^2 = 66.74, p < .001$) with expected frequency above threshold. This indicates that the fragmentarity of interrogatives is closely linked to their interactional function. Interrogative type also shows a significant association, but the minimum expected frequency falls below 5, requiring caution in interpretation. Formal markedness (presence vs. absence of syntactic or lexical markers) shows no significant association with intention ($p = .773$).

Fragment status	Information	Confirmation	Clarification	Total
Non-fragment	241 (77.5%)	53 (17.0%)	17 (5.5%)	311
Fragment	60 (43.8%)	34 (24.8%)	43 (31.4%)	137

Table 4. Fragment status and intention distribution

³ Marked interrogatives include constructions using *est-ce que*, subject–verb inversion, or overt *wh*-words.

Non-fragment interrogatives predominantly serve information-seeking functions (77.5%), whereas fragment interrogatives show a different distribution: only 43.8% are information-seeking, while clarification requests increase substantially (31.4% vs. 5.5%). For example, the fragment *pardon ?* in (3) functions as a clarification request, prompting the interlocutor to repeat their previous utterance:

- (3) #spk2: et toi comment fais-tu ton omelette?
 #spk1: pardon ?
 #spk2: comment fais-tu ton omelette ? (ID 433)
 #spk2: *how how do you make your omelette ?*
 #spk1: *sorry?*
 #spk2: *how do you make your omelette?*

This pattern suggests that fragment interrogatives function primarily as reactive moves in dialogue, often used to address comprehension problems or to seek clarification of preceding utterances.

5.3 Semantic alignment between interrogatives and responses

Following the analysis of response type and interactional intention, we now examine whether interrogatives and responses are semantically aligned. Semantic alignment was assessed by comparing the information requested in the interrogative with the information provided in the response. This allows us to evaluate not only whether a response is produced, but whether it satisfies the informational demand of the question. Cases with no semantic information (e.g. ‘ignorance’, or ‘question’ responses) were excluded from this analysis (N=469).

Factor	χ^2	df	p	Min expected
Interrogative type	4.47	2	.107	6.05
Fragment	1.13	1	.287	44.38
Markedness	8.76	1	.003	79.24
Intention	1.01	2	.603	16.23

Table 5. Factors associated with semantic match: Chi-square results

Among the four factors tested, only formal markedness shows a significant association with semantic match ($\chi^2 = 8.76$, $p = .003$), though with a small effect size (Cramér's $V = 0.137$). Unmarked interrogatives display a higher match rate (71.8%) than marked interrogatives (58.3%). Neither interrogative type, fragment status, nor intention shows a significant association. One possible interpretation concerns the nature of responses provided by marked interrogatives. In the ESLO interview setting, formally marked questions tend to project more elaborate responses. For instance, in (4), the interviewer requests a function label, but the interviewee provides a detailed description, resulting in a semantic mismatch:

- (4) #spk2: c'est taillandier
 #spk1: en quoi exactement ?
 #spk2: un taillandier c'est un forgeron spécialisé dans la fabrication des outils à taillant
 #spk2: *it's a toolsmith*

#spk1: *in what exactly?*

#spk2: *a toolsmith is a blacksmith specialized in the manufacture of cutting tools (ID 342)*

In the guided interview setting, formally marked questions tend to project elaborate responses that go beyond the targeted semantic category, resulting in lower match rates. In contrast, unmarked interrogatives, which rely on intonation and are typically more context-dependent, may elicit more concise and focused responses that correspond more directly to the requested information. However, these results should be interpreted with caution, since the semantic categories used in annotation (e.g., location, activity, time, cause) may be too fine-grained for reliable matching. Future work should develop a revised ontology with broader semantic groupings to better capture the relationship between questions and answers in spoken interaction.

6. Conclusion

This study examined interrogatives in spontaneous spoken French using a corpus-based approach that links form, function, and response behavior. The results show that 38.7% of interrogatives serve non-canonical functions such as confirmation or clarification requests, and that interrogative intention significantly shapes response type. Fragment status emerged as the strongest formal predictor of non-canonical use, while formal markedness was the only factor associated with semantic alignment between questions and answers.

Methodologically, this study contributes a multi-layer annotation framework that integrates formal properties, interactional intentions, and response behavior. This framework offers replicable categories for corpus-based dialogue studies and provides structured empirical data suitable for statistical modeling and NLP-based discourse classification.

Even though the present work constitutes the first steps of a larger work, we have identified interesting trends and provided quantitative evidence, which needs to be validated with data enlargement, and genre diversification. The annotation framework developed here is intended to enable testing more fine-grained hypotheses about how formal, semantic, and interactional factors interact, and is designed to be scalable to larger corpora, contributing to the broader goals of quantitative analysis of spoken language data.

References

- Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.
- Abeillé, A., Godard, D. and Marandin, J.-M. (2014). French questioning declaratives in question. In P. Hofmeister and E. Nordcliff (Eds.), *The Core and the periphery: Data-driven perspectives on syntax inspired by Ivan A. Sag*, Stanford: CSLI Publications.
- Barbedette, A. and Eshkol-Taravella, I. (2020). What speakers really mean when they ask questions. *Proceedings of LREC 2020*, 1159–1166.
- Baude, O. and Dugua, C. (2017). Les ESLO, du portrait sonore au paysage digital. *Corpus* [Online] 15. Online since 15 January 2017. Accessed 20 May 2026 at: <http://journals.openedition.org/corpus/2912>.
- Beyssade, C. and Marandin, J.-M. (2009). Commitment: une attitude dialogique. *Langue française*, 162(2), 89–107.
- Boritchev, M. (2021). *Modélisation dynamique des dialogues*, Thèse de doctorat, Université de Lorraine
- Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In D. Heylen, C. Pelachaud, M. Theune, I. van der Sluis, and A. Nijholt (Eds.), *Proceedings of the AAMAS 2009 Workshop on Functional Markup Language*, Budapest, 13–24.

- Core, M. G., & Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines* (Vol. 56, pp. 28-35).
- Coveney, A. (2011). L'interrogation directe. *Travaux de Linguistique*, 63(2), 112–145.
- Desmets, M., and Gautier, A. (2009). « Comment n'y ai-je pas songé plus tôt ? », questions rhétoriques en comment. *Travaux de Linguistique*, 58(1), 107–125.
- Druetta, R. (2018). Syntaxe de l'interrogation en français et clivage écrit-oral: une description impossible?. In M.-J. Béguelin, A. Coveney, and A. Guryev (Eds.), *L'interrogative en français*, Peter Lang, 19–50.
- Eshkol-Taravella, I., Barbedette, A., Liu, X., and Soumah, V.-G. (2022). Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral. *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*, 305–314.
- D. Farkas (2020). “Canonical and non-canonical questions”. In: University of California at Santa Cruz. Available online: <https://semanticsarchive.net/Archive/WU2ZjIwM/questions.pdf> (accessed on 20 May 2026).
- Gazdar, G. (1981). Speech act assignment. In A. Joshi, B. Webber, and I. A. Sag (Eds.), *Elements of Discourse Understanding*, Cambridge University Press, 64–83.
- Ginzburg, J. (2012). *The interactive stance: Meaning for conversation*. Oxford University Press.
- Ginzburg, J., Yusupujang, Z., Li, C., Ren, K., Kucharska, A., and Lupkowski, P. (2022). Characterizing the Response Space of Questions: Data and theory. *Dialogue and Discourse*, 13(2), 79–132.
- Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. *Proceedings of the 5th Linguistic Annotation Workshop*, 92–100.
- Gunlogson, C. (2002). Declarative questions. *Semantics and Linguistic Theory*, 12, 124–143.
- Hamblin, C. L. (1973). Questions in Montague English. In B. H. Partee (Ed.), *Montague Grammar*, Academic Press, 247–259.
- Hovy, E., and Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1), 13–36.
- Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Larivée, P., and Guryev, A. (2021). Variantes formelles de l'interrogation. Présentation. *Langue Française*, 212(4), 9–24.
- Lefevre, F. (2021). Les interrogatives averbales dans la presse, stratégies discursives récurrentes ? *Langue Française*, 212(4), 107–122.
- Riegel, M., Pellat, J.-C., and Rioul, R. (2018). *Grammaire méthodique du français*. PUF.
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Stenström, A.-B. (1984). *Questions and Responses in English Conversation*. CWK Gleerup.
- Tian, Y., Maruyama, T., and Ginzburg, J. (2017). Self-addressed questions and filled pauses: A cross-linguistic investigation. *Journal of Psycholinguistic Research*, 46(4), 905–922.
- Trotzke, A. (2023). *Non-canonical questions*. Oxford University Press.

The French Drama Revolution: Political Economy and Literary Production, 1700-1900

Thiago Dumont Oliveira¹

¹University of Tartu – thiago@ut.ee

Abstract

This paper investigates the changing nature of French drama between 1700-1900 using Latent Dirichlet Allocation and Jensen-Shannon Divergence. Results indicate that the topical distribution of French drama changed profoundly after the French Revolution, particularly between 1789 and 1850. Bourgeois themes emerged among the most prevalent topics since the late 18th century. To assess the coevolution of drama and economic growth, I plot the yearly prevalence of topics alongside French GDP between 1700-1900, and discuss these changes in light of the political and economic changes prompted by the French Revolution and the industrialization of the country.

Keywords: Political Economy, Cultural History, Computational Linguistics, Latent Dirichlet Allocation, Jensen-Shannon Divergence.

1. Introduction

This paper approaches the evolution of French drama from a *longue durée* perspective (Braudel, 1958), examining literary production through the lens of political economy. I assess which topics grew in importance within French drama between 1700-1900, and discuss these changes in light of the political and economic changes prompted by the French Revolution and the industrialization of the country, which drastically altered French society (Hobsbawm, 2010a,b). Using Latent Dirichlet Allocation and Jensen-Shannon Divergence, I show that the topical distribution of French drama changed profoundly during the ‘Age of Revolution’, as Hobsbawm dubbed the period 1789-1848. Topics containing bourgeois themes emerge among the most prevalent topics since the late 18th century and especially after 1820, with the acceleration of French industrialization.

I use French drama as a window of investigation into the relationship between literature and the social structure. Epistemologically, I take the position that literary production is embedded in political-economic relations (Granovetter, 1985), yet not fully determined by material conditions, since individuals simultaneously shape and are shaped by institutions (Agassi, 1975). The interplay between drama and the political-economic world can thus be approached as an instance of a question long debated among philosophers of science regarding the relationship between agent and structure (Oliveira and Suprinyak, 2018). I concur with Mandelbaum (1955) that individual action cannot be understood without reference to “societal facts”: “the actual behaviour of specific individuals towards one another is unintelligible unless one views their behaviour in terms of their status and roles, and the concepts of status and role are devoid of meaning unless one interprets them in terms of the organization of the society to which the individuals belong” (307-308). My position is that literature is deeply rooted in political-economic relations (or “societal facts”, to use Mandelbaum’s expression), yet literary

production is best approached as “total social facts” (?), and hence not ultimately reducible to political and economic material realities.

Topic modeling has been increasingly used within computational literary studies. Schöch (2017) uses Latent Dirichlet Allocation (LDA) to analyze French Drama of the Classical Age and the Enlightenment using plays from 1630 to 1789. Min and Park (2019) use Non-Negative Matrix Factorization (NMF), sentiment analysis, and network analysis to study Victor Hugo’s *Les Misérables*. Dahllöf and Berglund (2019) combine gender studies and Swedish literary history, using topic modeling to perform a gendered thematic analysis of Swedish classics (1821–1941) and recent bestsellers (2004–2017). Ginn and Hulden (2024) compare LDA and NMF and BERT-type models to examine Roman literature from the founding of Rome to its fall. Martinelli et al. (2024) use neural topic modeling to examine Classical Latin literature. My paper contributes to this literature by drawing on economic history, GDP data, and political economy to discuss the main shifts observable in French drama throughout the 18th and 19th centuries in light of the broader political-economic changes that transformed the country.

2. Data and Methods

My data are 1215 French plays published between 1700 and 1900. The corpus was built using the French-language corpora made available by the DraCor project (Fischer et al., 2019) and it was retrieved via their API (<https://dracor.org/doc/api>). Documents were processed using spaCy’s French language model and normalized through lemmatization because French is a highly inflected language. Preprocessing used a part-of-speech filter to retain only nouns, proper nouns, verbs, and adjectives. French stopwords were removed, as well as highly frequent light verbs (e.g. être, avoir, faire, dire) and address terms (e.g. monsieur, madame). In order to assess the changing nature of French Drama in the 18th and 19th century I use Latent Dirichlet Allocation (LDA) and Jensen-Shannon Divergence (JSD).

LDA is a machine learning method widely used in topic modeling to identify the latent semantic structure of texts. LDA is a generative probabilistic model that models documents as probability distributions over latent topics, while topics are modeled as probability distributions over the words of the vocabulary (Blei, 2012; Blei et al., 2003; Steyvers and Griffiths, 2007). Instead of reading *between* the lines to uncover authors’ purposes, LDA shifts the analysis to reading *above* the lines, such that the topical structure of the texts is derived from the entire constellation of documents to which they belong. It is thus a powerful instrument in approaching the coevolution of literature and economics from a “distant reading” perspective, rather than focusing on the exegesis of a few canonical plays (Moretti, 2013). Moreover, LDA is particularly well suited for my purposes, as it is an effective tool for exploring questions that bridge literary studies and socioeconomic perspectives on culture (DiMaggio et al., 2013).

One of the crucial issues with LDA is the selection of the number of topics, a hyperparameter of the model which must be decided apriori. The optimal number of topics depends on the characteristics of the corpus, on the number of tokens per document, and on the level of granularity adequate to the research question (Jockers, 2013; Rhody, 2012; Sbalchiero and Eder, 2020). Although there are numerous studies on the detection and selection of the number of topics (Gan and Qi, 2021; Griffiths and Steyvers, 2004; Newman et al., 2010; Röder et al., 2015), the optimal number of topics obtained via metrics such as perplexity and topic coherence is hardly a good benchmark, since the appropriate number of topics ultimately depends on topic interpretability and on the level of granularity required by the problem at hand (Wehrheim, 2019). Thus, while LDA is an unsupervised machine learning technique, historical analysis and

a theoretically informed understanding of the object of study remain essential in evaluating the coherence of topics and in selecting the number of topics (Mohr and Bogdanov, 2013).

The LDA model was fit using scikit-learn’s implementation, 30 iterations, and random seed set at 42. Topic prevalence curves and GDP were smoothed using locally weighted scatterplot smoothing (LOWESS). My choices regarding preprocessing and the number of topics were based on topic interpretability and coherence after experimenting with different POS strategies and number of topics. Since my main goal is to trace the rise of economic discourse in French Drama, a 10-topic model proved sufficient in yielding topics with a strong presence of economic discourse and bourgeois everyday life themes, as well as topics that are reminiscent of Greek tragedy and the aristocratic zeitgeist. Models with 20 and 30 topics yielded more fine-grained topics that were of little analytical value for the purposes at hand. Although such topics might be useful for literary scholars interested in identifying subgenres within French drama, from a political economy perspective they added little interpretative value. As a robustness check I also used NMF for the topic modelling analysis, which yielded similar results (see the Appendix).

In order to identify periods of rapid topical change in French Drama I use Jensen-Shannon divergence (Lin, 2002; Nielbo et al., 2023). JSD measures how much the topic distribution of documents changes between two consecutive years. It is defined by

$$\text{JSD}(P \mid Q) = \frac{1}{2}D(P \mid M) + \frac{1}{2}D(Q \mid M)$$

where $M = \frac{1}{2}(P + Q)$ and D is the Kullback–Leibler divergence:

$$D(P \mid Q) = \sum_{i:P_i>0} P_i \log_2 \frac{P_i}{Q_i}$$

The Jensen-Shannon divergence was computed from yearly LDA topic distributions. For each year, document-level topic proportions were averaged across all plays. These yearly distributions were normalized to sum to one, and Jensen-Shannon divergence was then calculated between each pair of consecutive available years using base-2 logarithms. The JSD curve was smoothed using LOWESS.

Lastly, I use multidimensional scaling to build a semantic map from the cosine distance between documents’ topic vectors. Three labels were added on the map, corresponding to three topics of particular interest to my argument, which I further explore throughout the paper. The labels were added at the centroid of the five documents with the highest topic proportions for each topic of interest. These five representative works per topic are listed in Table 1.

3. Results

Table 1 shows the average prevalence of topics between 1700 and 1900, the variation in prevalence over the period ($\Delta_{1700-1900}$), the 15 most important words, and 5 representative works per topic. $\Delta_{1700-1900}$ measures the change in topic prevalence over the period. It was obtained by calculating the slope of an OLS regression for each topic and multiplying it by 200, yielding the overall change of the fitted line over the period.

Topics 0 and 1 barely change over time and their slope is not significantly different from 0. OLS regressions for Topics 2, 3 and 7 yield coefficients that are statistically different from 0, but they are very small. Moreover, these topics are among the least prevalent topics in the table. Topic 6

Topic	Prevalence	$\Delta_{1700-1900}$	Top words (15)	Translation	Representative Works
0	0.07	≈ 0	français, auteur, france, art, talent, goût, ouvrage, paris, public, génie, nature, opéra, sujet, vérité, état	French, author, France, art, talent, taste, work, Paris, public, genius, nature, opera, subject, truth, state	Diderot: <i>Entretien entre d'Alembert et Diderot</i> Anonymous: <i>Ouverture de la Séance</i> Sacy: <i>L'Ile Déserte ou Le Naufrage</i> Diderot: <i>Rêve de d'Alembert</i> Sacy: <i>Le Testament</i>
1	0.13	≈ 0	mère, secret, sentiment, frère, fils, oncle, soeur, lettre, amitié, soin, âme, espérer, tromper, tendre, tendresse	mother, secret, sentiment, brother, son, uncle, sister, letter, friendship, care, soul, hope, deceive, tender, tenderness	Genlis: <i>La Mère Rivale, Comédie</i> Colleville: <i>Sophie et Derville</i> Carmontelle: <i>Les Faux Indifférents</i> Carmontelle: <i>Le Faux Empoisonnement</i> Graffigny: <i>Cénie</i>
2	0.04	+0.05	molière, pièce, théâtre, jouer, comédie, scène, rôle, acteur, comédien, auteur, poète, roi, don, talent, rire	Molière, play, theatre, play, comedy, stage, role, actor, comedian, author, poet, king, gift, talent, laugh	Jouy: <i>Les Bancs de la Promenade</i> Diderot: <i>Paradoxe sur le Comédien</i> Sacy: <i>La Modestie</i> Sacy: <i>La Répétition</i> Chalmeton: <i>A Jean Racine</i>
3	0.07	-0.09	encor, bien, époux, affaire, soin, sot, sage, hymen, tour, discours, fou, diable, rire, loi, vieux	again, good, husband, affair, care, fool, wise, marriage, turn, speech, madman, devil, laugh, law, old	Voltaire: <i>Le Dépositaire</i> Voltaire: <i>L'Enfant Prodigue</i> Regnard: <i>Démocrite</i> Regnard: <i>Les Ménechmes, ou Les Jumeaux</i> Barbier: <i>Le Faucon</i>
4	0.12	+0.37	pauvre, vieux, nuit, soir, roi, mère, fleur, mort, rêve, soleil, voix, peur, âme, bras, souvenir	poor, old, night, evening, king, mother, flower, death, dream, sun, voice, fear, soul, arm, memory	Beissier: <i>L'Oiseau Bleu, Saynète</i> Adenis: <i>Le Nouveau Né</i> Monselet: <i>Par la Poste</i> Crossonnois: <i>Un Monsieur qui ne Veut Plus Fumer</i> Beissier: <i>La Nuit de Noël, Comédie</i>
5	0.10	-0.24	affaire, épouser, mariage, lisette, mari, chevalier, marier, maîtresse, mère, adieu, fâcher, soeur, honnête, marquis, vérité	affair, marry, marriage, Lisette, husband, knight, wed, mistress, mother, farewell, anger, sister, honest, marquis, truth	Dancourt: <i>Colin-Maillard</i> Marivaux: <i>Le Dénouement Imprévu</i> Marivaux: <i>La Commère</i> Audiffret: <i>L'Epreuve</i> Marivaux: <i>La Double Inconstance</i>
6	0.07	-0.13	amant, tendre, doux, feu, aimable, ardeur, charmant, charme, objet, flamme, beauté, léandre, dou, désir, âme	lover, tender, sweet, fire, kind, ardor, charming, charm, object, flame, beauty, Leander, gentle, desire, soul	Voisenon: <i>Hilas et Zélis</i> Saint-Gilles Lenfant: <i>La Feinte Heureuse</i> La Motte: <i>Le Carnaval et la Folie</i> Favart: <i>L'Amant Déguisé</i> Hénault: <i>Le Temple des Chimères</i>
7	0.05	+0.09	mort, fils, terre, mère, ville, parole, demeure, roi, étranger, tuer, frère, sang, vieillard, ennemi, mer	death, son, land, mother, city, word, dwelling, king, foreigner, kill, brother, blood, old man, enemy, sea	Anonymous: <i>Oedipe à Colone</i> Crébillon: <i>Electre</i> Anonymous: <i>Oedipe Roi</i> La Harpe: <i>Philoctète</i> Beissier: <i>Antigone</i>
8	0.18	-0.27	roi, sang, fils, mort, sort, vertu, loi, peuple, crime, gloire, cruel, soin, fureur, encor, horreur	king, blood, son, death, fate, virtue, law, people, crime, glory, cruel, care, fury, again, horror	Chamfort: <i>Mustapha et Zéangir</i> Crébillon: <i>Pyrrhus</i> Voltaire: <i>Saül</i> Chabanon: <i>Eudoxie</i> Crébillon: <i>Atrée et Thyeste</i>
9	0.17	+0.23	argent, diable, affaire, pauvre, payer, mari, maison, franc, voyon, manger, dame, lettre, papa, garçon, boire	money, devil, affair, poor, pay, husband, house, franc, rogue, eat, lady, letter, father, boy, drink	Archambault: <i>Janot</i> Archambault: <i>Janot chez le Dégraisseur</i> Renard: <i>La Demande</i> Gueullette: <i>Le Chapeau de Fortunatus</i> Carmontelle: <i>Les Voyageurs</i>

Table 1: Topic Prevalence, Total Variation (1700–1900), 15 Top Words, Translations and 5 Representative Works.

is an intermediate case ($\Delta_{1700-1900} = -0.13$). While its average prevalence is low (0.07), it was among the most important topics in the early 18th century, although not as important as topics 5 and 8 – see Figure 1, which plots the evolution of the 10 topics over the period.

Topics 4 and 9 stand out as “hot topics” ($\Delta_{1700-1900} = +0.37$ and $+0.23$), whereas 5 and 8 are the “cold topics” ($\Delta_{1700-1900} = -0.24$ and -0.27). The aggregate prevalence of topics 5 and 8 falls from 0.5 in the first decade of the 18th century to 0.07 in the last decade of the 19th century. On the other hand, the aggregate weight of topics 4 and 9 increases from 0.11 to 0.51, making them the most prevalent topics of the late 19th century.

Three topics are of particular interest to my research question, namely topics 4, 8 and 9. Figure 1 shows that topics 4 and 9 grew significantly in importance after 1800, becoming the two most important topics by the end of the century, whereas topic 8 lost considerable weight after 1800. For the ease of interpretation, I label these topics as *Bourgeois Life* (topic 4), *Aristocratic Life* (topic 8) and *Household Economics* (topic 9). Topic 4, and especially topic 9, have a clear material dimension and fall under the umbrella of bourgeois comedies. I chose *Household Economics* as a label for topic 9 because of the presence of words such as *money*, *poor*, *pay*, *husband*, *house* and *franc* among its top 15 words. While the economic dimension is less explicit looking at the top 15 words of topic 4 (except for the word *poor*), an analysis of its representative works clearly indicates that this topic is mostly about the institutions and the social reproduction of the bourgeois family. Topic 8, on the other hand, focuses on questions of virtue, justice, and political authority which are reminiscent of the absolutist state. The top words of this topic include *king*, *virtue*, *law*, *people*, *crime*. This topic thus touches on issues regarding the political organization of the state, in modern parlance, that were central to political philosophers from Plato’s *Republic* and Aristotle’s *Politics* to Machiavelli’s *Prince* and Hobbes’s *Leviathan*.

Figure 2 shows the evolution of these three topics vis-à-vis French GDP. I use real GDP per capita in 2011 US dollars. The data were retrieved from the Maddison Project and are available from <https://www.rug.nl/ggdc/historicaldevelopment/maddison/>. The decreasing importance of *Aristocratic Life* since the 1750s seems to reflect the rise of anti-monarchic sentiment during this period, inspired by Montesquieu’s *The Spirit of Laws* (1748), Rousseau’s *The Social Contract* (1762), and Diderot & D’Alembert’s *Encyclopédie* (1751-1772).

There is a striking similarity between the evolution of GDP and *Household Economics*. Both series grew very slowly during the 18th century (although *Household Economics* oscillated more than GDP), and both series took off between 1825-1850. GDP growth rates increased considerably after 1825 as France initiated its industrialization, with the first railway being completed in 1827 (Saint-Étienne to Andrézieux).

The development of French capitalism thus marks a double shift in literary production: from tragedies to comedies, and from topics that reflect on the political authority of rulers to more mundane themes, reflecting the material reality of the nascent bourgeoisie. Continued growth and prosperity throughout the rest of the century, coupled with the proclamation of the Third Republic in 1870, completely shifted the politico-economic milieu in which authors and readers are socially embedded. The series *Bourgeois Life* is also essentially flat for most of the 18th century, but it considerably grows in importance between the French Revolution and 1850, remaining constant thereafter. Hence, this topic seems less influenced by the economic circumstances than *Household Economics*, and its evolution mirrors instead the changing political context prompted by the French Revolution and the novel political ideas originating from the French Enlightenment.

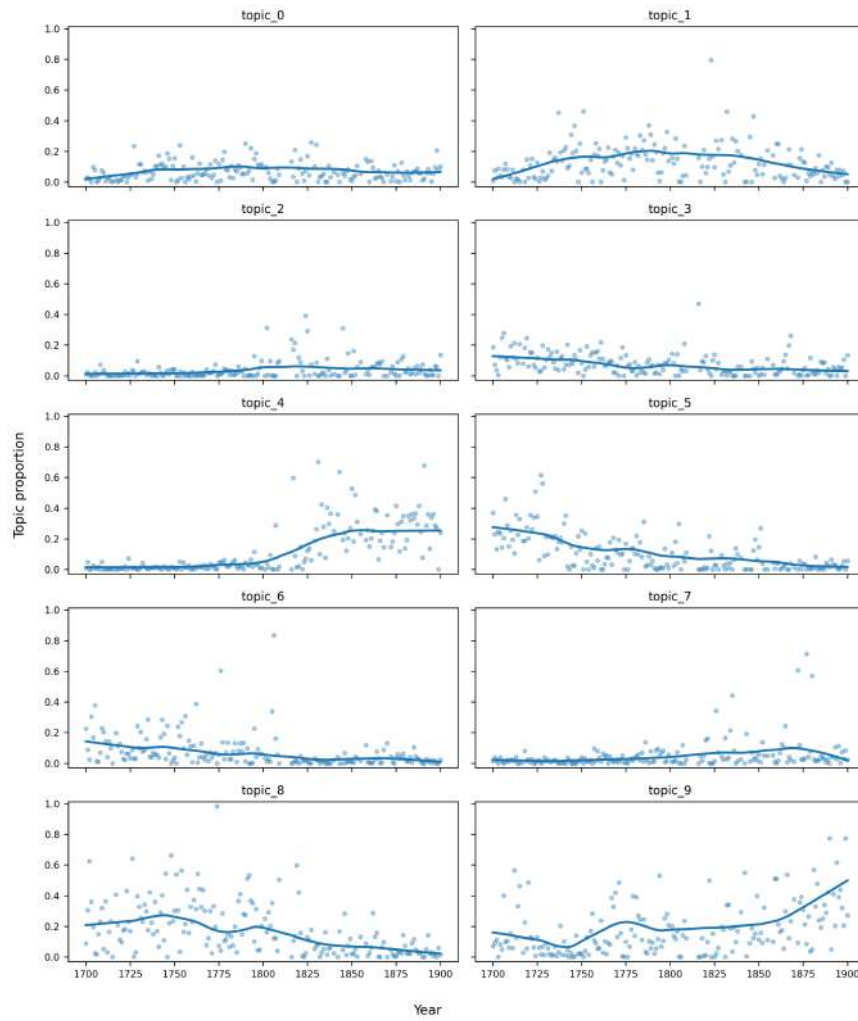


Figure 1: Temporal evolution of the topics

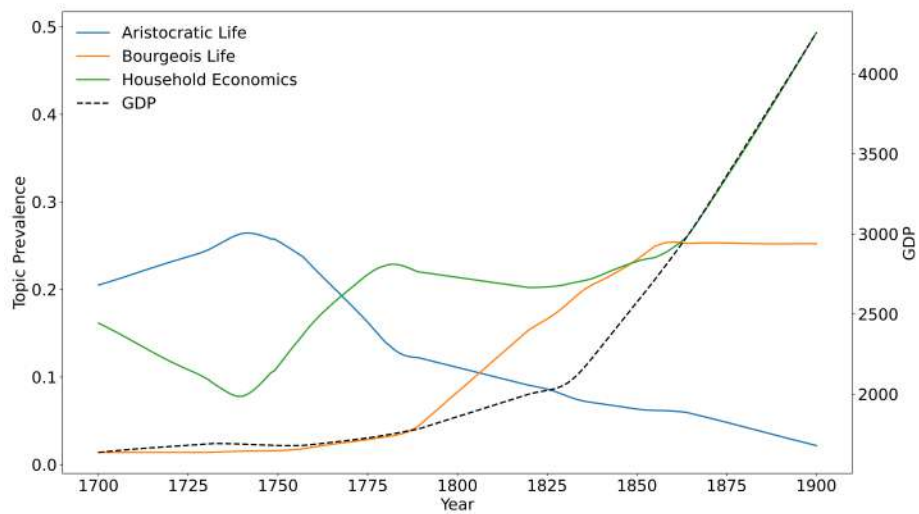


Figure 2: Topic Prevalence vs Real GDP per capita in 2011 US\$

The growing importance of these two series map rather neatly into Hobsbawm's periodization in *The Age of Revolution: 1789–1848* and *The Age of Capital: 1848–1875* (Hobsbawm, 2010a,b). While *Bourgeois Life* grows in importance during the Age of Revolution, but flattens during the Age of Capital, *Household Economics* would only increase in importance in the 1850s, precisely when Hobsbawm identifies the consolidation of capitalism. As French capitalism developed and a bourgeois class emerged in the country, the nature of French drama changed drastically.

Jensen-Shannon Divergence shows that the textual profile of French drama quickly changed in the first half of the 19th century (Figure 3, left). The peak in the JSD series reflects the concomitant decrease in importance of *Aristocratic Life* and increase in *Bourgeois Life* between 1800–1850. French plays shifted toward economic discourse and away from discussions of sovereignty, virtue and violence. This can also be seen by the fact that plays from the second half of the 19th century cluster in the lower left region of the semantic map, indicating that they share a textual profile that distinguishes them from plays from the 18th century (Figure 3, right).

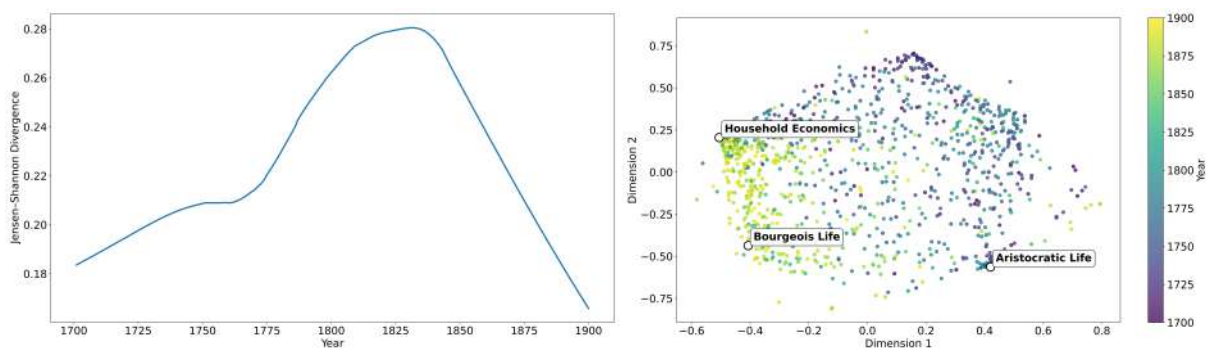


Figure 3: Jensen-Shannon Divergence (left) and semantic map: MDS of cosine distance between LDA topic distributions (right). The labels are at the centroid of the 5 representative documents.

4. Conclusion

This paper has investigated the changing nature of French drama between 1700 and 1900, situating literary production within the broader political-economic transformations which utterly reconfigured French society – and drama, as my results indicate. Applying LDA and JSD to a corpus of 1215 French plays, I have shown that the topical distribution of French drama changed profoundly between the late eighteenth and the mid-nineteenth centuries. In particular, the topic *Aristocratic Life*, which emphasizes sovereignty, virtue, law, and political authority, declined steadily in prevalence over time, while topics *Bourgeois Life* and *Household Economics* gained prominence after the French Revolution.

The thematic profile of French drama radically changed as the absolutist state weakened and capitalist social relations developed. Themes centered on political authority and aristocratic values gave way to more mundane, bourgeois themes, centered on everyday life, domestic relations, and economic concerns. The striking similarity between the series *Household Economics* and *GDP*, and more broadly the coevolution of literary production and economic development, deserve further investigation. More generally, this study illustrates how topic modeling can be used to trace large-scale changes in literary discourse over time, providing a quantitative complement to historical and theoretical approaches to the relationship between literature and political economy.

Acknowledgments

Funded by European Union project "The Center for Digital Text Scholarship" under 566 grant agreement ID 101186601.

References

- Agassi J. (1975). Institutional individualism. *The British journal of sociology*, 26(2):144–155.
- Blei D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.
- Blei D. M., Ng A. Y., and Jordan M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Braudel F. (1958). Histoire et sciences sociales: la longue durée. In *Annales. Histoire, sciences sociales*, volume 13, pp. 725–753. Cambridge University Press.
- Dahllöf M. and Berglund K. (2019). Faces, fights, and families: Topic modeling and gendered themes in two corpora of swedish prose fiction. In *DHN 2019, 4th Digital Humanities in the Nordic Countries 2019, University of Copenhagen, Copenhagen, Denmark, March 6–8, 2019*, pp. 92–111.
- DiMaggio P., Nag M., and Blei D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606.
- Fischer F., Börner I., Göbel M., Hechtel A., Kittel C., Milling C., and Trilcke P. (2019). Programmable corpora: introducing dracor, an infrastructure for the research on european drama. *Digital Humanities*, 2019:5.
- Gan J. and Qi Y. (2021). Selection of the optimal number of topics for lda topic model—taking patent policy analysis as an example. *Entropy*, 23(10):1301.
- Ginn M. and Hulden M. (2024). Historia magistra vitae: Dynamic topic modeling of roman literature using neural embeddings. *arXiv preprint arXiv:2406.18907*.
- Granovetter M. (1985). Economic action and social structure: The problem of embeddedness. *American journal of sociology*, 91(3):481–510.
- Griffiths T. L. and Steyvers M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.
- Hobsbawm E. (2010a). *The Age of Capital: 1848-1875*. Hachette UK.
- Hobsbawm E. (2010b). *Age of revolution: 1789-1848*. Hachette UK.
- Jockers M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
- Lee D. D. and Seung H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791.
- Lee D. D. and Seung H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.
- Lin J. (2002). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.
- Mandelbaum M. (1955). Societal facts. *The British journal of sociology*, 6(4):305–317.
- Martinelli G., Impiccihè P., Fersini E., Mambrini F., and Passarotti M. (2024). Exploring neural topic modeling on a classical latin corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6929–6934.
- Min S. and Park J. (2019). Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PloS one*, 14(12):e0226025.

- Mohr J. W. and Bogdanov P. (2013). Introduction—topic models: What they are and why they matter. *Poetics*, 41(6):545–569.
- Moretti F. (2013). *Distant reading*. Verso Books.
- Newman D., Lau J. H., Grieser K., and Baldwin T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 100–108.
- Nielbo K., Enevoldsen K., Baglini R., Fano E., Roepstorff A., and Gao J. (2023). Pandemic news information uncertainty—news dynamics mirror differential response strategies to covid-19. *Plos one*, 18(1):e0278098.
- Oliveira T. D. and Suprinyak C. E. (2018). The nature and significance of lionel robbins’ methodological individualism. *Economia*, 19(1):24–37.
- Rhody L. M. (2012). Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1), 2(1).
- Röder M., Both A., and Hinneburg A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408.
- Sbalchiero S. and Eder M. (2020). Topic modeling, long texts and the best number of topics. some problems and solutions. *Quality & quantity*, 54(4).
- Schöch C. (2017). Topic modeling genre: an exploration of french classical and enlightenment drama. *Digital Humanities Quarterly*, 11(2).
- Steyvers M. and Griffiths T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis*, pp. 439–460. Psychology Press.
- Wehrheim L. (2019). Economic history goes digital: topic modeling the journal of economic history. *Cliometrica*, 13(1):83–125.

Appendix

As a robustness check to the LDA analysis, I use Non-Negative Matrix Factorization (NMF). Instead of a probabilistic generative model, NMF is a deterministic linear-algebraic method and it is widely used in topic modeling (Lee and Seung (1999), Lee and Seung (2000)). Table 2 shows the 10 topics obtained using NMF. While there are some interesting differences considering the ten topics, the main results remain the same.

Notice that topic 3 is very similar to the topic *Household Economics* previously obtained. Not only are the top words quite similar (*money*, *franc*, etc), but also its average prevalence (0.12) and its prevalence variation (+0.37) resemble that of *Household Economics*. The main difference is that this topic also incorporates words such as *poor*, which were part of topic *Bourgeois Life*, hence my two previous topics that reflect the rise of the bourgeois class were combined into a single topic using NMF instead of LDA.

My other topic of interest, *Aristocratic Life*, also finds a counterpart in topic 1 on Table 2. Notice the presence of words such as *crime*, *people*, *virtue*, *law*, and *tyrant*. Its average prevalence (0.11) and its prevalence variation (-0.17) indicate that this topic was relatively important in the 18th century, but lost importance in the 19th century, similarly to our previous results regarding the topic *Aristocratic Life*.

Topic	Prevalence Δ		Top words (15)	Translation	Representative Works
0	0.12	≈ 0	marquis, oncle, julie, comtesse, chevalier, soeur, sophie, frère, affaire, baron, lettre, mariage, nièce, épouser, sentiment	marquis, uncle, Julie, countess, knight, sister, Sophie, brother, affair, baron, letter, marriage, niece, marry, sentiment	Carmontelle: <i>La Cloison, Beaucoup de Peine pour Rien</i> Carmontelle: <i>Les Moeurs du Temps</i> Carmontelle: <i>La Cloison, Comédie</i> Carmontelle: <i>Condamner sans Entendre, ou La Statue</i> Carmontelle: <i>La Statue</i>
1	0.11	-0.17	sang, fils, mort, fureur, crime, peuple, sort, horreur, cruel, loi, vertu, tyran, gloire, encor, destin	blood, son, death, fury, crime, people, fate, horror, cruel, law, virtue, tyrant, glory, again, destiny	Racine: <i>Polixène</i> Voltaire: <i>Zulime</i> Racine: <i>Bajazet</i> Guimond de La Touche: <i>Iphigénie en Tauride</i> Lebrun: <i>Les Druides</i>
2	0.09	+0.12	molière, théâtre, auteur, pièce, acteur, comédie, poète, scène, talent, comédien, jouer, art, rôle, génie, rire	Molière, theatre, author, play, actor, comedy, poet, stage, talent, comedian, perform, art, role, genius, laughter	Molière: <i>Mascarille</i> Sedaine: <i>Le Quinze Janvier</i> Collé: <i>L'Assemblée</i> Chalmeton: <i>A Jean Racine</i> Bernard: <i>Le Malade Réel</i>
3	0.26	+0.37	pauvre, maman, papa, argent, mari, vieux, franc, diable, vin, manger, soir, maison, lettre, affaire, garçon	poor, mother, father, money, husband, old, franc, devil, wine, eat, evening, house, letter, affair, boy	Labiche: <i>La Rose Rouge</i> Anicet-Bourgeois: <i>Les Aventures de Ninette</i> Courteline: <i>Le Nez du Général Suif</i> Léonce: <i>Le Lapin de la Portière</i> Richepin: <i>Cochon de Cocher !</i>
4	0.13	-0.19	amant, tendre, ardeur, aimable, doux, objet, feu, beauté, flamme, soin, époux, charme, voeu, déesse, léandre	lover, tender, ardor, kind, sweet, object, fire, beauty, flame, care, husband, charm, vow, goddess, Leander	Hénault: <i>Le Temple des Chimères</i> Saint-Gilles Lenfant: <i>La Feinte Heureuse</i> Favart: <i>L'Amant Déguisé</i> Vadé: <i>La Noce de Village</i> Rousseau: <i>Pan et Doris, Pastorale</i>
5	0.03	-0.03	rome, césar, sénat, romain, romains, caton, brutus, sylla, tyran, peuple, sextus, vertu, pompée, sang, tullie	Rome, Caesar, senate, Roman, Romans, Cato, Brutus, Sulla, tyrant, people, Sextus, virtue, Pompey, blood, Tullia	Voltaire: <i>La Mort de César</i> Voltaire: <i>Rome Sauvée</i> Voltaire: <i>La Mort de César</i> Crébillon: <i>Catiline</i> Deschamps: <i>La Mort de Caton</i>
6	0.06	-0.20	lisette, angélique, valère, bian, dorant, frontin, épouser, damis, mariage, lucile, mère, maîtresse, affaire, dorante, argante	Lisette, Angélique, Valère, Bian, Dorante, Frontin, marry, Damis, marriage, Lucile, mother, mistress, affair, Dorante, Argante	Regnard: <i>Crispin, rival de son maître</i> Regnard: <i>Le Neveu de la Marquise</i> Marivaux: <i>L'Épreuve</i> Marivaux: <i>Le Dénouement Imprévu</i> Marivaux: <i>La Mère Confidente</i>
7	0.10	+0.14	fils, mère, mort, zeus, terre, demeure, ville, tuer, frère, parole, oreste, étranger, vieillard, soeur, meurtre	son, mother, death, Zeus, earth, dwelling, city, kill, brother, speech, Orestes, foreigner, old man, sister, murder	Sophocle (tr. Leconte de Lisle): <i>Philoctète</i> Legouvé: <i>Agar dans le Désert</i> Sophocle (tr. Leconte de Lisle): <i>Les Trachiniennes</i> Eschyle (tr. Leconte de Lisle): <i>Électre</i> Eschyle (tr. Leconte de Lisle): <i>Les Choéphores</i>
8	0.04	≈ 0	abbé, chevalier, président, souper, vicomte, pièce, milord, dîner, cheval, opéra, chien, comtesse, carte, bernard, tragédie	abbot, knight, president, supper, viscount, play, milord, dinner, horse, opera, dog, countess, card, Bernard, tragedy	Collé: <i>L'Abbé de Coure-Diner</i> Carmontelle: <i>Les Voyageurs</i> Collé: <i>L'Après-Dinée</i> Carmontelle: <i>L'Histoire</i> Carmontelle: <i>Le Portrait</i>
9	0.06	≈ 0	roi, prince, reine, princesse, peuple, état, trône, cour, ministre, sire, majesté, france, soldat, guerre, couronne	king, prince, queen, princess, people, state, throne, court, minister, sire, majesty, France, soldier, war, crown	Collé: <i>L'Officier du Gobelet</i> Collé: <i>Roi et Reine</i> Regnard: <i>La Méprise</i> Regnard: <i>La Comédie</i> Vadé: <i>Bois-Rosé</i>

Table 2: Topic Prevalence, Total Variation (1700–1900), 15 Top Words, Translations and 5 Representative Works.

From Lyrics to Values: A Computational–Sociological Analysis of Cultural Change in Italian Popular Music

Ciro Clemente De Falco¹, Emilia Romeo²¹

¹University of Naples “Federico II” – ciroclemente.defalco@unina.it

²San Raffaele Roma University – emilia.romeo@uniroma5.it

Abstract

This paper analyses the evolution of cultural values conveyed by the top ten songs of the Sanremo Festival from 1990 to 2026 (370 songs, 1,907 lyrical sections). An LLM-based annotation protocol structured around eight bipolar value axes — with human validation on 10% of sections — serves as training data for a two-stage XLM-RoBERTa classification pipeline. Longitudinal analysis identifies four main patterns: a steady drift towards social conformity (A2), the structural dominance of the affective sphere (A4), the emergence of a melancholic register following the 2008 financial crisis (A3), and an exceptional concentration of value signals in the 2014–2018 period.

Keywords: cultural values; Italian music; Sanremo; NLP; longitudinal analysis; XLM-RoBERTa

Abstract

Il paper analizza l'evoluzione dei valori culturali veicolati dalle prime dieci canzoni classificate al Festival di Sanremo dal 1990 al 2026 (370 canzoni, 1.907 sezioni testuali). Su un corpus annotato tramite LLM secondo un protocollo originale a otto assi valoriali bipolari — con verifica umana sul 10% delle sezioni — viene addestrata una pipeline di classificazione automatica a due stadi basata su XLM-RoBERTa. L'analisi longitudinale identifica quattro pattern principali: una deriva costante verso la conformità sociale (A2), la dominanza strutturale della sfera affettiva (A4), l'emergere di un registro malinconico a partire dalla crisi del 2008 (A3) e una concentrazione valoriale eccezionale nel quinquennio 2014–2018.

Parole chiave: valori culturali; musica italiana; Sanremo; NLP; analisi longitudinale; XLM-RoBERTa

1. Musica, testi e mutamento culturale

La musica occupa una posizione centrale nei processi attraverso i quali le società producono, trasmettono e rielaborano significati collettivi (Madrussan, 2021). Le canzoni costituiscono un artefatto culturale in cui si intrecciano esperienza sociale, industria culturale e immaginario collettivo (Stazio, 2008). In particolare, i testi musicali offrono un accesso privilegiato ciò che, in un determinato contesto storico, appare desiderabile, legittimo, problematico o emotivamente saliente. Già nella tradizione degli studi sui canti popolari, il canto era inteso come espressione intima delle idee del popolo e come deposito di aspirazioni, affetti e rappresentazioni collettive (Busk, 1887). Le canzoni sono archivi culturali dinamici, utili a osservare le trasformazioni delle grammatiche morali e simboliche che attraversano una collettività nel tempo. Anche in Italia la canzone ha storicamente svolto una funzione che eccede la sfera dell'intrattenimento (Bovone e Ruggerone, 2003) costituendo un oggetto culturale di rilevanza storica e sociale, capace di mediare idee, mentalità e immaginari condivisi

¹ La fase di sviluppo del codice di classificazione e le attività di revisione e editing del testo sono state condotte con il supporto di Claude (Anthropic, modello Claude Sonnet 4.6). L'impostazione teorica, il protocollo di annotazione, l'interpretazione dei risultati e le scelte analitiche sono degli autori.

(Caon, 2020). Dalla canzone dell'emigrazione alle musiche della Resistenza, fino ai linguaggi contemporanei segnati da globalizzazione e migrazioni, la musica ha accompagnato alcuni dei principali snodi dell'esperienza storica italiana (Mattioli, 2016), contribuendo alla rappresentazione della vita quotidiana, delle relazioni e dell'identità nazionale. All'interno di questo quadro, il Festival di Sanremo rappresenta un osservatorio particolarmente rilevante. La sua centralità deriva dal ruolo storico di rituale mediatico nazionale, di dispositivo di selezione culturale e di spazio di legittimazione della canzone mainstream (Giannotti, 2005). Proprio per la sua collocazione al crocevia tra industria discografica, televisione, mercato culturale e consumo popolare, Sanremo può essere considerato una proxy ragionata della cultura musicale nazionale-popolare mediatizzata, vale a dire di quell'insieme di repertori simbolici che, in un determinato momento storico, ottengono ampia visibilità, circolazione e riconoscimento pubblico (Berezin, 1998). Nonostante la letteratura abbia già sottolineato come la canzone possa fungere da testimone storico e sociale, e come Sanremo occupi una posizione specifica nella storia della musica popolare italiana del dopoguerra (Facci et al., 2011), resta ancora relativamente poco sviluppata una lettura sistematica del Festival come corpus utile a osservare longitudinalmente l'evoluzione dei repertori valoriali che attraversano il mainstream culturale nazionale. Pertanto, l'obiettivo del presente lavoro è analizzare, in prospettiva diacronica, i repertori valoriali codificati nei testi delle prime dieci canzoni in classifica al Festival di Sanremo dal 1990 al 2026 — segmento più visibile e pubblicamente legittimato del mainstream festivaliero. A tal fine, è stato sviluppato un protocollo di annotazione originale su otto assi valoriali bipolari, applicato all'intero corpus tramite Large Language Model (LLM) con verifica umana sul 10% delle sezioni, e successivamente utilizzato come training set per una pipeline di classificazione automatica a due stadi basata su XLM-RoBERTa. Questo doppio impianto ha consentito sia di produrre un'analisi longitudinale fondata sull'annotazione manuale, sia di costruire uno strumento scalabile per future analisi comparative. Il paper è organizzato come segue. Nel secondo paragrafo viene presentata la metodologia adottata; nel terzo i risultati dell'analisi longitudinale; nell'ultimo le conclusioni e i limiti del lavoro.

2. Metodologia

Il presente studio persegue due obiettivi sequenzialmente connessi. Il primo è ricostruire le configurazioni e le traiettorie dei valori culturali veicolati dalla produzione musicale sanremese nel periodo 1990–2026 attraverso l'annotazione sistematica di un corpus di testi lirici. Il secondo è addestrare un sistema di classificazione automatica sui dati annotati, al fine di rendere scalabile la rilevazione dei valori culturali a corpus più ampi. I due obiettivi si sostengono a vicenda. Le annotazioni manuali prodotte nella prima fase costituiscono simultaneamente la base empirica dell'analisi sostantiva e il training set per il classificatore automatico. La validazione del classificatore consente di valutare in che misura le predizioni automatiche siano in grado di riprodurre i pattern identificati dall'annotazione manuale, fornendo uno strumento per future estensioni della ricerca.

2.1 Il corpus

Il corpus analizzato è composto dalle prime dieci canzoni classificate ogni anno al Festival della Canzone Italiana di Sanremo, per il periodo 1990–2026. La scelta di Sanremo come fonte del corpus è il risultato di una valutazione comparativa rispetto ad alternative disponibili. In una fase preliminare è stata considerata la classifica annuale della Federazione Industria Musicale Italiana FIMI dei singoli più venduti. Tuttavia, la separazione tra chart dei singoli e degli album non consentiva di definire criteri univoci di selezione dei brani. Il Festival di Sanremo, invece, adotta un formato concorsuale con canzoni identificabili in modo univoco e una classifica finale sistematica e replicabile, costituendo al contempo il principale evento della musica popolare

italiana. L'arco temporale considerato copre 37 anni e include 370 canzoni. I testi sono stati raccolti da fonti specializzate (Genius, AZLyrics) e sottoposti a verifica manuale. Per ciascuna canzone, il testo è stato segmentato in sezioni strutturali distinte — strofe (S), ritornelli (CH), bridge (BR), pre-chorus (PCH), intro (INT) e outro (OUT) — seguendo la struttura musicale effettiva del brano. Ogni sezione costituisce l'unità di analisi del presente studio. Il corpus risultante ammonta a 1.907 sezioni testuali, con una lunghezza media di 48 parole per sezione.

2.2 Il protocollo di annotazione

Per l'analisi dei valori culturali è stato sviluppato un protocollo di annotazione originale, strutturato attorno a otto assi bipolari. Ciascun asse rappresenta una dimensione valoriale rilevante per lo studio del cambiamento culturale nella società italiana contemporanea, operazionalizzata come tensione tra due orientamenti contrapposti. Gli assi sono i seguenti:

1. **A1** – *Status/Successo/Consumo* ↔ *Semplicità/Anti-materialismo*: celebrazione della ricchezza e del successo materiale vs. valorizzazione della semplicità e rifiuto del consumismo.
2. **A2** – *Autonomia/Indipendenza* ↔ *Conformità/Aspettative sociali*: affermazione dell'identità individuale vs. adesione a ruoli, tradizioni e norme condivise.
3. **A3** – *Edonismo/Piacere* ↔ *Controllo/Disciplina/Responsabilità*: celebrazione del piacere immediato e degli eccessi vs. valorizzazione dell'autocontrollo e del sacrificio.
4. **A4** – *Cura/Appartenenza* ↔ *Isolamento/Strumentalità relazionale*: relazioni autentiche e senso di comunità vs. solitudine, chiusura o relazioni superficiali.
5. **A5** – *Fiducia civica/Bene comune* ↔ *Cinismo istituzionale*: fiducia nelle istituzioni e nel cambiamento collettivo vs. sfiducia e disincanto sistemico.
6. **A6** – *Partecipazione/Impegno sociale* ↔ *Privatismo*: attivismo e responsabilità collettiva vs. ripiegamento nel privato e disinteresse per il sociale.
7. **A7** – *Sicurezza/Ordine* ↔ *Rischio/Trasgressione*: valorizzazione della stabilità e della protezione vs. celebrazione del rischio e della rottura delle regole.
8. **A8** – *Rispetto/Parità di genere* ↔ *Oggettivazione/Tradizionalismo di genere*: rappresentazione paritaria e autonoma vs. stereotipi di genere e relazioni asimmetriche.

Per ciascun asse, l'annotatore ha compilato tre campi per ogni sezione: a) valore dell'asse, codificato su scala nominale con quattro categorie: 0 = nessun polo presente, 1 = solo Polo A presente, 2 = solo Polo B presente, 4 = entrambi i poli presenti (ambivalente/tensione interna); b) stance del testo rispetto al valore presente, codificata come 0 = n/a, 1 = endorsement, 2 = critica/distanza, 3 = ironico/ambivalente; c) evidenza testuale a supporto della codifica (uno o due versi rappresentativi), utilizzata per documentare le configurazioni linguistiche e discorsive associate ai diversi orientamenti valoriali.

Da queste annotazioni viene derivata automaticamente la variabile *axis_score*: -1 se valore = 1 (Polo A), +1 se valore = 2 (Polo B), amb se valore = 4 (ambivalente, escluso dalla media annuale), 0 se valore = 0 (assente).

Nel presente studio la variabile *stance*, pur annotata sistematicamente, non è stata impiegata nella pipeline di classificazione automatica né nell'analisi longitudinale principale. Tuttavia, costituisce una risorsa disponibile per approfondimenti futuri sull'orientamento normativo dei testi rispetto ai valori rappresentati. L'annotazione è stata condotta a livello di sezione per catturare le tensioni valoriali interne ai singoli brani. Il corpus è stato interamente annotato tramite un modello di linguaggio di grandi dimensioni (Claude, Anthropic), seguendo il protocollo operativo descritto, con istruzioni dettagliate per ciascun asse e casi limite esplicitati. Al fine di verificare la qualità delle annotazioni automatiche, un campione pari al 10% del corpus (circa 190 sezioni, selezionate casualmente) è stato sottoposto a revisione manuale da

parte di un annotatore umano esperto. Il tasso di accordo con le annotazioni LLM è risultato pari al 97% dei casi, fornendo un'indicazione positiva sulla validità e coerenza del processo di annotazione.

2.3 La pipeline di classificazione automatica

Come anticipato la pipeline persegue due obiettivi distinti. Il primo è validare la coerenza interna delle annotazioni: un modello addestrato su una parte del corpus deve essere in grado di riprodurre le etichette assegnate manualmente sul resto, fornendo una misura di quanto il protocollo di annotazione sia stato applicato in modo coerente e generalizzabile. Il secondo obiettivo è costruire uno strumento applicabile a corpus non annotati.

2.3.1 Pipeline a due stadi

Due approcci preliminari hanno mostrato performance insufficienti — embeddings MiniLM con Logistic Regression multiclasse (F1-macro 0,44) e *fine-tuning* XLM-RoBERTa multiclasse (F1-macro 0,35) — a causa dello sbilanciamento estremo delle classi. Sulla base di queste evidenze, è stata adottata una pipeline a due stadi che decompone il problema in sottoproblemi più trattabili.

Stadio 1 – rilevamento della presenza (classificazione binaria). Per ciascun asse, un modello xlm-roberta-base *fine-tuned* predice se il valore è presente (1) o assente (0) nella sezione, indipendentemente dal polo. L'aggregazione dei casi positivi (Polo A + Polo B + ambivalente) consente di bilanciare il training set e di disporre di esempi sufficienti per l'apprendimento. Per compensare lo sbilanciamento residuo, vengono applicati class weights bilanciati e *threshold tuning*. La soglia di decisione ottimale (invece del default 0,5) viene determinata su un *validation set* interno per ogni *fold*, massimizzando l'F1-macro.

Stadio 2 – identificazione del polo (classificazione di polo). Esclusivamente sulle sezioni classificate come positive dallo Stadio 1, un secondo modello xlm-roberta-base *fine-tuned* predice quale polo è dominante (Polo A vs. Polo B). Per l'asse A4 (Cura↔Isolamento), che presenta 170 casi ambivalenti, è stata aggiunta una terza categoria. L'addestramento sui soli positivi consente di operare su dataset più bilanciati e di evitare la distorsione indotta dalla classe dominante assente.

Per entrambi gli stadi, il modello utilizzato è xlm-roberta-base (Conneau et al., 2020), addestrato con learning rate 2×10^{-5} , batch size 16, 3-4 epoche, *warmup ratio* 10% e gradient clipping a 1.0. La scelta di XLM-RoBERTa è motivata dalla sua robustezza su testi italiani informali, che risulta superiore rispetto a modelli addestrati prevalentemente su testi formali come dbmdz/bert-base-italian-cased.

2.4 Validazione della pipeline

La pipeline è stata validata attraverso una strategia di cross-validation stratificata a 5 *fold*, applicata separatamente a ciascuno stadio. Per ogni asse sono stati calcolati F1-macro e F1-weighted come metriche principali; l'F1-macro è preferito in quanto pesa equamente tutte le classi, penalizzando i modelli che ignorano le classi minoritarie. La Tabella 1 riassume le performance della pipeline per asse. Per lo Stadio 1 (classificazione binaria), il valore di F1-macro medio è pari a 0,645, con performance che variano da 0,543 (A7) a 0,751 (A5). Per lo Stadio 2 (classificazione di polo, calcolato sui soli positivi), l'F1-macro medio è pari a 0,534, con performance che variano da 0,439 (A5, A6) a 0,770 (A1). La variabilità tra gli assi riflette anzitutto differenze nella numerosità dei casi positivi nel corpus: A4 è presente nel 72% delle sezioni (1.373 su 1.907), A2 nel 33% (629 sezioni), A3 nel 18% (345 sezioni), mentre A1, A5, A6, A7 e A8 si attestano tutti tra il 7% e il 13%. Tale asimmetria è interpretabile sia in termini

di sbilanciamento delle classi — gli assi con meno di 200 esempi positivi risultano strutturalmente più difficili — sia in termini di distinguibilità linguistica dei valori: alcuni orientamenti valoriali tendono infatti a manifestarsi attraverso configurazioni lessicali e registri discorsivi più ricorrenti e riconoscibili di altri.

Tabella 1. Performance della pipeline per asse (5-fold stratified CV)

Asse	F1-macro S1	F1-macro S2	r (trend)	Affidabilità
A1 Status↔Semplicità	0.576	0.770	0.944	Ottima
A2 Autonomia↔Conformità	0.683	0.550	0.865	Buona
A3 Edonismo↔Controllo	0.555	0.605	0.788	Buona
A4 Cura↔Isolamento	0.745	0.499	0.697	Accettabile
A5 Fiducia↔Cinismo	0.751	0.439	0.212	Limitata
A6 Partecipazione↔Privatismo	0.646	0.439	0.611	Accettabile
A7 Sicurezza↔Trasgressione	0.543	0.523	0.382	Limitata
A8RispettoGenere↔Oggettivazione	0.663	0.443	0.433	Limitata

Nota: S1 = F1-macro Stadio 1 (binario); S2 = F1-macro Stadio 2 (polo); r = correlazione di Pearson tra trend manuale e predetto (v. sezione 3.1).

3. Risultati

3.1 Validazione della pipeline sui dati annotati

Al fine di verificare in che misura la pipeline a due stadi sia in grado di riprodurre, a livello aggregato, i trend ricavati dalle annotazioni manuali, per ciascuno degli otto assi, è stato calcolato l'*axis_score* medio annuale — media delle sezioni di ciascun anno, escludendo i casi ambivalenti e assegnando 0 alle sezioni prive di valore — sia a partire dalle annotazioni manuali (*ground truth*) sia a partire dalle predizioni della pipeline. Il confronto tra le due serie temporali è stato quantificato attraverso la correlazione di Pearson (*r*), riportata nella colonna *r (trend)* della Tabella 1. I risultati mostrano una variabilità sostanziale tra gli assi, coerente con le differenze di performance registrate in cross-validation. Gli assi A1 ($r = 0,944$), A2 ($r = 0,865$) e A3 ($r = 0,788$) presentano una corrispondenza eccellente tra trend manuale e predetto: le due serie temporali si sovrappongono quasi perfettamente, indicando che la pipeline cattura in modo affidabile le dinamiche longitudinali di questi valori. L'asse A4 ($r = 0,697$) mostra una corrispondenza accettabile, mentre A6 ($r = 0,611$) si colloca su un livello di accordo moderato. Per gli assi A5 ($r = 0,212$), A7 ($r = 0,382$) e A8 ($r = 0,433$), la corrispondenza tra trend manuale e predetto è più limitata. In questi casi, la pipeline tende a smorzare le oscillazioni del trend reale, producendo serie più piatte. Questo comportamento è attribuibile alla combinazione di bassa numerosità dei positivi e sbilanciamento tra i poli, che porta il classificatore a predire con maggiore cautela i casi non-zero. Un'osservazione metodologicamente rilevante emerge

dall'analisi complessiva. Anche negli assi con performance individuale più bassa (F1-macro), la corrispondenza dei trend aggregati è spesso superiore a quanto ci si attenderebbe. Questo risultato è coerente con la logica dell'aggregazione statistica, gli errori di classificazione individuali, se non sistematicamente orientati in una direzione, tendono a compensarsi nella media annuale. La pipeline è quindi più affidabile per l'analisi longitudinale aggregata che per la classificazione di singole sezioni. \

3.2 Analisi longitudinale dei valori culturali (1990–2026)

L'analisi longitudinale dei valori culturali veicolati dalle prime dieci canzoni di Sanremo consente di osservare l'evoluzione di otto dimensioni valoriali nel periodo 1990–2026. I risultati sono presentati nella Tabella 2, che riporta per ciascun asse e periodo storico l'*axis_score* medio calcolato sia sulle annotazioni manuali sia sulle predizioni della pipeline. Entrambe le stime sono sottoposte a doppio criterio di interpretabilità: test t a un campione ($\mu = 0$, $p < 0,05$) e soglia di rilevanza pratica $|\text{score}| \geq 0,10$. I valori che soddisfano entrambe le condizioni sono riportati senza parentesi; quelli sottosoglia o non significativi compaiono fra parentesi. La presenza di entrambe le stime nella stessa tabella consente di leggere simultaneamente i risultati sostantivi e la loro riproducibilità computazionale.

Tabella 2. Axis score per asse e periodo — ground truth (grassetto) e predetto dalla pipeline (corsivo)

Periodo	A1	A2	A3	A4	A5	A6	A7	A8
	Status↔Sempl.	Auton.↔Conf.	Edon.↔Contr.	Cura↔Isol.	Fiducia↔Cin.	Partec.↔Priv.	Sicur.↔Trasg.	Risp.↔Ogg.
Prima Rep. (1990–1993)	n.s.	-0.15 <i>-0.15</i>	n.s.	-0.50 <i>-0.31</i>	+0.10 <i>(+0.01)</i>	(-0.10) <i>-0.14</i>	n.s.	n.s.
Berlusconismo (1994–2000)	n.s.	-0.24 <i>-0.27</i>	n.s.	-0.44 <i>-0.43</i>	n.s.	n.s.	n.s.	n.s.
Centro-sin. (2001–2007)	n.s.	-0.22 <i>-0.27</i>	n.s.	-0.53 <i>-0.54</i>	n.s.	n.s.	n.s.	n.s.
Crisi (2008– 2013)	(+0.07) <i>+0.10</i>	-0.22 <i>-0.31</i>	-0.18 <i>-0.13</i>	-0.33 <i>-0.37</i>	n.s.	(-0.05) <i>-0.12</i>	n.s.	n.s.
Instab. (2014– 2018)	+0.15 <i>+0.19</i>	-0.40 <i>-0.46</i>	-0.27 <i>-0.21</i>	-0.50 <i>-0.30</i>	n.s.	n.s.	(+0.06) <i>+0.14</i>	-0.12 <i>(-0.01)</i>
Pandemia (2019–2022)	n.s.	-0.30 <i>-0.30</i>	-0.12 <i>-0.12</i>	-0.29 <i>-0.17</i>	-0.10 <i>(+0.04)</i>	n.s.	n.s.	n.s.
Destra gov. (2023–2026)	n.s.	-0.36 <i>-0.38</i>	-0.11 <i>-0.18</i>	-0.22 <i>(-0.04)</i>	n.s.	n.s.	-0.12 <i>(-0.02)</i>	n.s.

Nota: riga superiore = ground truth (annotazione manuale); riga inferiore in corsivo = predetto dalla pipeline. Valori senza parentesi = interpretabili ($p < 0,05$ e $|\text{score}| \geq 0,10$). Valori fra parentesi = sotto soglia o non significativi. n.s. = nessuna

stima interpretabile. Colorazione basata sul ground truth: verde = polo B; ambra = polo A debole/medio; corallo = polo A forte ($\geq 0,40$).

Valori strutturali: A4 e A2

Due assi mostrano traiettorie di lungo periodo robuste, continue e sostanzialmente indipendenti dalla periodizzazione politica: A4 (Cura/Appartenenza ↔ Isolamento) e A2 (Autonomia ↔ Conformità). Entrambi risultano interpretabili in tutti e sette i periodi, in entrambe le stime, e mantengono la stessa direzione dall'inizio alla fine del corpus.

A4 è l'asse dominante per eccellenza, attivo nel 72% delle sezioni. I valori si mantengono costantemente negativi con *ground truth* compreso tra $-0,22$ e $-0,53$: la cura, i legami affettivi e il senso di appartenenza dominano in modo schiacciante la produzione sanremese per tutto l'arco temporale. Il picco più basso si registra nel periodo 2001–2007 ($-0,53$), mentre i valori meno negativi si osservano durante la pandemia ($-0,29$) e nell'ultimo periodo ($-0,22$), suggerendo una lenta attenuazione della dominanza assoluta della cura. Il trend di lungo periodo è moderatamente positivo ($+0,007$ punti/anno): la sfera affettiva resta il nucleo semantico invariante della canzone sanremese, ma le sue rappresentazioni diventano progressivamente meno totalizzanti. Dal punto di vista linguistico, questo asse si manifesta attraverso un lessico della connessione e della prossimità — “amore”, “insieme”, “cuore”, “dentro”, “occhi” — che attraversa l'intero corpus in modo invariante. Tuttavia, il campo semantico che circonda questo nucleo centrale si trasforma nel tempo: se nei primi periodi prevalgono termini di fusione e appartenenza (“anima”, “bisogno”, “respiro”), dal 2008 in poi emergono con crescente frequenza parole di distanza e ambivalenza (“sola”, “niente”, “casa”, “male”), segnalando rappresentazioni della cura progressivamente meno totalizzanti e più introspettive.

A2 mostra il trend di lungo periodo sociologicamente più significativo dell'intera analisi: una deriva costante verso la Conformità di $-0,006$ punti per anno, per uno spostamento complessivo di circa 0,21 punti in trentasei anni. Il *ground truth* passa da $-0,15$ nel primo periodo a $-0,36$ nell'ultimo, attraversando tutti i contesti politici senza inversioni. La musica sanremese ha progressivamente abbandonato le narrazioni di affermazione identitaria autonoma per rappresentazioni aderenti ai ruoli sociali, alla tradizione e alle aspettative condivise, un cambiamento leggibile anche nel lessico dei testi: nei primi decenni del corpus il polo dell'autonomia si esprime attraverso termini come “libertà”, “strada” e formule di non-conformità (“non cambierò”, “diversi da loro”), mentre nei periodi più recenti il polo della conformità introduce parole più specifiche come “tradizione”, “ruoli” e formule di impegno condiviso (“prometto davanti a Dio”), segnalando un'evoluzione da un'identità costruita sulla rottura a una identità costruita sull'appartenenza e sul vincolo. Si tratta di un processo di ri-tradizionalizzazione culturale che precede, accompagna e sopravvive alle trasformazioni politiche del paese. Vale notare che per A2 il polo dell'Autonomia (575 sezioni) supera largamente quello della Conformità (42 sezioni): la deriva conformista è pertanto da attribuire alla progressiva riduzione delle narrazioni di affermazione autonoma più che a un aumento esplicito di contenuti conformisti.

Valori emergenti: A3 e la svolta del 2008

L'asse A3 (Edonismo/Piacere ↔ Controllo/Disciplina) presenta una discontinuità temporale precisa. Non interpretabile nei primi tre periodi (1990–2007), diventa sistematicamente significativo dalla crisi finanziaria del 2008 in poi. Entrambe le stime confermano questa svolta in modo coerente. Negli anni Novanta A3 oscilla intorno allo zero; è solo con la crisi del 2008 che si stabilizza in territorio negativo con valori compresi tra $-0,12$ e $-0,27$, indicando la progressiva dominanza di un registro malinconico, introspettivo e di elaborazione del dolore

rispetto all'estetica edonistica degli anni precedenti. Questo cambiamento è documentato anche a livello lessicale: nei periodi precedenti al 2008 i versi codificati per A3 concentrano termini di energia e piacere vitale ("notte", "voglia", "forza", "balla", "bruciare"), mentre dal 2008 emergono con frequenza crescente termini del registro della caduta e del dolore ("giù", "male"). La parola "notte" attraversa entrambi i periodi ma cambia segno: da contesto del piacere a contesto della sofferenza. La coincidenza temporale con la crisi finanziaria è difficile da ignorare. Il crollo del modello di crescita degli anni Novanta sembra aver prodotto una trasformazione duratura nel registro emotivo della musica mainstream italiana.

La finestra di rottura: il quinquennio 2014–2018

Il periodo compreso tra il governo Renzi e la fine della legislatura si distingue nettamente come l'unica fase in cui si concentrano contemporaneamente quattro assi interpretabili nel *ground truth*: A1, A2, A3 e A4, con l'aggiunta di A8. Non si tratta di un caso, è il momento di massima densità valoriale dell'intero corpus.

In questo quinquennio A2 raggiunge il suo minimo assoluto (−0,40) e A4 ritorna ai valori più negativi (−0,50). Appare per la prima volta A1 (Status ↔ Semplicità) con valore positivo (+0,15). In anni segnati dall'ascesa delle disuguaglianze e dal dibattito sui privilegi, la musica sanremese valorizza la semplicità come risposta simbolica alla crisi, un orientamento che nei testi si esprime attraverso lessico dell'essenzialità e del rifiuto dell'eccesso ("niente", "senza", "pane", "semplice"). Negli anni più recenti, con l'ingresso stabile del rap nel mainstream festivaliero, compare per la prima volta un lessico esplicito dello status ("soldi", "oro", "tuta", "gold") che introduce, pur senza raggiungere la soglia di interpretabilità statistica, una tensione valoriale inedita nel corpus. Emerge anche A8 (Rispetto di genere ↔ Oggettivazione) a −0,12, distribuito su 14 canzoni diverse. È l'unico periodo in cui le rappresentazioni di genere producono un segnale interpretabile e sufficientemente diffuso da non essere attribuibile a singoli artisti. Il 2014–2018 si configura come un momento di rottura rispetto alla relativa omogeneità valoriale degli altri periodi, probabilmente in risposta alla convergenza di crisi politica, mobilità sociale e nuove istanze culturali.

Valori episodici: A5, A6 e A7

Gli assi A5 (Fiducia civica ↔ Cinismo), A6 (Partecipazione ↔ Privatismo) e A7 (Sicurezza/Ordine ↔ Trasgressione) mostrano presenze troppo discontinue per essere interpretate come tendenze culturali sistematiche. A6 non raggiunge mai la soglia di interpretabilità nel *ground truth*. A5 risulta interpretabile solo in due frangenti circoscritti: nella Prima Repubblica (+0,10) e durante la pandemia (−0,10), quest'ultimo caso con un segnale concentrato in sole 6 canzoni su 40, probabilmente influenzato da pochi artisti specifici. A7 raggiunge la soglia solo nell'ultimo periodo (−0,12 nel *ground truth*), segnale interessante ma troppo recente per generalizzazioni; anche nei pochi casi interpretabili, i riferimenti alla sicurezza e alla stabilità emergono prevalentemente attraverso immagini di continuità affettiva e familiare ("una grande casa", "legati per la vita"), più che attraverso rappresentazioni esplicitamente politiche o collettive. La sistematica assenza di questi assi indica che la musica sanremese mainstream mantiene una distanza strutturale dalla sfera pubblica e civica, costruendosi quasi esclusivamente attorno alle dimensioni affettive e identitarie.

Convergenze e divergenze tra ground truth e predetto

Il confronto diretto tra le due stime riportate in Tabella 2 consente di valutare la robustezza dei risultati sostantivi in modo trasparente. La prima osservazione è quella più rilevante: sui valori strutturali — A2 e A4 — le due stime sono quasi perfettamente sovrapponibili in tutti e sette i periodi, sia in termini di direzione che di ordine di grandezza. Per A2 la differenza massima tra

ground truth e predetto in un singolo periodo è di 0,06 punti; per A4 la convergenza è eccellente fino al 2022. Analogamente, per A3 entrambe le stime individuano la stessa svolta temporale (2008) e mantengono valori negativi coerenti per tutti i periodi successivi. Questi risultati confermano che i trend strutturali identificati nell'analisi sono robusti e non dipendenti dalla fonte di stima. Le divergenze più marcate si concentrano in tre casi isolati, tutti già segnalati come assi con correlazione di trend più bassa. A4 nell'ultimo periodo (2023–2026) mostra un ground truth interpretabile a $-0,22$ mentre il predetto scende a $-0,04$, non interpretabile. La pipeline sottostima la persistenza della dominanza della cura nel periodo più recente. A5 durante la pandemia mostra il caso più netto di divergenza: il ground truth raggiunge la soglia a $-0,10$ (cinismo) mentre il predetto è in territorio positivo ($+0,04$); tuttavia, come già osservato, questo segnale è concentrato in pochissime canzoni e la sua interpretabilità sostantiva è già di per sé limitata. A8 nel quinquennio 2014–2018 è interpretabile nel ground truth ($-0,12$) ma non nel predetto ($-0,01$): la pipeline non riesce a cogliere il segnale sulle rappresentazioni di genere, coerentemente con la sua performance più debole su questo asse ($r = 0,43$).

In sintesi, le divergenze puntuali non alterano le conclusioni sostantive dell'analisi: i quattro pattern identificati — la deriva conformista di A2, l'attenuazione progressiva di A4, la svolta malinconica di A3 e la densità valoriale del quinquennio 2014–2018 — emergono in modo coerente da entrambe le stime. Le differenze si limitano a casi marginali, su assi già classificati come esplorativi, e non modificano né la direzione né l'ordine di grandezza dei trend principali.

4. Conclusioni e limiti

Il presente studio ha analizzato i valori culturali veicolati dalla produzione musicale del Festival di Sanremo nel periodo 1990–2026, attraverso l'annotazione sistematica di 1.907 sezioni testuali estratte dalle prime dieci canzoni classificate per ciascuna edizione. L'analisi longitudinale ha consentito di identificare quattro pattern valoriali distinti, caratterizzati da comportamenti temporali differenti.

Il primo e più robusto è la deriva costante dell'asse Autonomia/Indipendenza ↔ Conformità verso il polo della conformità. Questo dato suggerisce un processo di ri-tradizionalizzazione culturale nella musica sanremese che è indipendente dalla congiuntura governativa e che riflette trasformazioni più profonde nel modo in cui la produzione musicale mainstream rappresenta l'identità individuale: non più come rottura e affermazione autonoma, ma come adesione a ruoli, aspettative e appartenenze condivise. Il secondo pattern è la dominanza strutturale e tendenzialmente invariante dell'asse Cura/Appartenenza ↔ Isolamento con orientamento verso il polo della cura. La sfera affettiva e relazionale costituisce il nucleo semantico stabile della canzone sanremese: ciò che cambia nel corso dei decenni non è la centralità delle relazioni, ma la graduale attenuazione della loro dominanza assoluta, con rappresentazioni progressivamente meno totalizzanti e più ambivalenti soprattutto a partire dal periodo pandemico. Il terzo pattern è una discontinuità temporale precisa sull'asse Edonismo/Piacere ↔ Controllo/Disciplina. Assente o non interpretabile nei primi diciassette anni del corpus, questo asse diventa sistematicamente significativo a partire dalla crisi finanziaria del 2008 e mantiene valori negativi, indicativi di un registro malinconico, introspettivo e di elaborazione del dolore, per tutti i periodi successivi. L'estetica edonistica degli anni Novanta sembra essersi esaurita in coincidenza con la crisi, lasciando il posto a una sensibilità emotiva più cupa e riflessiva. Il quarto pattern è la concentrazione valoriale eccezionale del quinquennio 2014–2018, l'unico periodo in cui quattro o cinque assi risultano simultaneamente interpretabili. In questo frangente, caratterizzato da profonda instabilità politica e da nuove istanze sociali, la produzione sanremese mostra la sua maggiore densità valoriale: oltre ai trend strutturali già in

atto, emergono per la prima volta segnali interpretabili sulla valorizzazione della semplicità (A1) e sulle rappresentazioni di genere (A8).

Sul piano computazionale, la pipeline di classificazione a due stadi ha dimostrato una capacità di riprodurre i trend aggregati del ground truth superiore a quanto suggerito dalle metriche di performance individuale. Per gli assi strutturali (A2, A4) e per quello emergente (A3), le predizioni della pipeline convergono con le annotazioni manuali sia in termini di direzione che di ordine di grandezza, validando il suo utilizzo per future estensioni del corpus.

Sul piano metodologico, tre limiti principali meritano di essere segnalati. L'annotazione è stata condotta tramite LLM con verifica umana su un campione del 10% (98% di accordo), ma in assenza di un calcolo formale dell'accordo inter-annotatore multi-annotatore sull'intero corpus. La pipeline mostra performance differenziate: per A5, A7 e A8 ($r < 0,45$) le predizioni andrebbero trattate con cautela in applicazioni future, a causa della bassa numerosità dei positivi. I trend identificati non sono generalizzabili all'intera produzione musicale italiana né all'intero repertorio delle canzoni in concorso al Festival, poiché il corpus considera esclusivamente le prime dieci classificate, che costituiscono però una scelta metodologica difendibile in quanto rappresentano i contenuti che ottengono la massima legittimazione e visibilità nell'intersezione tra pubblico, stampa e industria. Va tuttavia riconosciuto che il meccanismo di voto composito del Festival — variato nel tempo nella ponderazione tra giuria tecnica, televoto e giuria demoscopica — ha potuto influenzare i repertori valoriali osservati, e che canzoni con testi valorialmente più densi o sperimentali potrebbero essere concentrate nelle posizioni inferiori della classifica. Tra gli sviluppi futuri prioritari figurano pertanto l'estensione del corpus all'intero repertorio in gara — resa oggi possibile dal classificatore validato — e l'analisi comparativa tra canzoni ai vertici e ai margini della classifica finale, al fine di verificare se i pattern identificati siano una caratteristica della produzione premiata o dell'intera offerta festivaliera.

Bibliografia

- Berezin, M. (1998). I rituali pubblici e la rappresentazione dell'identità politica. *Rassegna italiana di sociologia*, vol.(39): 359-386.
- Bovone, L., e Ruggerone, L. (2003). *Intorno a produzione e consumo: percorsi nell'industria culturale* (Vol. 15). FrancoAngeli.
- Busk, R. H. (1887). *The folk-songs of Italy*. S. Sonnenschein, Lowrey & Company.
- Caon, F. (2020). Un genere letterario potenzialmente motivante: la canzone. In Marin T. (a cura di), *Insegnare la Civiltà italiana con la C maiuscola*, Roma, Edilingua
- Facci, S., Soddu, P., e Piloni, M. (2011). *Il Festival di Sanremo: parole e suoni raccontano la nazione*. Carocci.
- Giannotti, M. (2005). *L'enciclopedia di Sanremo: 55 anni di storia del festival dalla A alla Z*. Gremese Editore.
- Madrussan, E. (2021). *Formazione e musica: L'ineffabile significante nel quotidiano giovanile*. Mimesis.
- Mattioli, V. (2016). *Superonda: Storia segreta della musica italiana*. Baldini & Castoldi.
- Midolo, E. D. (2007). *Sound Matters: orizzonti sonori della cultura contemporanea* (Vol. 6). Vita e Pensiero.
- Nobile, S. (2018). Testi musicali e contesti sociali: un'indagine empirica. *Sociologia e ricerca sociale*, vol.(117): 45-66.
- Piccini, M. P. (2016). L'analisi testuale e l'analisi delle corrispondenze lessicali. *Università Pontificia Salesiana-Facoltà di Scienze della Comunicazione Sociale*, vol.(19).
- Stazio, M. (2008). Il futuro alle spalle: canzone napoletana fin de siècle e industria culturale. *Studi sulla canzone napoletana classica*, 379-430.

S05

**Specialized Session
Quantitative Text Analysis Methods for
Scientific Literature**

09-lug
12:00 - 13:30
Aula 11

Mapping the Conceptual Landscape of Funded Computer Science Research in Spain

M.J. Cobo¹, V. Duarte-Martinez², L. Castillo-Chávez³, J.A. Moral-Munoz⁴

¹University of Granada – Spain – mjcobo@decsai.ugr.es

²ESPOL Polytechnic University – Ecuador – vealduar@espol.edu.ec

³Instituto Politécnico Nacional (IPN) – Mexico – lcastilloc2100@alumno.ipn.mx

⁴University of Cadiz – Spain – joseantonio.moral@uca.es

14 maggio 2026

Abstract (in English)

Public research funding plays an important role in shaping the development of scientific fields, yet the thematic structure of funded research is rarely examined from the perspective of the scientific output it generates. This contribution analyzes the conceptual landscape of Spanish computer science research projects funded in the 2019 national call. Funded projects were identified from the awarded grants database of the Spanish Agencia Estatal de Investigación, and their associated scientific publications were retrieved from the Web of Science through funding information. The resulting corpus, comprising 128 funded projects and 1,611 publications, was analyzed through a science mapping workflow implemented in SciMAT. A co-word analysis was applied to detect research themes, which were then represented through strategic diagrams and thematic networks. The results reveal a thematic landscape strongly articulated around artificial intelligence and data-driven research, with deep learning, machine learning, and multi-objective optimization occupying central positions in the conceptual structure. These core themes coexist with other relevant lines related to decision support, fuzzy systems, hardware-oriented research, hyperspectral imaging, recommender systems, virtual reality, and Internet of Things. Overall, the funded research portfolio reflects a broad but coherent landscape, organized around a strong computational intelligence backbone from which multiple specialized and application-oriented directions emerge.

Keywords: science mapping analysis, funded research projects, computer science, SciMAT

1. Introduction

Every year, the Spanish government funds a large number of research projects through competitive calls managed by the *Agencia Estatal de Investigación* (AEI). These calls support a broad range of scientific domains, including computer science, and give rise to a substantial body of scientific output in the form of journal articles and conference papers. From a research policy and science of science perspective (Fortunato et al., 2018), this funded activity can be studied not only at the level of individual projects, but also as a collective research system. Doing so makes it possible to examine what kinds of topics are actually being developed through public funding and, more broadly, to identify the conceptual priorities and directions promoted within a national scientific landscape.

Previous studies have examined publicly funded research through bibliometric approaches, especially by linking grants and publications to analyse productivity, scientific impact, and accountability (Corsini and Pezzoni, 2023; Dzieżyc and Kazienko, 2022; Perianes-Rodríguez et al.,

2024). However, these analyses have focused predominantly on performance-related dimensions, while paying less attention to the thematic landscape and conceptual organization of the research output derived from funded projects. In this sense, there is still room to better understand publicly funded research not only in terms of how much it produces or how much impact it achieves, but also in terms of what it is conceptually about.

Addressing this question is not straightforward. Funded projects usually span multiple sub-areas, employ heterogeneous terminology, and generate a large volume of publications. Under these conditions, manual inspection does not scale well and makes it difficult to obtain a global and interpretable view of the conceptual structure of the corpus. In this context, science mapping analysis (Börner et al., 2003; Cobo et al., 2011b) provides a suitable methodological framework. By means of co-word analysis (Batagelj and Cerinšek, 2013; Callon et al., 1983) and strategic diagrams (Cobo et al., 2011a; He, 1999), science mapping makes it possible to detect thematic areas, represent their relationships, and distinguish between central, developed, transversal, or more peripheral topics within a research corpus.

Against this background, the objective of this contribution is to identify and characterize the thematic areas addressed by Spanish computer science research projects funded in the 2019 national call. More specifically, rather than analyzing project summaries alone, the study focuses on the scientific publications derived from those projects, since these constitute their most visible and consolidated research output. Accordingly, the paper addresses the following research question: what thematic areas are covered by Spanish publicly funded research projects in computer science, as reflected in their associated scientific publications?

To answer this question, a corpus linking funded projects and scientific outputs was constructed. Funded projects were identified from the AEI awarded grants database, and their funding identifiers were subsequently used to retrieve the associated publications from the Web of Science. The resulting corpus was then analyzed through a science mapping workflow implemented in SciMAT. Through co-word analysis, thematic clusters were detected and later represented using strategic diagrams and thematic networks. In this way, the study provides a structured overview of the conceptual landscape of Spanish computer science research projects and illustrates the value of science mapping for examining the thematic organization of publicly funded scientific activity.

2. Methodology

Science mapping or bibliometric mapping is a spatial representation of how disciplines, fields, specialties, and documents or authors are related to one another (Small, 1999). It has been widely used to show and uncover the hidden key elements (documents, authors, institutions, topics, etc.) in different research fields.

Science mapping analysis can be done with different software tools (Cobo et al., 2011b). Particularly, SciMAT was presented in Cobo et al. (2012) as a powerful tool that integrates most of the advantages of available science mapping software tools (Cobo et al., 2011b). SciMAT was designed according to the science mapping analysis approach presented in Cobo et al. (2011a), combining both performance analysis tools and science mapping tools to analyze a research field and detect and visualize its conceptual subdomains (particular topics/themes or general thematic areas) and its thematic evolution.

Therefore, in this contribution, SciMAT was employed to develop a conceptual science mapping analysis (Börner et al., 2003; Cobo et al., 2011b) based on co-words bibliographic networks (Batagelj and Cerinšek, 2013). Thus, the analysis was carried out in three stages:

1. *Detection of the research themes.* In each period studied, the corresponding research themes are detected by applying a co-word analysis (Batagelj and Cerinšek, 2013; Callon et al., 1983) to raw data for all the published documents in the research field, followed by clustering of keywords to topics/themes, which locates keyword networks that are strongly linked to each other, and that correspond to centers of interest or to research problems that are the subject of significant interest among researchers. The similarity between the keywords is assessed using the equivalence index (Callon et al., 1991). Thus, at the end of this phase, the whole co-words network is split into themes (i.e. clusters). Each cluster represents a topic covered in the research field during the time span. The theme comprises a sub-co-words network, formally formed by keywords strongly related, that is, keywords that used to appear together in papers.
2. *Visualizing research themes and thematic network.* In this phase, the detected themes are visualized using two different visualization instruments: strategic diagram and thematic network (Cobo et al., 2011a; He, 1999). Each theme can be characterized by two measures (Callon et al., 1991): *centrality* and *density*. Centrality measures the degree of interaction of a network with other networks. On the other hand, density measures the internal strength of the network. Given both measures, a research field can be visualized as a set of research themes, mapped in a two-dimensional strategic diagram (Figure 1), and classified into four groups:

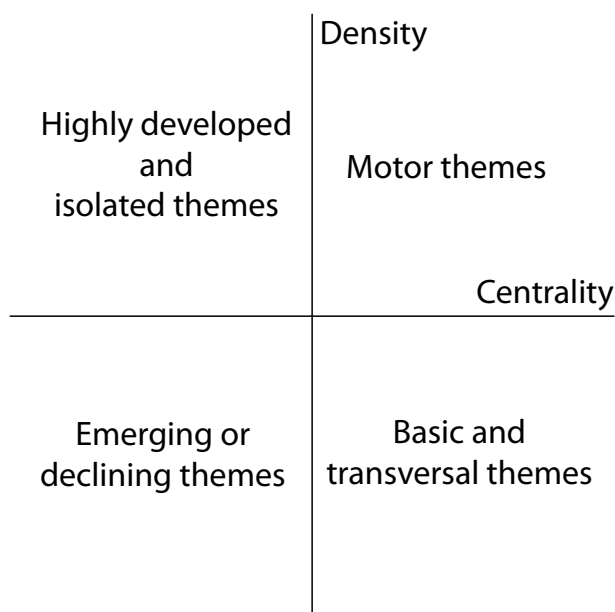


Figure 1: The strategic diagram.

- (a) Themes in the upper-right quadrant are both well-developed and important for the structure of the research field. They are known as the *motor-themes* of the specialty, given that they present strong centrality and high density.
- (b) Themes in the upper-left quadrant have well-developed internal ties but unimportant external ties, so they are of only marginal importance for the field. These themes are very *specialized and peripheral*.
- (c) Themes in the lower-left quadrant are both weakly developed and marginal. The themes in this quadrant have low density and low centrality and mainly represent

either *emerging or disappearing* themes. In a longitudinal framework, whether they are emerging or declining can be determined by observing their position in the diagram over the years. In the absence of such a longitudinal framework, such a determination should be made by expert knowledge.

- (d) Themes in the lower-right quadrant are important for a research field but are not developed. This quadrant contains *transversal and general*, basic themes.

3. *Performance analysis*. In this phase, the relative contribution of the research themes to the whole research field is measured (quantitatively and qualitatively) and used to establish the most prominent, most productive, and highest-impact subfields. Some bibliometric indicators used are the number of published documents, citations, and different types of h-index (Hirsch, 2005). The performance measures for each theme are computed considering the associated documents. Thus, for instance, the h-index is computed using the citations of the theme's documents.

3. Dataset

The dataset used in this study was constructed in two stages. On the one hand, the set of funded research projects was identified from the *Agencia Estatal de Investigación* (AEI). On the other hand, the scientific publications associated with these projects were retrieved from the Web of Science (WoS) through their funding information. This strategy is consistent with previous studies that have used funding acknowledgement data and related metadata to connect research funding with scientific publications, while also noting the technical limitations and quality issues of these data sources (Álvarez-Bornstein and Montesi, 2020; Álvarez-Bornstein et al., 2017; Paul-Hus et al., 2016).

The starting point of the analysis was the 2019 AEI call “*Proyectos de I+D+I (Generación de Conocimiento y Retos Investigación)*”. Within this call, the projects corresponding to the research area “*Tecnologías de la información y de las comunicaciones*” and, the subarea “*Ciencias de la computación y tecnología informática*” were selected. This search yielded a total of 128 funded projects.

Once the project set was defined, the funding identifiers associated with these projects were used to search for the corresponding scientific output in the Web of Science database. The retrieval process was restricted to documents indexed as *article* and *review*, since these document types represent the most consolidated research outputs for the purposes of this analysis. As a result, a total of 1,611 publications were obtained.

4. Conceptual analysis

To analyze the conceptual structure of the scientific output associated with the funded projects, a science mapping analysis based on co-word networks was applied to the whole corpus of publications retrieved from the 2019 AEI call. Since the objective of this contribution is to characterize the thematic landscape of that call as a whole, the corpus was analyzed as a single period. Following the SciMAT methodology, research themes were detected from the co-occurrence of keywords, and then represented by means of a strategic diagram and thematic networks. In addition, performance indicators were computed for each detected theme in order to identify the most productive and influential thematic areas.

According to the strategic diagram shown in Figure 2, the funded scientific output is organized around a heterogeneous set of themes that reflects the breadth of Spanish computer science

research. The most relevant themes are concentrated in the motor and basic quadrants. On the one hand, themes such as *data-models*, *group-decision-making*, *fuzzy-sets*, *recommender-systems*, *virtual-reality*, and *assistive-technology* appear as well-developed and structurally relevant themes, that is, as motor themes of the thematic configuration. On the other hand, *deep-learning*, *machine-learning*, *multi-objective-optimization*, *internet-of-things*, *FPGA*, and *hyperspectral-imaging* are located in the basic and transversal area, suggesting that they play a central role in the field while showing a more open or cross-cutting development. This distribution reveals a conceptual landscape in which artificial intelligence and data-driven topics coexist with optimization, hardware-oriented research, and application-oriented lines.

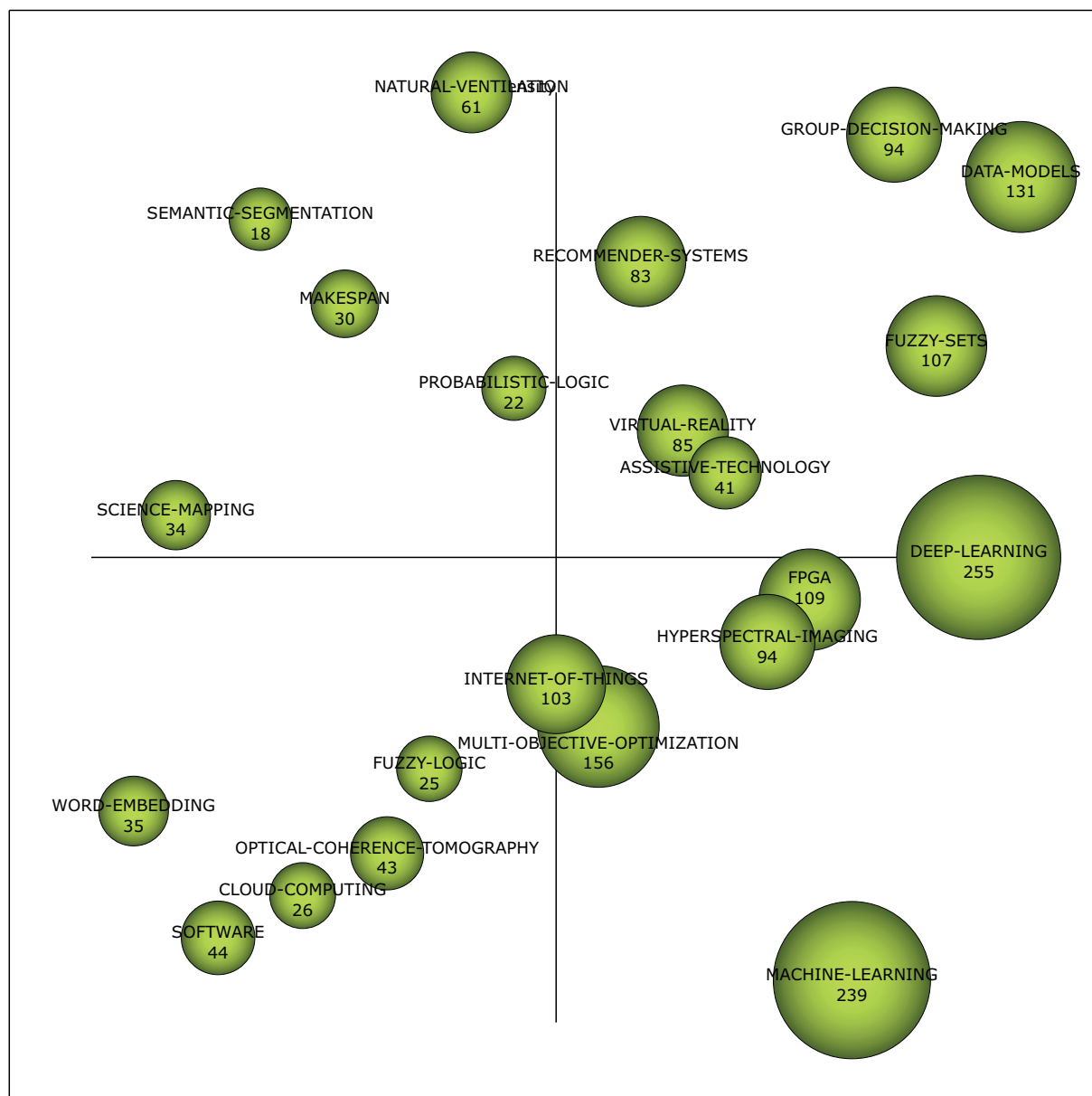


Figure 2: Thematic structure - Strategic diagram.

The performance measures (Table 1) reinforce this interpretation. In terms of production, the largest themes are *deep-learning*, *machine-learning*, *multi-objective-optimization*, *data-models*, *FPGA*, *fuzzy-sets* and *internet-of-things*, which confirms their prominence within the corpus. In terms of scientific impact, *deep-learning* stands out clearly as the leading theme, with 5,369

citations and an h-index of 36, followed by *machine-learning* (4,307 citations; h-index = 31). High impact is also observed in *data-models* (3,391 citations; h-index = 27), *group-decision-making* (3,216 citations; h-index = 30), and *hyperspectral-imaging* (2,682 citations; h-index = 24). Therefore, beyond their conceptual position in the strategic diagram, these results indicate that the corpus is strongly driven by a combination of highly productive artificial intelligence themes and several mature, high-impact specialized areas.

Theme	#documents	#citations	h-index
DEEP-LEARNING	255	5369	36
MACHINE-LEARNING	239	4307	31
MULTI-OBJECTIVE-OPTIMIZATION	156	1904	20
DATA-MODELS	131	3391	27
FPGA	109	1270	18
FUZZY-SETS	107	1835	22
INTERNET-OF-THINGS	103	1301	21
GROUP-DECISION-MAKING	94	3216	30
HYPERSPPECTRAL-IMAGING	94	2682	24
VIRTUAL-REALITY	85	1129	17
RECOMMENDER-SYSTEMS	83	1374	20
NATURAL-VENTILATION	61	1053	21
SOFTWARE	44	1103	12
OPTICAL-COHERENCE-TOMOGRAPHY	43	657	14
ASSISTIVE-TECHNOLOGY	41	428	12
WORD-EMBEDDING	35	455	12
SCIENCE-MAPPING	34	771	14
MAKESPAN	30	352	12
CLOUD-COMPUTING	26	356	9
FUZZY-LOGIC	25	257	8
PROBABILISTIC-LOGIC	22	366	9
SEMANTIC-SEGMENTATION	18	333	10

Table 1: Performance of the themes

The following themes were selected for closer inspection in order to illustrate the diversity of the conceptual landscape, combining central methodological areas, high-impact applied topics, and more isolated lines of research.

Deep-learning (Figure 3a) is one of the most productive and influential themes in the corpus. Its conceptual structure is centered on core deep learning architectures and their application to image-based problems. The thematic network is articulated around subtopics such as *convolutional-neural-networks*, *neural-networks*, *transfer-learning*, *images*, *revised-U-Net*, and *segmentation*, together with application-oriented terms such as *medical-images*, *computer-aided-diagnosis*, and *collaborative-filtering*. Overall, the theme combines foundational deep learning models with their deployment in computer vision and medical image analysis, while also extending toward related intelligent data analysis tasks.

Machine-learning (Figure 3b) is likewise among the largest and most cited themes. Rather than pointing to a narrowly delimited specialty, its conceptual structure reflects a broad methodological space organized around the development and application of machine learning techniques in different problem settings. The thematic network includes subtopics such as *prediction*, *big-*

data, *decision-trees*, and *accuracy*, together with broader terms such as *artificial-intelligence* and *edge-computing*. It also incorporates application-oriented nodes such as *disease*, *graphs*, and *energy-consumption*. Taken together, this configuration portrays machine learning as a transversal methodological area linked to prediction, decision support, and data-intensive applications across diverse domains.

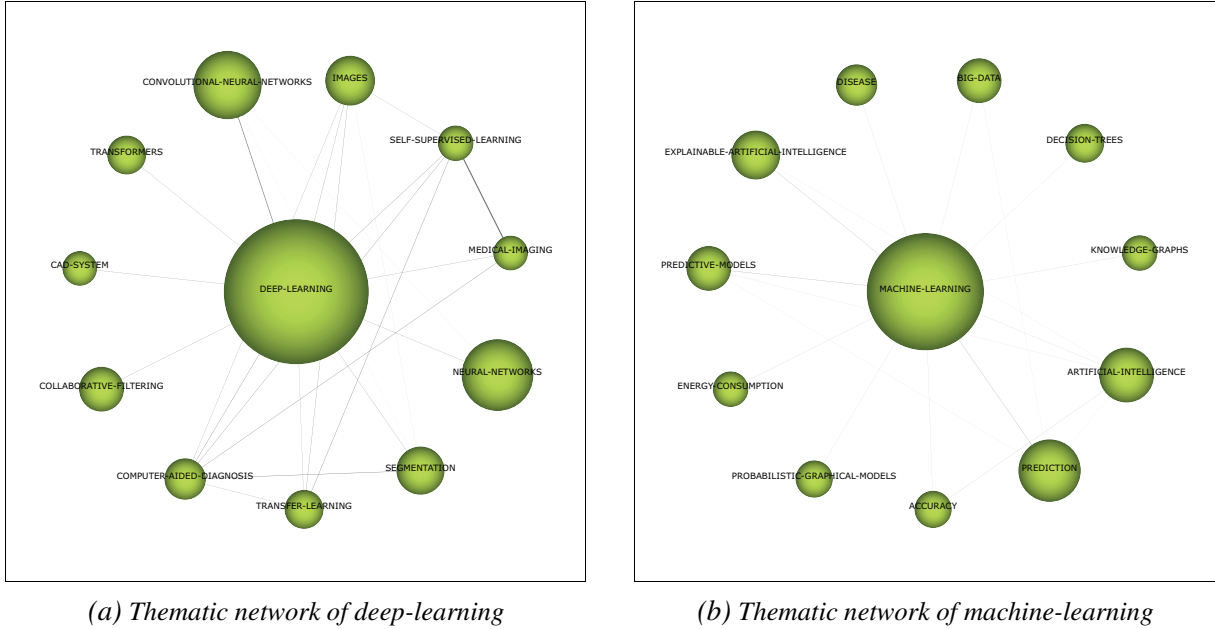


Figure 3: Selected thematic networks.

As a basic theme with notable production and visibility, multi-objective-optimization captures an important methodological line within the corpus. Its conceptual structure is centered on optimization methods designed to handle multiple criteria simultaneously. The thematic network is organized around subtopics such as *genetic-algorithms*, *evolutionary-algorithms*, *feature-selection*, and *selection*, together with related terms such as *binary-codes*, *vector-based* representations, and *performance*. The presence of application-oriented nodes such as *cancer* and *bioinformatics* further indicates that these approaches are not restricted to methodological development, but are also deployed in data-intensive domains. Altogether, the theme reflects a research line focused on evolutionary and metaheuristic solutions to complex optimization problems, especially in feature selection and related applied scenarios.

Internet-of-things (Figure 4a) appears as a productive basic theme with a clear application-oriented profile. Its conceptual structure revolves around connected devices and sensor-based environments, particularly in health and care settings. The thematic network is articulated around subtopics such as *sensors*, *device*, and communication-related terms, together with application-oriented nodes such as *e-health*, *care*, *older-adults*, and *adults*. The presence of terms associated with intervention environments and computing infrastructures suggests an interest in the deployment of IoT solutions in real-world assisted living and healthcare scenarios. In this sense, the theme reflects research on monitoring, communication, and support technologies in human-centered environments.

Among the motor themes, group-decision-making (Figure 4b) stands out especially for its scientific impact. Its conceptual structure is clearly oriented toward collective decision processes and the computational mechanisms that support them. The thematic network is organized around subtopics such as *consensus*, *consensus-reaching*, *social-networks*, *aggregation*, *consistency*,

and *ranking*, together with other terms related to adjustment and checking mechanisms in decision models. Altogether, this configuration points to a line of research focused on supporting and improving collective decision processes, particularly in contexts where preferences must be aggregated, consistency evaluated, and consensus achieved among multiple participants.

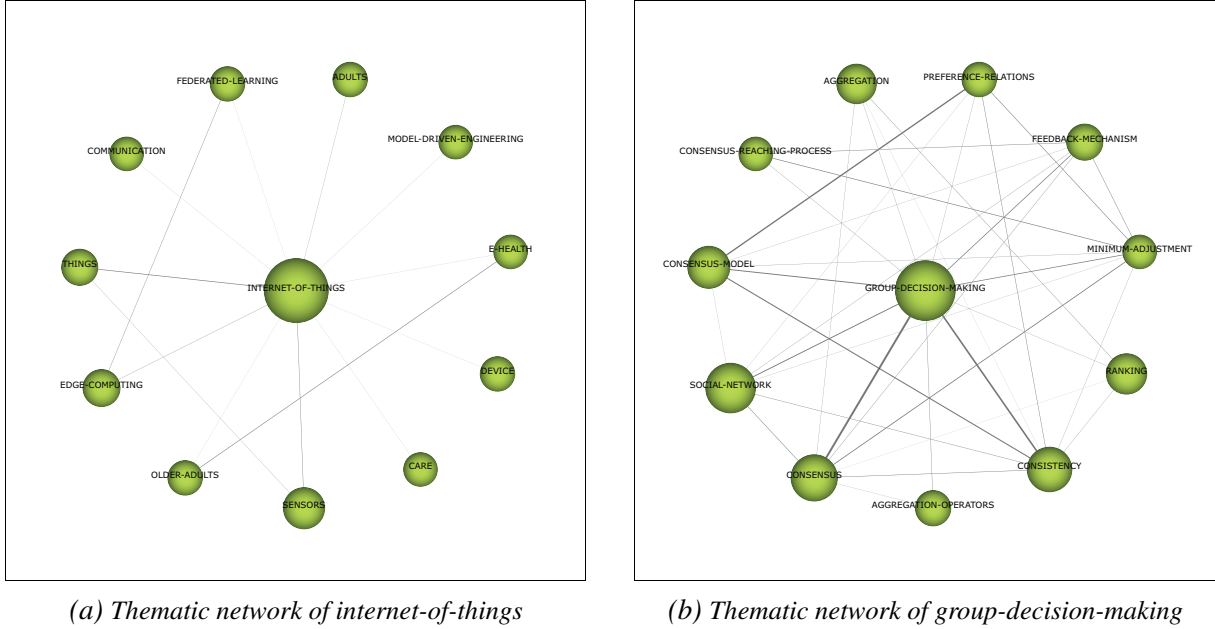


Figure 4: Selected thematic networks.

Hyperspectral-imaging represents one of the most visible applied themes in the corpus. Its conceptual structure is centered on the analysis and processing of hyperspectral images through advanced computational methods for classification and representation. The thematic network includes subtopics such as *kernel* methods, *representation*, *graphs*, *anomaly-detection*, and *super-pixel* approaches, together with terms related to dimensionality reduction and remote sensing. Overall, the theme reflects the use of machine learning and pattern recognition techniques in hyperspectral image analysis, especially for feature representation, classification, and anomaly detection.

Science-mapping is an isolated theme within the corpus and does not belong to the main backbone of the field. It can therefore be interpreted as an example of applied computer science, in which computational and artificial intelligence techniques are used to analyze a specific domain. More precisely, this theme covers the use of computer science and AI methods to discover hidden patterns, structures, and relationships within scientific and technological corpora, including the identification of thematic areas, conceptual structures, and knowledge dynamics through quantitative analysis.

5. Conclusions

This contribution examined the conceptual structure of Spanish computer science research projects funded in the 2019 national call, by analyzing the scientific publications derived from those projects through a science mapping workflow implemented in SciMAT. By approaching the funded projects as a collective research system rather than as isolated cases, the analysis made it possible to identify the main thematic areas emerging from publicly funded activity and to characterize how they are organized within the broader conceptual space of the field.

The results reveal a thematic landscape that is broad but conceptually well articulated. Funded research is strongly organized around an artificial intelligence and data-driven backbone, with deep learning, machine learning, and optimization-related themes occupying central positions in the structure. Around this backbone, the landscape incorporates several mature and visible areas linked to decision support, fuzzy systems, hardware-oriented research, image analysis, recommender systems, virtual reality, and Internet of Things. Rather than suggesting fragmentation, this configuration reflects a coherent research portfolio in which transversal computational paradigms support multiple specialized and application-oriented directions.

This thematic configuration also highlights a defining feature of contemporary computer science: the close interplay between methodological development and domain-oriented application. Part of the funded research is concentrated on enabling computational approaches with wide applicability, while another part is oriented toward the use of those approaches in specific contexts such as health, sensing, optimization, and collective decision making. In this sense, the funded portfolio portrays computer science not only as a field driven by methodological innovation, but also as one with a strong capacity to project those methods across diverse scientific and technological problems. More peripheral or isolated themes should therefore be understood not as marginal in value, but as expressions of the diversity and specialization that complement the main conceptual backbone.

Overall, the study shows that science mapping provides a useful perspective for examining the thematic organization of publicly funded research and for making visible the conceptual priorities reflected in project outputs. In the case of Spanish computer science, the funded landscape appears as a coherent but diverse system, structured around a strong computational intelligence core and extended through multiple specialized lines of application. Future work could expand this approach to additional calls, compare different funding periods, or incorporate longitudinal analyses to study thematic evolution over time.

Acknowledgments

This paper has been part of the project C-ING-165-UGR23, co-funded by the Regional Ministry of University, Research and Innovation and the European Union under the Andalusia ERDF Programme 2021–2027.

References

- Álvarez-Bornstein B. and Montesi M. (2020). Funding acknowledgements in scientific publications: A literature review. *Research Evaluation*, 29(4):469–488.
- Álvarez-Bornstein B., Morillo F., and Bordons M. (2017). Funding acknowledgments in the web of science: completeness and accuracy of collected data. *Scientometrics*, 112(3):1793–1812.
- Batagelj V. and Cerinšek M. (2013). On bibliographic networks. *Scientometrics*, 96(3):845–864.
- Börner K., Chen C., and Boyack K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255.
- Callon M., Courtial J. P., and Laville F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1):155–205.
- Callon M., Courtial J. P., Turner W. A., and Bauin S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2):191–235.
- Cobo M. J., López-Herrera A. G., Herrera-Viedma E., and Herrera F. (2011a). An approach for detecting,

- quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, 5(1):146–166.
- Cobo M. J., López-Herrera A. G., Herrera-Viedma E., and Herrera F. (2011b). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7):1382–1402.
- Cobo M. J., López-Herrera A. G., Herrera-Viedma E., and Herrera F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 63(8):1609–1630.
- Corsini A. and Pezzoni M. (2023). Does grant funding foster research impact? evidence from france. *Journal of Informetrics*, 17(4):101448.
- Dziężyc M. and Kazienko P. (2022). Effectiveness of research grants funded by european research council and polish national science centre. *Journal of Informetrics*, 16(1):101243.
- Fortunato S., Bergstrom C. T., Börner K., Evans J. A., Helbing D., Milojević S., Petersen A. M., Radicchi F., Sinatra R., Uzzi B., Vespignani A., Waltman L., Wang D., and Barabási A.-L. (2018). Science of science. *Science*, 359:eaa0185.
- He Q. (1999). Knowledge Discovery Through Co-Word Analysis. *Library Trends*, 48(1):133–159.
- Hirsch J. E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.
- Paul-Hus A., Desrochers N., and Costas R. (2016). Characterization, description, and considerations for the use of funding acknowledgement data in web of science. *Scientometrics*, 108(1):167–182.
- Perianes-Rodríguez A., Olmeda-Gómez C., Delbianco N. R., and Grácio M. C. C. (2024). Public funding accountability: a linked open data-based methodology for analysing the scientific productivity and influence of funded projects. *Scientometrics*, 129:5841–5868.
- Small H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9):799–813.

Modeling Emotion Constructs in Consumer Reviews: A Word2Vec–ELIta Hybrid Method Combined with PLS-SEM

Cosima Valentino¹, Luca D’Aniello², Massimo Aria², Mario Angelelli¹, Enrico Ciavolino¹

¹University of Salento, Data Science Research Center – cosima.valentino@unisalento.it, mario.angelelli@unisalento.it, enrico.ciavolino@unisalento.it

²University of Naples, Federico II – luca.daniello@unina.it, massimo.aria@unina.it

Abstract

The analysis of consumer reviews has become an important tool for assessing perceived service quality, particularly in public healthcare systems, where traditional satisfaction surveys often fail to capture the emotional dimension of user experience. Although sentiment analysis is widely used to extract evaluative information from textual data, it commonly reduces opinions to simple polarity measures, overlooking the multidimensional structure of emotions expressed in natural language. This study proposes a hybrid methodology for emotion detection in textual reviews aimed at constructing statistically grounded emotional constructs suitable for structural equation modeling. The proposed approach combines word embeddings, trained on a corpus of hospital reviews, with the ELIta lexical emotion dictionary, which provides expert-based emotional annotations for Italian language terms. Word embeddings are used to capture contextual semantic similarity, while emotional information is transferred from the lexical resource to semantically related words through cosine similarity. At the document level, reviews are represented as aggregated embedding vectors, allowing the estimation of emotion-specific scores for each text. This procedure enables the detection of eight basic emotions and the identification of dominant emotional patterns in user-generated content. Unlike purely descriptive emotion mining approaches, the extracted emotion scores are subsequently modeled as latent constructs within a Partial Least Squares Structural Equation Modeling (PLS-SEM) framework. This integration supports a causal modeling perspective, enabling the analysis of relationships between emotional responses, perceived service dimensions, and overall satisfaction. Overall, the results show how emotional information extracted from free-text reviews can be formalized into latent constructs and incorporated into multivariate statistical models for the analysis of service evaluation data.

Keywords: Emotion Detection; Word2Vec; Lexical Resources; Textual Data Analysis; Structural Equation Modeling; PLS–SEM; Partial Least Squares

1. Introduction

Digital transformation has shifted patient feedback from traditional surveys to spontaneous on-line reviews. Platforms like Google Maps now provide a rich repository of patient experience data that, unlike structured questionnaires, captures the critical emotional nuances of healthcare delivery (Ranard et al., 2016). Despite its importance for National Health Service strategic planning, patient satisfaction is still largely measured through static tools that fail to account for the complex emotional drivers of user perception (Donabedian, 1988).

To address this gap, this paper leverages Natural Language Processing (NLP) to move beyond simple sentiment polarity toward granular emotion detection (Liu, 2020; Mohammad, 2016). We propose an integrated methodological approach that extracts emotional information from unstructured Italian hospital reviews and formalizes it into latent constructs for Partial Least Squares Structural Equation Modeling (PLS-SEM) (Angelelli et al., 2025; Ciavolino et al., 2022; Hair, 2014; Hair Jr et al., 2023; Sarstedt et al., 2022). This study seeks to extend existing methodologies for assessing healthcare quality, offering further instruments to support decision-making and service improvement in the public sector.

2. Method

This study adopts a quantitative, exploratory, and predictive research design aimed at analyzing the emotions expressed in online reviews of Italian public hospitals, in line with previous studies applying text mining and computational methods to user-generated content and service evaluation (Khan et al., 2014; Sanchez-Franco et al., 2019). The methodology is based on a hybrid pipeline that integrates NLP techniques for extracting emotional information from unstructured texts with a PLS-SEM framework to examine its relationship with user satisfaction. The unit of analysis is the individual online review, consisting of a free-text comment and a numerical rating. The dataset was collected from Google Maps and includes 19,695 reviews referring to 49 Italian public hospitals, of which 11,722 contain textual content and 7,973 report only the numerical rating. For each observation, territorial and temporal information were also considered and later used for comparative analyses. From a computational perspective, the textual data underwent a preliminary preprocessing phase including normalization, tokenization, and filtering of terms that could not be represented in the vector space (Jurafsky and Martin, 2019). The reviews were then represented through word embeddings obtained with Word2Vec, trained directly on the corpus (Goldberg and Levy, 2014; Mikolov et al., 2013). The semantic representation of each document was constructed by aggregating the vectors of the words contained in the text, following a compositional strategy based on average word embeddings (Arora et al., 2017). Emotion detection was performed through a hybrid approach combining the distributional model with ELItaGOLDEN, an emotion lexicon for the Italian language based on human annotation (Di Palma, 2024). Starting from the eight basic emotions of Plutchik's model (Plutchik, 1980), emotional scores were computed at the review level through semantic similarity measures between document vectors and the emotion-related terms included in the lexicon. The resulting scores were then aggregated into three first-order dimensions: positive emotions, negative emotions, and ambivalent/neutral emotions (Kim and Klinger, 2018; Russell, 1980). On this basis, a higher-order emotional construct was defined and used as a predictor of the rating within the PLS-SEM model (Cheah et al., 2019).

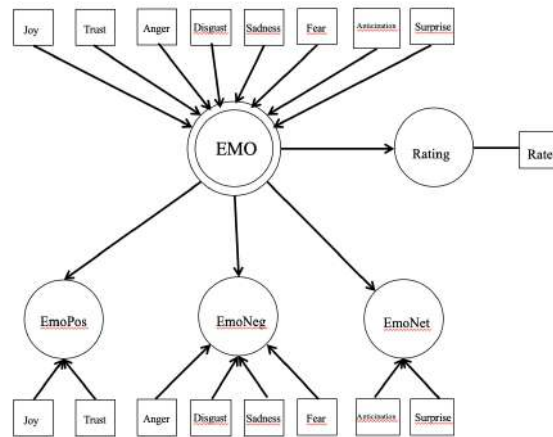


Figure 1: Path Model

This approach was selected because of its suitability for complex constructs, non-normally distributed data, and prediction-oriented objectives. Model assessment concerned both the measurement model and the structural model, through the evaluation of indicator weights, path coefficients, explained variance, and out-of-sample predictive ability, following established PLS-SEM guidelines for hierarchical component models and formative specifications, (Becker et al., 2012; Ingusci et al., 2024; Roldán and Sánchez-Franco, 2012; Sarstedt et al., 2022). Finally, to test for possible territorial and temporal heterogeneity, the analysis was extended through MICOM and Multi-Group Analysis (MGA) procedures (Henseler et al., 2016; Sarstedt et al., 2011), comparing the model across macro-geographical areas and between the pre- and post-pandemic periods.

3. Results

This section discusses the results of the measurement model, the structural model, and the multi-group analyses carried out through MICOM and MGA.

3.1. Validation of the emotion construct

The assessment of the measurement model confirmed the robustness of the hierarchical structure adopted. First, no multicollinearity issues emerged among the indicators, as all VIF values remained below the critical threshold. Second, the outer weights of the main emotional dimensions were overall significant, confirming the contribution of discrete emotions to the construction of the first-order constructs. In particular, anger and disgust emerged as the main contributors within the negative dimension, while trust and joy were the most relevant in the positive dimension. Within the ambivalent area, anticipation showed the strongest contribution, whereas surprise was not statistically significant. In general, these findings support the construction of a higher-order general emotional construct (EMO).

	VIF	IC 2.5%	IC 97.5%
Anticipation → EMO	1.681	1.608	1.765
Anticipation → EmoNet	1.105	1.088	1.122
Disgust → EMO	2.627	2.532	2.735
Disgust → EmoNeg	2.303	2.220	2.393
Trust → EMO	1.021	1.014	1.028
Trust → EmoPos	1.216	1.196	1.240
Joy → EMO	1.021	1.014	1.028
Joy → EmoPos	1.172	1.155	1.194
Fear → EMO	2.123	2.052	2.199
Fear → EmoNeg	1.903	1.841	1.967
Anger → EMO	1.946	1.887	2.006
Anger → EmoNeg	2.387	2.316	2.468
Surprise → EMO	2.191	2.112	2.274
Surprise → EmoNet	1.105	1.088	1.122
Sadness → EMO	2.127	2.030	2.235
Sadness → EmoNeg	2.026	1.934	2.125

Table 1: VIF Statistics

	Weights	p-val	IC 2.5%	IC 97.5%
Anticipation → EMO	0.469	0.000	0.455	0.482
Anticipation → EmoNet	1.008	0.000	0.980	1.028
Disgust → EMO	0.112	0.000	0.085	0.136
Disgust → EmoNeg	0.425	0.000	0.361	0.481
Trust → EMO	0.217	0.000	0.203	0.232
Trust → EmoPos	0.682	0.000	0.650	0.714
Joy → EMO	0.277	0.000	0.263	0.290
Joy → EmoPos	0.641	0.000	0.608	0.671
Fear → EMO	-0.411	0.000	-0.434	-0.386
Fear → EmoNeg	-0.833	0.000	-0.882	-0.778
Anger → EMO	0.338	0.000	0.319	0.357
Anger → EmoNeg	0.810	0.000	0.765	0.852
Surprise → EMO	0.021	0.218	-0.013	0.057
Surprise → EmoNet	-0.027	0.539	-0.111	0.061
Sadness → EMO	-0.099	0.000	-0.123	-0.075
Sadness → EmoNeg	-0.294	0.000	-0.354	-0.232

Table 2: Outer Weights

3.2. Structural model results

The structural model, figure 2, shows that the higher-order construct EMO exerts a positive and statistically significant effect on the overall service rating ($\beta = 0.489$, $p < 0.05$).

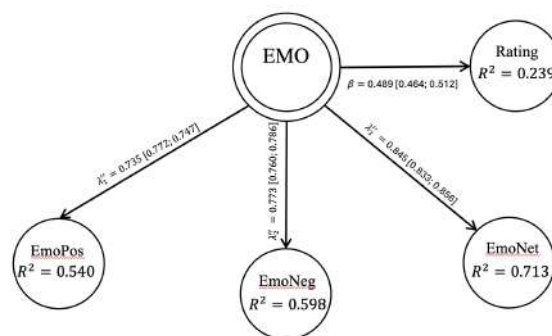


Figure 2: Results Structural Model

Table 3 reports the estimated path coefficients for the relationships among the latent constructs included in the model. All coefficients are positive and statistically significant. In particular, EMO shows strong links with the three first-order dimensions, supporting the hierarchical specification of the model. This finding confirms that the emotional component extracted from online

reviews represents a relevant determinant of user evaluations. The model shows an explanatory power of $R^2 = 0.239$ (p-value < 0.05, [0.215 ; 0.262]), indicating that about 24% of the variance in rating is explained by the emotional component. Given the unstructured nature of textual data, this value can be considered satisfactory. The effect size, in table 4 is also substantial ($f^2 = 0.314$), indicating a meaningful impact of the emotional construct on the dependent variable. From a predictive perspective, the PLSpredict analysis shows, in table 4s, a positive Q^2 value (0.057), suggesting out-of-sample predictive relevance. Overall, these results confirm that the emotions extracted from the texts are not only statistically significant, but also contribute to predicting users' overall evaluations.

	λ	Path Coefficient	p_values	IC 2.5%	IC 97.5%
EMO $\rightarrow$ EmoNeg	0.773		0.000	0.760	0.786
EMO $\rightarrow$ EmoNet	0.845		0.000	0.833	0.856
EMO $\rightarrow$ EmoPos	0.735		0.000	0.722	0.747
EMO $\rightarrow$ Rating		0.489	0.000	0.464	0.512

Table 3: Path Coefficient

	f^2	Q^2
EMO $\rightarrow$ Rating	0.314	0.057

Table 4: Effect size (f^2) and predictive relevance (Q^2)

3.3. Multi-group analysis

To test the stability of the model, measurement invariance (MICOM) and multi-group analysis (MGA) were conducted along two dimensions: territorial and temporal. From a territorial perspective, the results, in table 5 show that, although there are differences in the average levels of emotional expression and ratings across Northern, Central, and Southern Italy, the structure of the model remains substantially invariant. Although the Rating construct shows a p-value < 0.05 in the MICOM procedure, the multi-group comparison was still considered appropriate because Rating is a single-item construct. In all geographical comparisons, the path coefficients do not show statistically significant differences, as reported in table 6. This indicates that the mechanism through which emotions influence ratings is stable across the national territory.

A different pattern emerges along the temporal dimension, in table 5. The comparison between the pre- and post-pandemic periods reveals a more marked discontinuity. The MICOM procedure does not fully confirm compositional invariance, suggesting a change in the meaning attributed to some emotional dimensions after the pandemic.

Table 5: MICOM Summary: Geographical and Temporal Comparisons

Groups	Variables	Step 2 (Comp. Inv.)		Step 3a (Means)		Step 3b (Variance)	
		Corr.	p-val	Diff.	p-val	Diff. Var.	p-val
Southern vs Northern	EMO	1.000	0.720	0.006	0.805	0.003	0.926
	EmoNeg	0.999	0.821	0.001	0.981	-0.031	0.559
	EmoNet	1.000	0.911	-0.030	0.248	0.042	0.300
	EmoPos	1.000	0.725	0.044	0.085	-0.025	0.657
	Rate	1.000	0.000	-0.119	0.000	0.053	0.001
Southern vs Central	EMO	1.000	0.643	0.102	0.000	0.064	0.121
	EmoNeg	0.997	0.709	0.105	0.000	0.021	0.683
	EmoNet	1.000	0.624	0.026	0.340	0.108	0.032
	EmoPos	1.000	0.789	0.083	0.003	-0.071	0.255
	Rate	1.000	0.000	0.044	0.117	-0.041	0.006
Northern vs Central	EMO	1.000	0.666	0.096	0.000	0.057	0.089
	EmoNeg	0.999	0.805	0.107	0.000	0.044	0.309
	EmoNet	1.000	0.465	0.052	0.017	0.061	0.124
	EmoPos	1.000	0.458	0.040	0.051	-0.044	0.364
	Rate	1.000	0.000	0.163	0.000	-0.094	0.000
Pre vs Post COVID	EMO	0.998	0.019	0.054	0.026	0.013	0.701
	EmoNeg	0.950	0.098	0.055	0.023	-0.001	0.992
	EmoNet	0.999	0.042	-0.011	0.649	0.052	0.222
	EmoPos	1.000	0.005	0.056	0.016	0.001	0.988
	Rate	1.000	0.106	0.159	0.000	-0.153	0.000

Consistently, the MGA reveals a significant difference in the main structural relationship: the EMO $\rightarrow$ Rating coefficient increases from 0.354 in the pre-Covid period to 0.508 in the post-Covid period ($p < 0.05$), as illustrated by the comparison between the pre-Covid and post-Covid path models, figures 3 and 4.

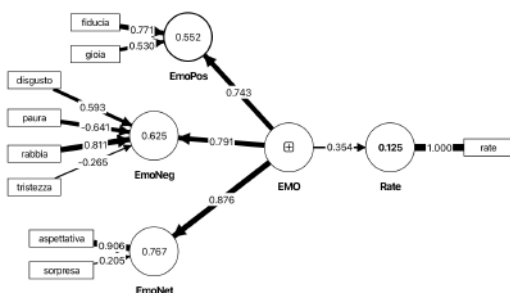


Figure 3: Result MGA Pre-Covid

3.4. Summary of findings

4. Discussion

that, despite cultural and communicative differences in expressive intensity, the mechanism through which emotional experience is translated into a synthetic evaluation remains substantially invariant. This result is consistent with the literature on unobserved heterogeneity, which highlights the importance of distinguishing between differences in mean levels and differences in structural mechanisms (Becker et al., 2012; Henseler et al., 2016). Even more relevant is the temporal result. The lack of full compositional invariance and the stronger EMO $\rightarrow$ Rating coefficient in the post-pandemic period suggest that Covid-19 acted as a shock capable of reshaping not only satisfaction levels, but also the meaning attributed to emotions expressed in reviews. In this sense, the pandemic may be interpreted as a semantic and evaluative turning point, in line with studies highlighting the transformative effects of the pandemic crisis on consumer behavior and user expectations (Donthu and Gustafsson, 2020; Sheth, 2020).

5. Limitations and Future Research

Among the main limitations of the study, the analysis is based exclusively on online reviews, which may under-represent less digitally engaged population groups. In addition, although effective, the emotion detection approach remains subject to the limits of domain dependence in healthcare language. Finally, the model identifies the emotions expressed by users, but does not precisely link them to the specific topics generating them. Future research could therefore integrate the current framework with topic modeling techniques in order to connect emotions with the thematic content of reviews and provide more actionable insights for healthcare management.

References

- Angelelli, M., Ciavolino, E., Ringle, C. M., Sarstedt, M., and Aria, M. (2025). Conceptual structure and thematic evolution in partial least squares structural equation modeling research. *Quality & Quantity*, 59(3):2753–2798.
- Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.
- Becker, J.-M., Klein, K., and Wetzels, M. (2012). Hierarchical latent variable models in pls-sem: guidelines for using reflective-formative type models. *Long range planning*, 45(5-6):359–394.
- Cheah, J.-H., Ting, H., Ramayah, T., Memon, M. A., Cham, T.-H., and Ciavolino, E. (2019). A comparison of five reflective–formative estimation approaches: reconsideration and recommendations for tourism research: J.-h. cheah et al. *Quality & quantity*, 53(3):1421–1458.
- Ciavolino, E., Aria, M., Cheah, J.-H., and Roldán, J. L. (2022). A tale of pls structural equation modelling: episode i—a bibliometrix citation analysis. *Social Indicators Research*, 164(3):1323–1348.
- Di Palma, E. (2024). Elita: A new italian language resource for emotion analysis. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 297–307.
- Donabedian, A. (1988). The quality of care: how can it be assessed? *Jama*, 260(12):1743–1748.
- Donthu, N. and Gustafsson, A. (2020). Effects of covid-19 on business and research.
- Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.
- Hair, J. F. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM)*. sage.
- Hair Jr, J., Hair Jr, J. F., Sarstedt, M., Ringle, C. M., and Gudergan, S. P. (2023). *Advanced issues in partial least squares structural equation modeling*. saGe publications.
- Henseler, J., Ringle, C. M., and Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *International marketing review*, 33(3):405–431.

- Ingusci, E., Angelelli, M., Sternativo, G. A., Catalano, A. A., De Carlo, E., Cortese, C. G., Demerouti, E., and Ciavolino, E. (2024). A higher-order life crafting scale validation using pls-cca: the italian version. *Behaviormetrika*, 51(1):359–387.
- Jurafsky, D. and Martin, J. H. (2019). *Speech and language processing* (3rd draft ed.).
- Khan, K., Baharudin, B., Khan, A., and Ullah, A. (2014). Mining opinion components from unstructured reviews: A review. *Journal of King Saud University-Computer and Information Sciences*, 26(3):258–275.
- Kim, E. and Klinger, R. (2018). A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.
- Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.
- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.
- Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., Asch, D. A., Ungar, L. H., and Merchant, R. M. (2016). Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs*, 35(4):697–705.
- Roldán, J. L. and Sánchez-Franco, M. J. (2012). Variance-based structural equation modeling: Guidelines for using partial least squares in information systems research. In *Research methodologies, innovations and philosophies in software systems engineering and information systems*, pages 193–221. IGI Global Scientific Publishing.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.
- Sanchez-Franco, M. J., Cepeda-Carrion, G., and Roldan, J. L. (2019). Understanding relationship quality in hospitality services: A study based on text analytics and partial least squares. *Internet Research*, 29(3):478–503.
- Sarstedt, M., Hair, J. F., Pick, M., Liengaard, B. D., Radomir, L., and Ringle, C. M. (2022). Progress in partial least squares structural equation modeling use in marketing research in the last decade. *Psychology & Marketing*, 39(5):1035–1064.
- Sarstedt, M., Henseler, J., and Ringle, C. M. (2011). Multigroup analysis in partial least squares (pls) path modeling: Alternative methods and empirical results.
- Sheth, J. (2020). Impact of covid-19 on consumer behavior: Will the old habits return or die? *Journal of business research*, 117:280–283.

Mapping Communicative Purposes in Academic Discourse: designing an analytic linguistic pattern

Alessandra Belfiore^{1,4}, Walter Giordano^{2,4}, Massimo Aria^{2,4}, Corrado Cuccurullo^{2,3,4}

¹ Università Telematica Pegaso – alessandra.belfiore@unipegaso.it

² Università degli Studi di Napoli Federico II - walter.giordano@unina.it

³ Università della Campania Luigi Vanvitelli – corrado.cuccurullo@unicampania.it

⁴K-Synth, academic spin-off

Abstract

This study presents an empirical analysis aimed at systematizing the mapping of communicative purposes in academic discourse through the construction of an analytical repertoire of linguistic patterns, with a particular focus on rhetorical moves. Starting from a random corpus of academic texts, we identified recurring linguistic forms that link each functional linguistic and syntactic strategy to specific communicative purposes, providing operational definitions, identification criteria, and typical linguistic frames. This pilot study makes the rhetorical functions underlying authors' linguistic choices explicit and facilitates a more refined and replicable analysis of the structure of scientific texts. To demonstrate the functionality and effectiveness of our work, we present examples of its application, highlighting how it can guide the reading and interpretation of various sections of academic articles, from the introduction to the discussion. Looking ahead, we aim at expanding this work by developing a semi-automated system based on scoring or ranking algorithms, capable of assisting in the identification of potential rhetorical moves present in texts. This approach offers innovative tools to support both the production and evaluation of academic discourse, paving the way to computational analysis of scholarly language and for training researchers and students in effective argumentative strategies. Overall, the project bridges theoretical analysis, descriptive tools, and practical applications, contributing to a more systematic understanding of rhetorical structures in academic writing.

Keywords: Rhetorical moves, Academic Discourse, Discourse markers, Syntactic strategy, Functional Discourse Analysis

1. Introduction

In today's global research landscape, academic writing is no longer merely a means to spread knowledge; it has become a highly competitive and standardized arena. The proliferation of scientific publications and the pressure of the "publish or perish" system have led to a crystallization of textual genres, turning academic discourse into an ecosystem governed by strict rhetorical conventions. In this context, the effectiveness of an article does not depend solely on the intrinsic value of the data presented, but on the author's ability to negotiate their position within the relevant scientific community. The command of rhetorical moves has thus

become the fundamental prerequisite for ensuring the visibility and authority of research, transforming the text into a sophisticated tool of persuasion and strategic positioning.

However, despite the centrality of these dynamics, a critical gap persists between theoretical awareness of the content to be conveyed and practical proficiency in manipulating the linguistic structures necessary to convey it. Although researchers and students clearly identify what to write (the subject of the research), they often encounter insurmountable obstacles in mastering how to write (functional syntax). The communicative functions underlying lexical and syntactic choices frequently remain implicit or are left to the individual's stylistic intuition, making the analysis and production of scientific text an opaque and difficult-to-replicate process. To date, there is no systematic analytical tool that translates the abstractions of genre theory into tangible and operational linguistic patterns capable of guiding the author in the construction of a rigorous and transparent argumentative discourse.

To address this methodological gap, and based on our firm conviction that interdisciplinary collaboration between studies on statistical methods and research on linguistic and discourse interpretation is extremely fruitful, we have started this thread of study, which proposes a systematic approach based on mapping communicative purposes through the construction of an analytical repertoire of linguistic patterns. The main objective is to unambiguously link specific syntactic frames to precise rhetorical moves. Through the identification of textual regularities extracted from an extensive corpus, the research offers operational definitions and identification criteria that make the authors' argumentative strategies explicit. In fact, our research question can be formulated as follows: RQ1: Is it possible, through empirical observation and manual annotation of the introductions to research articles, to extract recurring linguistic patterns that can be systematized into a functional repertoire aligned with the structure of the rhetorical moves in the CARS model? In the light of this, our work is not limited to a theoretical description but provides a practical and scalable tool, laying the groundwork for future integrations with computational analysis systems and scoring algorithms aimed at supporting both the production and the assisted evaluation of academic discourse.

This paper is structured as follows. Section 2 outlines the theoretical framework; Section 3 describes the methodology and analytical process; Section 4 presents the applied analyses; to conclude, Section 5 discusses the implications for research and practice.

2. Literature Review

At the very heart of the entire process of writing a high-quality research article lies an understanding of the genre and subgenre to which the article belongs. In scientific literature, genre refers to the type of article (in this case, an academic or scientific article), while subgenre defines the specific type of article. Correctly identifying the genre and subgenre helps establish the tone, style, and expectations for the content, thereby making it easier to adhere to disciplinary conventions. The distinction between genre and subgenre is essential for avoiding formatting errors and ensuring that the reader can understand and appreciate the work. Each genre and subgenre affects the structure of the article by determining the organization and inclusion of sections; it affects writing style by influencing register, tone (formal or informal), voice (active or passive), and level of detail; and it affects content by defining what information to include in each section and the level of depth required.

The concept of "genre" in academic discourse has undergone a radical transformation thanks to the pioneering work of John Swales (1990). Moving beyond a purely stylistic view, Swales defines genre as a set of communicative purposes recognized by the academic community. In

the CARS model—an acronym for Create A Research Space—he establishes a framework for structuring the introduction of academic and research articles. This model is structured in three phases: 1. Establishing the territory, 2. Defining a niche, 3. Occupying the niche. In the first phase, the author sets the stage by introducing the general context and the significance of the research topic. This may involve describing the broader field of study, providing essential background information, and highlighting the importance of the topic. In the second phase, the author identifies a gap in existing research or a specific problem that has not been adequately addressed. This step is crucial as it justifies the need for the current study and prepares the ground for the research question or hypothesis. In the final phase, the author explicitly states the objective, purpose, or thesis of the current research. This may include describing the research questions, hypotheses, or objectives. The author may also mention the methodology and potential contributions of the research, thereby nestling the new research within the identified niche.

The literature following Swales (Bhatia, 2004; Hyland, 2005) has further explored the concept of rhetorical moves. A rhetorical move is a functional unit of text that contributes to the realization of the genre’s overall communicative purpose. The literature has highlighted how rhetorical competence is intrinsically linked to the ability to use “metadiscourse” and positioning strategies (hedging and boosting). However, the main challenge remains the objective identification of rhetorical moves. The literature often focuses on the macro-textual function, leaving in the shadows the linguistic micro-signals that allow us to distinguish, for example, a “description of method” from a “methodological justification.”

Corpus linguistics offers an indispensable empirical methodology. The use of large databases of academic texts makes it possible to identify what Sinclair (1991) refers to as “syntactic regularities” or “lexical bundles.” The underlying hypothesis is that every communicative purpose corresponds to a constellation of preferred linguistic signals. Integrating Genre Analysis with corpus analysis means moving from a qualitative description of the text to a quantitative and replicable mapping. The analytical approach to recurring linguistic structures is rooted primarily in the study of discourse markers—that is, the parts of speech that function as signaling devices for specific sets of meaning (Hyland 1998; Bazzanella 2002; Biber et al. 1999; di Ferrante 2021), in their capacity to serve as semantic links (Fraser 2011) and their functional relevance depending on their position and sequential frequency (Haselow 2019). Discourse analysis based on discourse markers has recently yielded results primarily related to the structural study of rhetorical moves, as well as to discursive characteristics (Egbert and Plonsky 2015; Halleck and Connor 2006). Di Ferrante brilliantly (2023) identifies the functional implications of syntactic structures and applies them to an analysis of academic article titles. The basic idea, then, is that meaning does not lie in the individual word, but in the linguistic pattern (the recurring syntactic formula) that acts as a “fingerprint” of the rhetorical function.

3. Metodologia

3.1 Corpus Construction

The corpus was constructed using a stratified random sampling approach. Articles were retrieved using the openalexR package (Aria et al., 2024), an R interface for the OpenAlex API. To ensure that the corpus of linguistic models was drawn from texts of proven rhetorical effectiveness, the following inclusion criteria were applied: (1) only journal articles (excluding

review articles, editorials, or conference proceedings); (2) a citation count exceeding 1,000, ensuring that the texts had successfully passed peer review and exerted significant influence in their respective fields; (3) availability in Open Access format, allowing for the retrieval of the full text in PDF format. Following this protocol, 100 articles were systematically downloaded using the full-text URLs provided by OpenAlex metadata.

3.2 Corpus Partitioning

To develop and subsequently validate the analytical model, the corpus was divided into two subsets. Fifty articles were assigned to the training set, which was used for the manual extraction of models and categorization. Of these 50 articles, 43 were usable after excluding texts with damaged PDFs, incomplete introductions, or hybrid section structures that could not be reliably segmented. The remaining 50 articles constitute a test set reserved for the experimental verification of the scalability and accuracy of the syntactic features identified in future work.

3.3 Annotation Procedure

An expert linguist annotated the training set using a “function-first” analytical protocol (Lu et al., 2021) in three phases.

In the first phase, each introduction was segmented according to the CARS model (Swales, 2004), isolating Move 1, Move 2, and Move 3 as primary units of analysis. The annotator followed the diagnostic criteria established in the literature on move analysis: Move 1 was identified through statements of centrality, generalization of the topic, and review of previous research; Move 2 through the identification of gaps, adversative constructions, and statements of limitation; Move 3 through statements of purpose, announcements of the current research, and structural patterns. Each sentence was assigned a primary rhetorical move label; sentences that fulfilled two simultaneous functions received dual labels with a designated primary function.

In the second phase, syntactic patterns were extracted through a careful reading of the annotated segments. This phase focused on identifying fixed or semi-fixed syntactic formulas (e.g., “In this article, we present...”; “Despite the growing importance of...”) and on recurring clause types: assertive propositions (direct statements about the world or the research), concessive propositions (constructions introducing a presupposed premise with an adversative implication), and incidental propositions (parenthetical or non-restrictive clauses that add supplementary information).

In the third phase, functional mapping was performed. Each extracted pattern was cataloged and assigned a syntactic function label (what it does) and a rhetorical category (which move it belongs to).

4. Findings

4.1 Distribution of Rhetorical Moves

In each article included in the training set, the introductory section was isolated and annotated to verify the presence of the three CARS rhetorical figures (M1, M2, or M3). The results are summarized in Table 1.

Table 1. Distribution of rhetorical moves across 43 annotated introductions.

Condition	No M1	No M2	No M3	M1+M2 fused	No moves	Full or partial presence of moves
N (out of 43)	1	1	1	7	6	37
Percentage	2.3%	2.3%	2.3%	16.3%	14.0%	86.0%

The results reveal that 37 out of 43 articles (86.0%) feature rhetorical moves, either all of them or a partial presence. Six articles (14.0%) do not feature any identifiable moves, suggesting an atypical structural organization or insufficient rhetorical explicitness. Move 1, Move 2, and Move 3 are each absent in one case. Specifically, in 7 cases (16.3%), Move 1 and Move 2 appear merged into a single undifferentiated treatment, without an explicit segment signaling a discontinuity to separate them. This pattern is consistent with the observation by Ramadhanty and Sutrisno (2022) that Move 2 functions as an optional move in English-language academic writing, with authors sometimes preferring to justify their research through the research problem itself rather than articulating an explicit gap.

4.2 Syntactic Clause Types by Rhetorical Move

An in-depth analysis of the 37 articles that feature all three strategies provided data on the distribution of recurring linguistic forms across the various rhetorical strategies, as shown in Table 2.

The most common syntactic forms consist of a large number of assertive clauses, followed by a similarly high number of concessive clauses, and a smaller but still significant number of parenthetical or incidental clauses.

Table 2. Distribution of syntactic sentences types across rhetorical moves

Sentences type	M1	M2	M3
Assertive	0	0	47
Concessive	0	28	4
Incidental	3	6	6

As can be seen, the massive presence of assertive clauses in rhetorical move M3 is significant, since, by its very nature, it serves as a strategic tool to “fill the research gap.” Move 3 serves to delineate the scope of the research by announcing the objectives, methods, and expected contributions of the present study—functions that inherently require declarative constructions with a high degree of certainty. Assertiveness is therefore strongly justified in “making room”

for certainty at the expense of uncertainty. In fact, for example, the opposite strategies—those of hedging—might rightly be found in the results discussion section, where it makes sense to take a cautious and discretionary approach in presenting the results.

Some examples of strongly assertive statements found in Section M3:

Our method thus enables exploration of a vast state space of possible models,

The method described here differs from previous methods in several key respects,

We believe that internationalization is the product of a series of incremental decisions,

This article describes four foundational principles — Findability, Accessibility, Interoperability, and Reusability — that serve to guide data producers

Concessive clauses appear almost exclusively in M2 ($n = 28$), consistent with their rhetorical function of highlighting a gap. Concessive constructions introduce a positive, presupposed state of the literature before moving on to its limitations or omissions. In fact, concessive clauses are those that express a conditional premise (despite the fact that..., although there are..., etc.); they fit well with the strategic function that M2 aims to convey. In fact, by stating that “current studies do not consider,” “few scholars have focused on,” etc., space is created for premises that reveal the scientific gap that will then be “filled” in the rhetorical M3.

Some examples of concessive clauses found in M2:

Despite the increasing importance of this approach,

While some evidence suggests that better-qualified teachers may make a difference,

Despite these indications of correlations with cognitive ability measures.

Yet time trends in consumption are largely unknown,...

To round out the set of recurring propositions, there are also a fair number of parenthetical or incidental propositions. Incidental clauses are distributed more evenly across the moves (M1: 3; M2: 6; M3: 6), reflecting their semantically neutral function of providing supplementary information rather than promoting a specific rhetorical purpose. They serve to supplement and expand understanding of semantic content. Incidental propositions enrich communication by providing further information about the message’s content. Summarizing, rhetorical move M1 has a neutral strategic content, as it does nothing more than describe the state of the art in that branch of science. M1 and M2, on the other hand, due to their nature as described here, have a markedly higher density of strategic communication.

4.3 Discourse Connectors by Rhetorical Move

A second level of analysis examined the distribution of discourse connectors across the different phases of the discourse. Table 3 presents the results.

Table 3. Distribution of discourse connectors across rhetorical moves.

Connector type (examples)	M1	M2	M3
Causal-consecutive (thus, consequently)	—	9	—
Additive (furthermore, moreover, in addition)	1	9	3

Connector type (examples)	M1	M2	M3
Adversative (however, yet)	—	21	5
Conclusive (finally)	—	4	1
Intensifying (indeed)	—	4	1
Exemplificative (for example)	—	6	—
Other (instead, unfortunately, etc.)	—	23	4

The data confirm the strategic poverty of Move 1, in terms of rhetorical power, from the perspective of discourse markers. Only a single additive connector is found, consistent with the descriptive function and low argumentative density of this move. Move 2 accounts for the vast majority of connectors of all types, with adversative connectors (*however, yet*) representing the most frequent category by far ($n = 21$). This result confirms the rhetorical function of Move 2 described by Swales (2004): adversative connectors are the primary lexical mechanism through which the authors signal a shift from established territory to its problematic dimensions. Additive connectors (*furthermore, moreover, in addition*) in Move 2 perform a complementary function, accumulating evidence of the inadequacy of the existing literature.

Noteworthy is the presence of exemplifying connectors (for example, $n = 6$) exclusively in Move 2. This suggests that authors use the concrete instantiation of the gap as a rhetorical strategy, grounding the abstract claim of insufficiency in specific, observable cases.

5. Discussion and Conclusions

The results obtained in our preliminary work are intended to serve as the foundation for a methodological evolution of this study. The data collected, in fact, appear to confirm the initial empirical observation of the presence of semantic tools that support rhetorical emotions in their functions. In this way, our analysis provides a perspective that focuses on functional syntactic structures (types of propositions), which can undoubtedly suggest methodological innovations, primarily related to automated methods of analysis.

Our results are consistent with the existing literature on corpus-based move analysis and expand upon it in three ways. First, they demonstrate that it is the syntactic types of propositions that constitute systematic functional correlates of CARS moves. The concentrated distribution of assertive propositions in M3 and concessive propositions in M2 suggests that clause-level syntax encodes rhetorical intent in structured and learnable ways.

Second, the near-total absence of discourse connectors in M1, and their concentration in M2, provides empirical support for the theoretical claim that M1 exhibits low strategic communicative density, while M2 is the most rhetorically marked and argumentatively complex move in the introduction.

Third, the merging of M1 and M2 in 16.3% of the articles raises methodological issues for future annotation studies. Automatic classifiers trained exclusively on the three-step CARS architecture may systematically misclassify introductions that incorporate gap signaling within

the domain definition, suggesting the need for a hybrid annotation scheme capable of accommodating this model.

Our investigation has important implications for research and practice.

Related to research, our work enables the training of students and early-career researchers by allowing them to work directly with the data to extract the hidden structure within academic documents. Academic writing is often perceived as a “black box” where conventions are learned through osmosis or trial and error. The tool makes the correlation between syntactic choice and rhetorical intention explicit, reducing the cognitive load during the text production phase. For non-native speakers, having access to typified linguistic frames associated with clear communicative purposes (e.g., gap-filling or announcing purposes) represents a fundamental step toward equitable participation in the global scientific conversation.

In practical terms, this study serves as the foundation for developing a tool that automates this type of analysis. The tool will assist authors and editors of scientific journals during the pre-screening phase by quickly identifying articles that contain gaps in their arguments or lack a clear structure of fundamental rhetorical moves.

Future research will focus on automating the recognition of rhetorical devices. In fact, the next logical step is to implement algorithms capable of assigning a “rhetorical coherence” score to a text. By analyzing the density and sequence of the detected patterns, the system will be able to verify whether an introduction, for example, meets the requirements of the CARS model. We plan to develop a prototype capable of scanning files and suggesting to the user the possible rhetorical move for each paragraph. This system would serve as augmented peer review, providing immediate feedback on the logical structure of the manuscript. Integration with Generative Artificial Intelligence models represents the game-changer of this project. Instead of relying on generic text generation, our repertoire of patterns can be used to train AIs to produce structurally rigorous scientific texts, acting as a “functional constraint” that prevents stylistic drift.

In conclusion, the bridge built between linguistic analysis and computational applications opens up new possibilities for a deeper understanding of scientific language, transforming writing from a bureaucratic hurdle into a transparent and effective argumentative strategy.

References

- Aria, M., Le, T., Cuccurullo, C., Belfiore, A., & Choe, J. (2024). openalexR: An R-Tool for Collecting Bibliometric Data from OpenAlex. *R J.*, 15(4), 167-180.
- Bazzanella, Carla (2001) Segnali discorsivi e contesto. In Heinrich, Wilma and Heiss, Christine (eds.) *Modalità e Fraser*, Bruce (2011) The sequencing of contrastive discourse markers in English. *Baltic Journal of English Language, Literature, and Culture* 1, 29–35. substandard. Bologna: Clueb, 41–64. urgh: Pearson Education Limited
- Bhatia, V. (2014). *Worlds of written discourse*.
- Biber, Douglas, Johansson, Stig Leech, Geoffrey, Conrad, Susan, Finegan, Edward and Quirk, Randolph (1999) *Longman Grammar of Spoken and Written English*. Edinburg
- Di Ferrante, Laura (2023). (Re)presenting science in research articles and press releases: a contrastive analysis of titles and headlines. *Iperstoria*, 22, 187-214. <https://doi.org/10.13136/2281-4582/2023.i22.1350>

- Di Ferrante, Laura (2021). Transitioning between small talk and work talk through discourse markers: Evidence from a workplace spoken corpus. *Brno Studies in English*, 47(2), 7-30. <http://doi.org/10.5817/BSE2021-2-2>
- Egbert, J., & Plonsky, L. (2015). Success in the abstract: Exploring linguistic and stylistic predictors of conference abstract ratings. *Corpora*, 10(3), 291–313.
- Halleck, G. B., & Connor, U. M. (2006). Rhetorical moves in TESOL conference proposals. *Journal of English for Academic Purposes*, 5, 70–86.
- Haselow, Alexander (2019) Discourse marker sequences: Insights into the serial order of communicative tasks in real-time turn production. *Journal of Pragmatics* 146, 1–18.
- Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse studies*, 7(2), 173-192.
- Sinclair, J. M. (1992, January). The automatic analysis of corpora. In *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* (Vol. 82, pp. 379-397).
- Swales, J. M. (1990). *Genre analysis*. Cambridge University Press.

Enhancing leader-based community detection in scientific collaboration using publications metadata

Sara Geremia¹, Domenico De Stefano²

¹University of Trieste – sara.geremia@phd.units.it

²University of Trieste – ddestefano@units.it

Abstract (in English)

Leader-based approaches to community detection provide a powerful framework for uncovering group structures in collaboration networks characterized by degree heterogeneity, but are typically grounded in purely structural information, which limits their ability to capture thematic cohesion in scientific communities. This work introduces an attribute-driven extension of the Density-based Community Detection (DeCoDe) framework, designed for collaboration networks where communities emerge jointly from network topology and publication metadata. Focusing on scientific collaboration among Italian scholars as a case study, we construct a network from the Italian Ministry of University and Research and Scopus data, where nodes represent authors and edges encode co-authorship ties, and we characterize each node with attributes derived from publication-level textual descriptors. We represent the scientific content of papers using their topics, which are aggregated at the author level and transformed into dense semantic embeddings via the all-MiniLM-L6-v2 SentenceTransformer model, yielding a high-dimensional attribute space tailored to quantitative text analysis. In this space, we estimate local attribute density using a density estimator to identify leaders as nodes located in high-density regions of the embedding space, rather than via structural centrality alone as in traditional leader-based approaches. These attribute-based densities are then integrated into an extended DeCoDe procedure, which defines communities as connected components in upper-level sets of the density function, thereby enforcing both structural connectivity and thematic homogeneity. Results on empirical data indicate that attribute-based density estimation supports the identification of communities that are both densely interconnected and thematically interpretable, highlighting leaders whose profiles are representative of the dominant themes of their groups.

Keywords: Co-authorship networks, Community detection, Attributed networks, Homophily, Leader-based attraction.

1. Introduction

Scientific collaboration networks are typically analyzed by means of co-authorship data (Cugmas et al., 2019; De Stefano et al., 2023; Fabbrucci Barbagli et al., 2025), where nodes represent researchers and links indicate joint publications. This representation allows the study of structural properties such as topology (De Stefano et al., 2013; Newman, 2001), centrality (Lu and Feng, 2009), or community formation (Geremia et al., 2025a; Ghosh and Kshitij, 2017; Zhang et al., 2023), revealing how knowledge is produced and diffused within scientific communities (Lužar et al., 2014).

However, relying solely on relational data may overlook the semantic dimension of collaboration and the context in which scholars choose how to collaborate with others. For this reason, recent research increasingly integrates textual data—such as titles, abstracts, and keywords—into network analysis (Li et al., 2025). Approaches like word co-occurrence analysis and topic modeling enable the identification of research themes and their evolution, capturing the content of scientific production and collaboration patterns (Lozano et al., 2019). Recent applications combine network structure with textual information (e.g., author–topic models or heterogeneous

graph methods), allowing the analysis of topical communities (Schäfermeier et al., 2023). This integration is particularly relevant for detecting emerging research areas, mapping interdisciplinary collaborations, and analyzing the evolution of scientific fields.

In this study we follow this route and our aim is to identify tightly connected groups of Italian academic scholars in the fields of statistics, business administration, and sociology by analyzing both patterns of scientific collaboration and the topics addressed in their articles published in international journals indexed in the Scopus database. Identifying such groups (or communities) is fundamental in the study of scientific collaboration, especially when considering not only co-authorship ties—as is common in the literature—but also attributes derived from bibliographic and individual data. To this end, we adopt AttDeCoDe, a novel community detection approach for leader-based community detection in attributed networks proposed in Geremia et al. (2025b), which is suitable for networks where both leader-based mechanisms and homophily may govern group formation. These relational mechanisms are particularly relevant in scientific collaboration networks, where groups are often shaped by prominent authors (for instance, senior researchers) and by homophily in terms of similar scientific fields or research interests or other personal characteristics (Horta et al., 2022).

The paper is organized as follows: Section 2. illustrates the data retrieved on the target population under study and the individual and bibliographic data that are used to enhance the collaboration network structure and the related attributed network data; Section 3. describes the community detection method used to cluster the derived attributed network; Section 4. shows the results and concludes.

2. Scientific collaboration data on Italian scholars

We collect bibliographic data of Italian Academic scholars belonging to three disciplines, namely Statistics, Management, and Sociology. We consider the so-called Settori Scientifico Disciplinari (hereafter, SSD or sectors), which represent the finest official classification of scientific disciplines according to the Italian Ministry of University and Research (MUR, Ministero dell’Università e della Ricerca). Each SSD corresponds to a specific research field and is used to organize academic positions, publications, and evaluation procedures. In this study, we focus on a subset of SSDs spanning these three disciplines: the statistical area includes SECS-S/01 (Statistics), SECS-S/02 (Statistics for experimental and technological research), SECS-S/03 (Economic Statistics), SECS-S/04 (Demography), and SECS-S/05 (Social Statistics). Management is represented by only scholars belonging to SECS-P/08 sector (Management), while the sociological domain includes SPS/07 (General Sociology) and SPS/09 (Economic Sociology). These sectors include both methodological disciplines and applied social sciences, allowing us to investigate collaboration patterns and thematic similarities across related research areas.

The data on scholars were collected from the MUR database (Ministero dell’Università e della Ricerca, 2025). For each year from 2014 to 2024, we obtained information on individual researchers, including their scientific sector, Italian geographical area of the hiring university (North-Western, North-East, Center, South, and Island), and their academic role (Researcher, Associate Professor or Full Professors). Bibliographic and co-authorship data were collected from Scopus using each scholar’s Scopus IDs, with several checks conducted to verify the correspondence between names, surnames, and their respective universities of affiliation. The final dataset was created using Python and the Pybliometrics library (Rose and Kitchin, 2019), including publication unique identifiers, keywords, references, and co-authors’ Scopus IDs. The dataset allows us to construct a two-mode network (authors-by-papers) including 1,809 pa-

pers and 2,691 authors, as well as the corresponding binary one-mode co-authorship network (authors-by-authors). The resulting one-mode network is quite sparse (the overall density is 0.09) but exhibits a moderately high clustering tendency (transitivity equal to 0.26).

Table 1 reports the distribution of the retrieved scholars by sector, role, and location, along with key network statistics. Statisticians (30.7% of the network) show the highest connectivity and bridging role (average degree 5.06, average betweenness 5883.28, internal density 0.20), indicating strong within-group interconnections, whereas scholars in Management and Sociology are more weakly connected (edge densities of 0.05 and 0.04, respectively). The degree distribution is right-skewed, reflecting substantial heterogeneity in collaboration patterns and the presence of a few highly connected hub authors, or leaders. This motivates the leader-based community detection approach adopted in the following analysis to identify clusters of scholars. Beyond network topology, the availability of individual-level and bibliometric attributes for each node further motivates an attributive enrichment of the network structure.

3. Methodology

The clustering problem in our context is framed in terms of detecting high-density regions on the network, where nodes are grouped according to a node-wise measure of density (Menardi and De Stefano, 2022). Two nodes belong to the same cluster if their densities exceed a given threshold and they are connected through a path in G . This idea extends modal clustering to a relational setting: node-wise densities $\hat{\delta}(v_1), \dots, \hat{\delta}(v_i), \dots, \hat{\delta}(v_n)$ are estimated, and communities form by aggregating nodes around modal actors, i.e., local maxima of the density function.

In unweighted networks, for any threshold λ , the upper-level set $V_\lambda = \{v_i \in V : \hat{\delta}(v_i) \geq \lambda\}$ induces a subgraph $G_\lambda = (V_\lambda, E_\lambda)$, where $E_\lambda = \{e_{ij} \in E : v_i, v_j \in V_\lambda\}$. The connected components of G_λ represent density-based groups. As λ decreases, these components merge, producing a cluster tree that captures the hierarchical organization of communities. The number of network clusters K is not fixed a priori, but is determined by the structure of this tree: clusters are identified as the connected components of G_λ at the lowest λ levels where they appear as leaves

		N	%	Average Degree	Average Betweenness	Density (%)
	Total	2691		2.34	2917.98	0.09
Sector	Statistics	827	30.73	5.06	5883.28	0.20
	Business	821	30.51	1.26	1489.36	0.05
	Sociology	1043	38.76	1.02	1691.31	0.04
Role	Full Professor	748	27.80	3.20	4662.07	0.12
	Associate	1134	42.14	2.25	2696.91	0.08
	Researcher	809	30.06	1.67	1615.27	0.06
Location	North-West	687	25.53	2.23	2546.70	0.08
	North-East	591	21.96	2.29	3251.66	0.09
	Center	689	25.60	2.39	3369.73	0.09
	South	517	19.21	2.49	2618.23	0.09
	Islands	207	7.69	2.28	2442.49	0.09

Table 1: Co-authorship network characteristics

of the cluster tree. Low-density nodes may either remain unassigned or be attached to nearby cluster cores. In addition to the detected number of clusters $\hat{K}$ and the cluster membership $C(v_i) \in \{1, \dots, K\}$, the algorithm outputs an indicator $r(v_i)$ specifying whether node v_i acts as a cluster core ($r(v_i) = 1$) or as a non-core member ($r(v_i) = 0$).

Building on this framework, AttDeCoDe (Geremia et al., 2025b) generalizes DeCoDe to node-attributed networks $\{G = (V, E), \mathbf{X}\}$, where node densities are computed from the attribute matrix $\mathbf{X}$ rather than from network structural properties. This approach identifies clusters that emerge around actors located in high-density regions of the attribute space. Under homophily, such representative nodes act as focal points, around which others with comparable attributes tend to cluster. Structural connectivity is preserved by requiring paths in the original network to link nodes, ensuring that communities reflect both attribute similarity and topological integrity. AttDeCoDe therefore shifts the focus from purely structural density to attribute-informed density, allowing the clustering process to reflect meaningful variation in the attributes that characterize each node.

3.1. Application to scientific collaboration data

The AttDeCoDe perspective is consistent with the logic of scientific collaborations: scholars engaged in similar research areas tend to cluster around shared priorities, while differences in expertise or sector generate boundaries that limit the formation of ties. To identify clusters driven by homophilous influence, we construct node-level attributes by assigning each researcher the complete set of topic keywords attached to the scientific papers in which it collaborates. This yields a standardized thematic profile for each author, allowing us to assess whether proximity in scientific domains is reflected in the observed clustering patterns. By grounding community detection in paper-level attributes, our framework identifies leaders that are both thematically central—exhibiting high attribute density—and structurally well positioned to attract collaborations, thereby accumulating a broad and diverse keyword profile.

We represent each node through a dense semantic embedding of its associated keywords and estimate local concentration in the resulting numerical attribute space. Specifically, we generate sentence embeddings from the concatenated keyword lists using the all-MiniLM-L6-v2 SentenceTransformer model (Hugging Face, 2024), which yields 384-dimensional vectors in a dense semantic space. This embedding approach allows us to capture higher-order semantic relationships among keywords beyond simple term co-occurrence, and to represent nodes in a continuous attribute space suitable for density estimation.

For estimating the attribute-based density, we considered a non-parametric kNN density estimator with $k = 5$. This relatively small value of k allows for capturing local variations in attribute concentration while maintaining robustness against noise.

4. Results and discussion

By applying AttDeCoDe algorithm, we detect 269 communities ranging from a minimum size of 1 up to 54 (the largest cluster). The average size of the detected cluster is 5. The distribution of communities is highly skewed, with most clusters being small and tightly knit, while a handful of larger communities emerge around topic representatives. Figure 1 illustrates the density structure of the network: the t-SNE projection (van der Maaten and Hinton, 2008) of node attributes (panel a) reveals spatial separation among groups of scholars, with higher-density regions corresponding to nodes that are representative in terms of research topics, while panel (b) confirms that node degree and local density are skewed, consistent with the heterogeneous

collaboration patterns described in Section 2., while kNN attribute density seems to capture local thematic concentration.

We focus the discussion on the four largest clusters — Cluster 70 (54 members), Cluster 66 (35 members), Cluster 150 (28 members), and Cluster 94 (27 members) — whose direct connections are visualized in Figure 2. Table 2 reports the distribution of scholars by area, sector, and role within each cluster, with shaded cells indicating the categories characterizing cluster leaders, while Table 3 presents the top-10 keywords by frequency for each cluster, with shading indicating overlap with the leaders' keyword profiles.

Cluster 70. The largest cluster (54 members, 6 leaders) is dominated by statisticians from SECS-S/01 and SECS-S/03 (each accounting for 31.5% of members), with a non-negligible presence of Management scholars (SECS-P/08, 25.9%). Geographically, members are concentrated in the North-East (31.5%) and North-West (22.2%), consistent with the leaders' geographical profile. Senior scholars are well represented, with full professors (37.0%) and associate professors (35.2%) together making up nearly three quarters of the cluster. The keyword profile reflects a dual focus on firm-level economics and applied statistical methods: *Firm Size Distribution*, *Gibrat's Law*, *Spatial Econometrics*, and *Panel Data Models* are among the most frequent terms and are shared with cluster leaders, suggesting that these topics drive collaboration within the group. The presence of *COVID-19* and *Maximum Likelihood Method* further points to a cluster engaged in applied econometric and statistical modeling, including pandemic-related research.

Cluster 66. The second largest cluster (35 members, 3 leaders) stands out for its strong interdisciplinary character, combining demographers (SECS-S/04, 40.0%), statisticians (SECS-S/01 and SECS-S/05, 20.0% each), and sociologists (SPS/07 and SPS/09). Leaders span the Center and North-East, and the cluster is geographically the most diverse, with a notable share of Central Italian institutions (34.3%). The keyword profile is distinctly demographic and sociological, centered on family dynamics and social support: *Childlessness*, *Birth Rate*, *Postponement*, *Divorce*, *Cohabitation*, *Grandparents*, and *Caregiving* dominate the frequency list.

Cluster 150. This cluster (28 members, 3 leaders) is the most homogeneous in terms of disciplinary composition, with SECS-S/01 accounting for 71.4% of members and SECS-S/02 contributing a further 17.9%. It is geographically spread across the North-East (28.6%), North-West (28.6%), and South (25.0%), with leaders present in the last two macro-areas. The keyword profile is strongly oriented towards spatial and geostatistical modeling: *Spatio-Temporal*, *Variogram*, *Covariance Function*, *Disease Mapping*, and *INLA* are all prominent, and most are shared with the leaders' keyword profile, indicating a tightly focused research agenda.

Cluster 94. The fourth largest cluster (27 members, 2 leaders) is also heavily concentrated in SECS-S/01 (70.4%), with a distinctive geographical profile: nearly 90% of members are located in the North-West (44.4%) or the Islands (44.4%), and no members from the North-East or South are present. The cluster has the highest share of researchers (44.4%), suggesting a younger scholarly community relative to the other top clusters. The research agenda is oriented towards financial statistics and econometrics: *Systemic Risk*, *Interbank Markets*, *Time Series Analysis*, *Banking*, and *Volatility Clustering* characterize the cluster's output, with substantial

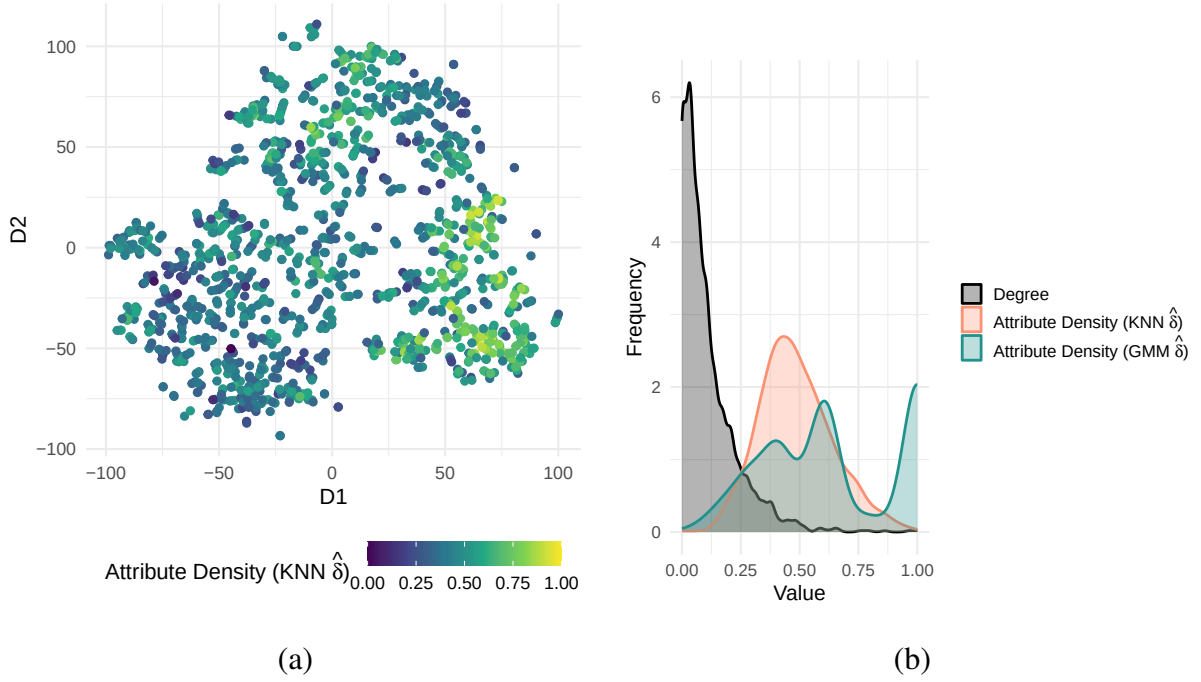


Figure 1: Distribution of node density in the co-authorship network (a) two-dimensional t-SNE projections of node attributes with lighter colours indicating higher kNN density; (b) normalised distributions of kNN density, node degree, and local density.

overlap between cluster-wide and leader keywords. This coherence between the thematic profiles of leaders and members points to a well-structured community with clear intellectual leadership around financial risk and market dynamics.

Cross-cluster patterns. Several patterns emerge from comparing the four clusters. First, senior academics (full and associate professors) consistently occupy leadership roles across all clusters, confirming that seniority and academic visibility are strong predictors of structural centrality in scientific collaboration networks. Second, *COVID-19* appears among the top keywords in all four clusters, reflecting the pervasive influence of pandemic-related research across disciplines during the observation period. Finally, while disciplinary homophily is evident in most clusters, Cluster 66 demonstrates that meaningful cross-disciplinary collaboration can emerge at the intersection of demography, sociology, and statistics, particularly around shared societal concerns such as family change and population aging.

Overall, these findings highlight the ability of the proposed approach to uncover meaningful and interpretable collaboration structures by jointly leveraging relational and attribute information. At the same time, the analysis is subject to some limitations, including the static nature of the observed network, which may not fully capture the temporal evolution of collaborations. Future research could extend this framework by incorporating dynamic network perspectives and more refined textual representations to further enhance the understanding of scientific collaboration patterns.

Table 2: Role, Sector, and Area Distribution Across the Top 4 Clusters. Shaded cells indicate the categories that characterize cluster leaders.

Variable	Level	Cluster 70		Cluster 66		Cluster 150		Cluster 94	
		N	%	N	%	N	%	N	%
Area	Center	11	20.4	12	34.3	4	14.3	3	11.1
	Islands	11	20.4	1	2.9	1	3.6	12	44.4
	North-East	17	31.5	13	37.1	8	28.6	0	0.0
	North-West	12	22.2	7	20.0	8	28.6	12	44.4
	South	3	5.6	2	5.7	7	25.0	0	0.0
Sector	p_08	14	25.9	0	0.0	0	0.0	2	7.4
	SECS-S/01	17	31.5	7	20.0	20	71.4	19	70.4
	SECS-S/02	0	0.0	0	0.0	5	17.9	2	7.4
	SECS-S/03	17	31.5	0	0.0	3	10.7	1	3.7
	SECS-S/04	2	3.7	14	40.0	0	0.0	0	0.0
	SECS-S/05	3	5.6	7	20.0	0	0.0	2	7.4
	SPS/07	1	1.9	5	14.3	0	0.0	1	3.7
	SPS/09	0	0.0	2	5.7	0	0.0	0	0.0
Role	Associate prof.	19	35.2	8	22.9	13	46.4	9	33.3
	Full professor	20	37.0	14	40.0	9	32.1	6	22.2
	Researcher	15	27.8	13	37.1	6	21.4	12	44.4
Total (cluster size)		54		35		28		27	
<i>of which: leaders</i>		<i>6</i>		<i>3</i>		<i>3</i>		<i>2</i>	

Table 3: Top 10 Keywords by Frequency Across the Top 4 Clusters. Keywords also present in the top-10 of cluster leaders are shaded.

Cluster 70		Cluster 66		Cluster 150		Cluster 94	
Keyword	N	Keyword	N	Keyword	N	Keyword	N
Firm Size Distr.	12	Caregiving	12	ARIMA	8	Estimator	7
Gibrat's Law	12	Grandchild	12	Math. Modeling	8	Quantile Regression	7
COVID-19	14	Grandparents	12	COVID-19	12	Time Series An.	7
Firm	15	Divorce	14	Bayes Theorem	13	Vary. Coeff. Model	7
Panel Data Model	16	Family Characteristic	15	Covariance Fct	13	China	8
Spatial Econom.	16	Cohabitation	16	Disease Mapping	13	COVID-19	10
Estimator	17	COVID-19	17	Estimator	13	High-Dimensional	10
China	19	Postponement	17	INLA	13	Banking	12
Max. Lik. Method	21	Birth Rate	18	Spatio-Temporal	13	Interbank Mkt	12
Cluster Analysis	— [†]	Childlessness	19	Variogram	13	Systemic Risk	12

Note: Abbreviations: INLA = Integrated Nested Laplace Approximation; Math. = Mathematical; Vary. Coeff. = Varying Coefficient; Firm Size Distr. = Firm Size Distribution; Spatial Econom. = Spatial Econometrics; Max. Lik. = Maximum Likelihood; Interbank Mkt = Interbank Markets; Covariance Fct = Covariance Function.

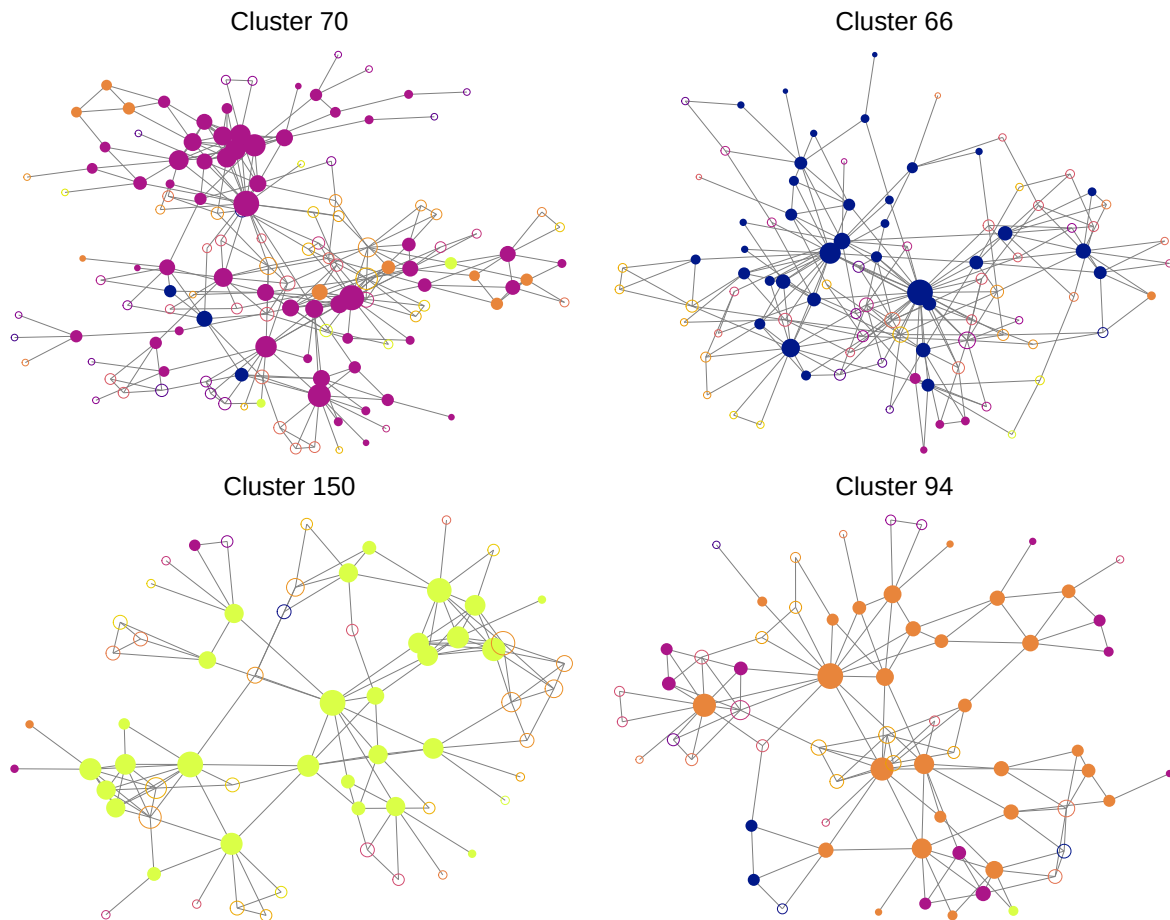


Figure 2: Networks representing direct connections of the members of the four largest clusters. Point size represents node degree, and node colour indicates community membership.

References

- Cugmas M., Ferligoj A., and Kronegger L. (2019). Scientific co-authorship networks. *Advances in network clustering and blockmodeling*, pp. 363–387.
- De Stefano D., Fuccella V., Vitale M. P., and Zaccarin S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, 35(3):370–381.
- De Stefano D., Fuccella V., Vitale M. P., and Zaccarin S. (2023). Quality issues in co-authorship data of a national scientific community. *Network Science*, 11(1):98–112.
- Fabbrucci Barbagli A. G., De Stefano D., Santelli F., and Zaccarin S. (2025). Unveiling collaboration persistence and interactions among italian academic statisticians through relational hyperevent models. *Statistical Methods & Applications*, pp. 1–19.
- Geremia S., De Stefano D., and Fop M. (2025a). Community-level core-periphery structures in collaboration networks.
- Geremia S., Fop M., and De Stefano D. (2025b). A density-based framework for community detection in attributed networks. *arXiv preprint arXiv:2512.24336*.
- Ghosh J. and Kshitij A. (2017). Examining the emergence of large-scale structures in collaboration networks: Methods in sociological analysis. *Sociological Methods & Research*, 46(4):821–863.
- Horta H., Feng S., and Santos J. M. (2022). Homophily in higher education research: A perspective based on co-authorships. *Scientometrics*, 127(1):523–543.

- Hugging Face (2024). all-MiniLM-L6-v2.
- Li Y., Zhang X., Bai X., Bai S., and Jiang Z. (2025). Predicting co-word links via heterogeneous graph convolutional networks. *Scientific Reports*, 15(1):23143.
- Lozano S., Calzada-Infante L., Adenso-Díaz B., and García S. (2019). Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature. *Scientometrics*, 120(2):609–629.
- Lu H. and Feng Y. (2009). A measure of authors' centrality in co-authorship networks based on the distribution of collaborative relationships. *Scientometrics*, 81(2):499–511.
- Lužar B., Levnajić Z., Povh J., and Perc M. (2014). Community structure and the evolution of interdisciplinarity in slovenia's scientific collaboration network. *Plos one*, 9(4):e94429.
- Menardi G. and De Stefano D. (2022). Density-based clustering of social networks. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 185(3):1004–1029.
- Ministero dell'Università e della Ricerca (2025). Official ministerial italian higher education database (mur). <https://cercauniversita.cineca.it/php5/docenti/cerca.php>. Accessed: 2026-02-19.
- Newman M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.
- Rose M. E. and Kitchin J. R. (2019). pybliometrics: Scriptable bibliometrics using a python interface to scopus. *SoftwareX*, 10:100263.
- Schäfermeier B., Hirth J., and Hanika T. (2023). Research topic flows in co-authorship networks. *Scientometrics*, 128(9):5051–5078.
- van der Maaten L. and Hinton G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(nov):2579–2605. Pagination: 27.
- Zhang Y., Pan R., Wang H., and Su H. (2023). Community detection in attributed collaboration network for statisticians. *Stat*, 12(1):e507.

Acknowledgements

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union –NextGenerationEU–Project Title “Methods for the analysis of scientific collaboration networks” –CUP J53D23011540006 - Grant Assignment Decree adopted on by the Italian Ministry of University and Research (MUR).

Assessing BERTopic stability on a specialized scientific corpus: seed sensitivity, outlier reduction, and temporal semantic drift

Carolina Vasconcelos¹, Sandro Mendonça², Bruno Damásio¹

¹NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa ,
bdamasio@novaims.unl.pt

²Business Research Unit (BRU-IUL), ISCTE – Lisbon University Institute

Abstract

Neural topic models such as BERTopic have gained popularity for analyzing large text corpora, yet their stability properties remain underexplored. We evaluate BERTopic robustness along three dimensions using a corpus of 20,565 peer-reviewed abstracts published by banking institutions between 1980 and 2023. First, we measure seed sensitivity by fitting the same model specification across 11 random seeds and comparing document assignments, topic-level consistency, and word-level overlap. Document-level agreement ranges from 51% to 94%, while top-word Jaccard similarity is more stable (57–86%). Second, we compare five outlier reduction strategies and document a coherence–coverage trade-off: the baseline model achieves $c_v = 0.72$ with 43% outliers, whereas the best post-hoc strategy reduces outliers to near zero at $c_v = 0.66$. Third, we assess temporal semantic stability by re-estimating topic representations within four decade windows and computing pairwise cosine similarities, which range from 0.76 to 0.99. These results show that BERTopic produces topics whose semantic content is more robust than their document assignments suggest, and that outlier handling and seed choice are first-order methodological decisions.

Keywords: BERTopic, topic model stability, seed sensitivity, outlier reduction, temporal semantic drift

1. Introduction

Topic models extract latent thematic structures from text corpora and have become a standard tool in computational text analysis (Blei et al., 2003; Vayansky and Kumar, 2020). Neural approaches—particularly BERTopic (Grootendorst, 2022)—combine pre-trained sentence embeddings with dimensionality reduction and density-based clustering, and often outperform classical models such as LDA on coherence benchmarks (Egger and Yu, 2022). Yet the stochastic components of this pipeline (UMAP initialization, HDBSCAN border-point assignment) raise questions about replicability that the literature has not systematically addressed.

Three stability concerns are especially relevant for applied researchers. First, UMAP (McInnes et al., 2018) uses random initialization, meaning different seeds can produce different low-dimensional representations and, consequently, different cluster assignments. Second, HDBSCAN (Campello et al., 2013) classifies a large share of documents as noise (outliers), and BERTopic offers multiple post-hoc strategies to reassign them—each with different trade-offs for coherence and coverage. Third, when a corpus spans several decades, users must decide whether topics identified on the full corpus retain stable meaning across time periods.

Prior work on topic model stability has focused on LDA (Chuang et al., 2015) or on coherence metrics (Hoyle et al., 2021; Röder et al., 2015), with limited attention to the specific instabilities introduced by the embedding–UMAP–HDBSCAN pipeline. We address this gap by evaluating

BERTopic along three dimensions: (i) seed sensitivity at the document, topic, and word levels; (ii) the coherence–coverage trade-off across five outlier reduction strategies; and (iii) temporal semantic drift measured via decade-specific topic representations.

Our test corpus consists of 20,565 peer-reviewed article abstracts published by central, development, and investment banks between 1980 and 2023, retrieved from the Web of Science. This corpus is well-suited for stability analysis: it is large enough to produce reliable topic estimates, specialized enough that topics have identifiable referents, and spans four decades, enabling temporal comparisons.

2. Data and methods

2.1. Corpus

The dataset comprises 20,565 abstracts of peer-reviewed articles authored by researchers affiliated with banking institutions, extracted from three Web of Science queries (central banks, development banks, investment banks). Documents span 1980–2023 and were deduplicated by abstract text.

This corpus provides an ideal testbed for stability analysis for several reasons. First, the substantial size (20K+ documents) enables reliable statistical assessment of stability metrics across multiple random seeds. Second, the institutional specialization creates interpretable topic boundaries: central banks focus on monetary policy, development banks on social issues, and investment banks on computational methods. This thematic coherence facilitates meaningful evaluation of semantic drift over time. Third, the four-decade span (1980–2023) allows temporal stability assessment across meaningful time windows while maintaining sufficient document density within each period.

The abstracts average 156 words (SD=89), with 95% falling between 50 and 350 words. The vocabulary contains 47,892 unique terms after standard preprocessing (lowercasing, punctuation removal, stop word filtering). Document lengths are well-suited for sentence transformer encoding, avoiding both the truncation issues common with full texts and the sparse representations that characterize very short documents such as titles or keywords.

2.2. BERTopic pipeline

We encode abstracts using the all-MiniLM-L6-v2 sentence transformer (Reimers and Gurevych, 2019), reduce dimensionality with UMAP ($n\_neighbors=100$, $min\_dist=0.1$, cosine metric), and cluster with HDBSCAN ($min\_cluster\_size=150$). Topic representations use c-TF-IDF, refined with KeyBERT and Maximal Marginal Relevance (MMR). This specification was selected from a grid search over 150 hyperparameter combinations, optimizing the product of c_v coherence and topic diversity.

2.3. Stability assessment

Seed sensitivity. We fit the model with 11 random seeds (2, 23, 24, 29, 39, 42, 48, 63, 90, 96, 97), holding all other parameters fixed. Using seed 42 as reference, we compute: (a) document-level stability—the share of documents assigned to the same topic across seeds; (b) topic-level stability—per-topic agreement rates; (c) word-level stability—Jaccard similarity of the top-10 c-TF-IDF words per topic across seeds.

The choice of 11 seeds balances computational cost with statistical power. Preliminary analysis with 50 seeds showed that agreement patterns stabilize after approximately 10 seeds, suggesting

that 11 provides sufficient coverage of the initialization space. The seeds were selected from different ranges (single digits, teens, twenties, etc.) to avoid potential clustering in the random number generator’s sequence.

Document-level stability is computed as the proportion of documents whose topic assignment matches the reference across seed pairs. Topic-level stability measures the preservation of cluster composition: for each topic in the reference model, we identify the best-matching topic in the comparison model (highest overlap) and compute the proportion of shared documents. This metric is asymmetric and measures how well each reference topic is preserved rather than overall assignment agreement.

Outlier reduction. BERTopic’s baseline model assigns 43% of documents to an outlier class (*topic* = −1). We compare five post-hoc reassignment strategies: c-TF-IDF similarity (assigns outliers to the topic with highest c-TF-IDF cosine similarity), class probabilities (uses HDBSCAN’s soft clustering probabilities), topic–document distributions (finds the most frequent topic in each outlier’s neighborhood), embedding distances (assigns to nearest cluster centroid in embedding space), and a chained strategy (c-TF-IDF followed by distributions). For each, we measure c_v coherence, topic diversity (share of unique words across topics), and remaining outlier share.

Temporal semantic stability. We partition the corpus into four decade windows (1990s: $n=1,847$; 2000s: $n=4,253$; 2010s: $n=7,892$; 2020s: $n=6,573$) and re-estimate c-TF-IDF topic representations within each window. For each topic, we compute a weighted-average word-embedding vector per decade using the all-MiniLM-L6-v2 embeddings and measure pairwise cosine similarities across all six decade pairs (1990s–2000s, 1990s–2010s, 1990s–2020s, 2000s–2010s, 2000s–2020s, 2010s–2020s). This approach captures semantic drift in topic vocabularies while controlling for changes in document composition across decades.

3. Results

3.1. Seed sensitivity

Table 1 reports document-level stability across seeds. Agreement with the reference seed clusters into two groups: four seeds achieve above 90% agreement, while six seeds fall between 51% and 57%. This bimodal pattern reflects UMAP’s sensitivity to initialization—small perturbations can flip the global orientation of the embedding space, reshuffling cluster assignments even when the local structure is preserved.

Table 1: Document-level stability: share of documents assigned to the same topic as seed 42.

Seed	24	96	90	29	2	23	97	48	39	63
Agreement	.57	.91	.93	.53	.51	.51	.55	.94	.52	.94

Topic-level stability varies from 18% (Poverty and Living Standards) to 95% (the outlier class). The lowest-stability topics tend to be those with fewer documents or more diffuse semantic boundaries. Poverty and Living Standards ($n=863$ documents) sits in a crowded region of the UMAP space near other development topics, making border assignment particularly sensitive to initialization. By contrast, Inflation and Monetary Policy ($n=3,847$ documents) achieves 89% topic-level stability, reflecting both its larger size and clearer semantic boundaries.

Word-level stability tells a different story: Jaccard similarity of top-10 words ranges from 0.57

to 0.86, and eight of twelve topics exceed 0.72. Even the least stable topic (Algorithms and Computing, Jaccard=0.57) retains core terms like “algorithm”, “network”, and “optimization” across all seeds. The most stable topic (Education and School Outcomes, Jaccard=0.86) consistently produces words like “education”, “school”, “student”, “teacher”, and “achievement”.

The gap between document-level and word-level stability indicates that while UMAP rotations reassign individual documents across runs, the resulting topics capture similar semantic content. This decoupling reflects the hierarchical nature of the BERTopic pipeline: UMAP preserves local neighborhood structure even when global orientations vary, so cluster centroids remain semantically consistent even as border documents migrate between topics. For applied researchers, this means topic *interpretations* are more robust than topic *assignments*.

3.2. Outlier reduction

Table 2 summarizes the five strategies. The baseline achieves the highest coherence ($c_v = 0.72$) but leaves 43% of documents unclassified. All post-hoc strategies reduce outliers to near zero. The distributions strategy offers the best trade-off: it retains $c_v = 0.66$ with only one remaining outlier, yielding the highest coherence $\times$ diversity score (0.66) among complete-assignment strategies. The probabilities strategy achieves full coverage but at the steepest coherence cost ($c_v = 0.57$). The chained strategy produces identical results to the distributions strategy alone, suggesting that the initial c-TF-IDF step has no incremental effect when followed by distributional reassignment.

Table 2: Outlier reduction strategies: coherence, diversity, and outlier share.

Strategy	c_v	Diversity	$c_v \times \text{Div.}$	Outliers (%)
Baseline	0.72	0.98	0.70	43.2
c-TF-IDF	0.64	0.99	0.64	<0.01
Probabilities	0.57	0.98	0.56	0.0
Distributions	0.66	0.99	0.66	<0.01
Embeddings	0.66	0.97	0.64	<0.01
Chain	0.66	0.99	0.66	<0.01

The coherence drop from 0.72 to 0.66 reflects a composition effect: documents that HDBSCAN classified as noise tend to sit between topic clusters in embedding space. Reassigning them dilutes within-topic word specificity, which lowers c-TF-IDF concentration and, consequently, c_v scores. This trade-off is inherent to density-based clustering and should be reported in any BERTopic application.

3.3. Temporal semantic stability

Table 3 presents cosine similarities between decade-specific topic representations. Most topics maintain high cross-decade similarity (> 0.85), confirming that the 12-topic structure captures durable thematic areas. Education (topic 8) is the most stable topic, with all pairwise similarities above 0.96. Algorithms (topic 5) and Governance (topic 11) show the steepest drift: their 1990s–2020s similarity falls to 0.76 and 0.78, respectively. Inspecting decade-specific top words reveals the source of this drift: the Algorithms topic shifts from “road, network, knowledge” in the 1990s to “machine learning, quantum, neural” in the 2020s—the topic’s core function (computational methods) persists, but the vocabulary tracks technological change.

A recurrent pattern emerges: adjacent-decade similarities (e.g., 2010s–2020s) tend to be higher than distant-decade similarities (e.g., 1990s–2020s), consistent with gradual vocabulary evolu-

Table 3: Temporal semantic stability: cosine similarity of decade-specific topic vectors. Topics sorted by 1990s–2020s similarity (ascending).

Topic	90–00	00–10	90–10	90–20	10–20	00–20
5-Algorithms	.86	.91	.80	.76	.91	.85
10-Governance	.85	.85	.79	.78	.93	.86
11-Labor	.87	.93	.78	.78	.88	.86
6-Housing	.94	.88	.81	.80	.99	.88
9-Poverty	.83	.76	.83	.82	.85	.80
8-Fiscal Policy	.87	.96	.85	.86	.95	.92
4-Agriculture	.97	.91	.89	.94	.88	.97
5-Climate	.95	.92	.91	.90	.98	.91
0-Inflation	.95	.94	.92	.92	.98	.94
1-Trade	.86	.91	.92	.92	.95	.93
2-Health	.98	.96	.94	.91	.91	.91
7-Education	.96	.98	.97	.97	.99	.98

tion rather than abrupt semantic shifts.

4. Methodological implications

The stability assessment reveals systematic patterns that inform best practices for BERTopic applications. The bimodal distribution of seed sensitivity (four seeds with >90% agreement, six seeds with 51–57% agreement) suggests that UMAP initialization space contains distinct attractor basins. This has three practical implications.

First, researchers should run multiple seeds and inspect both high-agreement and low-agreement cases. High-agreement seeds indicate robust local optima, while low-agreement seeds may reveal alternative but equally valid topic structures. Rather than simply averaging or selecting based on coherence, researchers should examine whether different seed groups capture substantively different topic organizations.

Second, the word-level stability findings suggest that topic interpretation should rely on multiple seeds’ word lists rather than a single model. The core vocabulary that persists across seeds (typically 60–80% of top words) represents the most reliable topic descriptors, while seed-specific words may reflect noise or borderline semantic associations.

Third, the coherence-coverage trade-off in outlier reduction creates a methodological choice that should align with research goals. Applications prioritizing interpretability (exploratory analysis, domain mapping) may prefer the baseline model’s higher coherence despite substantial outliers. Applications requiring complete document classification (downstream modeling, quantitative analysis) should use the distributions strategy, which minimizes coherence loss while achieving near-complete coverage.

The temporal stability results provide guidance for longitudinal applications. Cross-decade cosine similarities above 0.85 indicate stable semantic cores suitable for trend analysis. Similarities between 0.75–0.85 suggest gradual evolution that may require period-specific topic representations. Similarities below 0.75 indicate substantial domain shifts where single-model analysis may obscure meaningful change.

5. Discussion and conclusion

We assessed BERTopic stability along three axes and found a consistent pattern: topic *meaning* is more stable than topic *membership*. Seed variation produces bimodal document-level agreement (51–94%), but top-word overlap remains above 57% for all topics and above 72% for most. This decoupling reflects the geometry of the embedding–UMAP–HDBSCAN pipeline: UMAP rotations can reassign border documents while preserving cluster cores, so the semantic content captured by c-TF-IDF is less affected than individual assignments.

The outlier reduction comparison reveals a coherence–coverage trade-off that researchers should report explicitly. The 43% baseline outlier rate is typical for HDBSCAN on high-dimensional text embeddings, and ignoring it—or choosing a reduction strategy without justification—can substantially affect downstream analysis. The distributions strategy offers a principled default, achieving near-complete coverage with the smallest coherence loss.

Temporal analysis confirms that BERTopic topics can capture durable thematic structures over 30+ years, even when the vocabulary within topics evolves. The least stable topics are those where the underlying domain underwent rapid change (computing, governance), not those where the model fails. This distinction matters: low temporal stability can be a feature (reflecting genuine domain evolution) rather than a defect.

The temporal patterns also reveal methodological insights about BERTopic’s behavior on evolving corpora. Topics with consistent cosine similarities above 0.90 (Education, Health, Trade) represent domains where core concepts remain stable even as specific applications change. Topics with gradual decline (Algorithms: 0.86→0.91→0.80→0.76 across adjacent decades) capture domains undergoing steady vocabulary evolution. Sharp drops followed by recovery (Housing: 0.94→0.88→0.81→0.99 for 90s-00s, 00s-10s, 90s-10s, 10s-20s) suggest period-specific disruptions (in this case, the 2008 financial crisis) that temporarily altered discourse patterns.

These findings have practical implications for longitudinal topic modeling studies. Researchers working with evolving domains should expect vocabulary drift and interpret low cross-period similarities as potentially meaningful rather than problematic. The embedding-based approach of computing decade-specific cosine similarities provides a more nuanced stability measure than simple word overlap, capturing semantic relationships that pure lexical methods would miss.

Three practical recommendations follow. First, researchers should report results across multiple seeds, or at minimum acknowledge seed sensitivity. Second, outlier reduction strategy and its coherence cost should be documented. Third, for corpora spanning long periods, decade-specific word lists provide a useful diagnostic for distinguishing vocabulary drift from topic instability.

6. Application to institutional research analysis

The stability assessment informed the analysis of banking institution research patterns, demonstrating how methodological insights translate to substantive findings. The high word-level stability (>0.72 for most topics) supported interpretation of the 12-topic structure as representing durable research domains: central banks focusing on monetary policy, development banks on social issues, and investment banks complementing central bank research.

The outlier reduction analysis proved crucial for institutional comparison. The baseline model’s 43% outlier rate would have excluded nearly half the corpus from institutional attribution, fun-

damentally biasing conclusions about research specialization. The distributions strategy enabled complete institutional mapping while maintaining interpretable topic boundaries (coherence drop from 0.72 to 0.66).

Temporal stability findings validated the four-decade analysis window. Topics like Inflation and Monetary Policy (0.92 similarity across 1990s–2020s) and Education (0.97 similarity) showed sufficient stability to support claims about persistent institutional specialization. Topics like Algorithms (0.76 similarity) required careful interpretation: the vocabulary evolution from “network optimization” to “machine learning” reflects genuine domain change rather than model instability.

The seed sensitivity results shaped the robustness claims. Document-level assignment differences (51–94%) meant that specific institutional collaborations varied across seeds, but word-level consistency (0.57–0.86 Jaccard similarity) ensured that topic characterizations remained stable. This distinction allowed strong claims about *what* institutions study while exercising appropriate caution about *specific* research partnerships.

These applications demonstrate that BERTopic stability assessment enables more nuanced and defensible substantive conclusions. Rather than treating topic models as black boxes, systematic stability evaluation provides the methodological foundation for confident interpretation of domain-specific findings.

References

- Blei D. M., Ng A. Y., and Jordan M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Campello R. J. G. B., Moulavi D., and Sander J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, pp. 160–172. Springer.
- Chuang J., Roberts M. E., Stewart B. M., Weiss R., Tingley D., Grimmer J., and Heer J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 529–539.
- Egger R. and Yu J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7:886498.
- Grootendorst M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Hoyle A., Goel P., Peskov D., Hian-Cheong A., Boyd-Graber J., and Resnik P. (2021). Is automated topic model evaluation broken? The incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34.
- McInnes L., Healy J., and Melville J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Reimers N. and Gurevych I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics.
- Röder M., Both A., and Hinneburg A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 399–408.
- Vayansky I. and Kumar S. A. P. (2020). A review of topic modeling methods. *Information Systems*,

94:101582.

C16

Andrea Camilleri and Sellerio corpora

09-lug
12:00 - 13:30
Aula 12

Camilleri’s Sicilian: From Stopwords To Topic Modelling

Vincenzo Giuseppe Genova¹, Andrea Sciandra²

¹Department of Economics, Business, and Statistics. University of Palermo –
vincenzogiuseppe.genova@unipa.it

²Department of Philosophy, Sociology, Education, and Applied Psychology & Department of
Statistical Sciences. University of Padua – andrea.sciandra@unipd.it

Abstract

Andrea Camilleri represents a peculiar case in Italian narrative for the conscious use of a mixed language, which combines Italian and Sicilian. This linguistic choice generates a unique lexicon, distant from the conventions of standard Italian and only partially attributable to codified Sicilian dialect; this represents a challenge for computational linguistics. In this work we present the construction of a stop-words list dedicated to the Sicilian lexicon present in Camilleri’s works, aimed at the automatic analysis of topics. Starting from a corpus that collects the 70 works of Camilleri published by Sellerio, Sicilian lemmas were identified through a manual annotation procedure that combined native-speaker expertise with systematic consultation of the Sicilian vocabulary of the *Centro di Studi Filologici e Linguistici Siciliani* and the *Dizionario Camilleriano/Italiano* by Mario Genco. The objective of the constructed list is to distinguishing dialectal forms from Italian lexicon. Subsequently, stopwords were identified and words with semantic relevance were selected. The result is a Camilleri’s Sicilian lexicon, which collects morphosyntactic information. This resource constitutes a useful contribution both to the automatic processing of natural language in the context of the Sicilian dialect and to authorial characterisation and the study of linguistic variability in literary dialectal contexts. As a first application, we estimated a Structural Topic Model incorporating this dialect-specific information obtaining interpretable topics that align with the series. Our results show that a targeted author-specific dialectal annotation with structural topic models and document-level covariates offers a useful starting point for studying other Italian regional and dialectal literary varieties where standard-language tools are inadequate.

Keywords: Sicilian dialect lexicon, Stopwords annotation, Computational linguistics resource, Camilleri’s Sicilian

1. Introduction

Andrea Camilleri has often been described as a paradigmatic case of contemporary Italian narrative intersecting standard Italian, regional Italian of Sicily, and Sicilian dialect, where spelling, phonology, morphology, and phraseology are systematically remodelled to construct a recognisable idiolect (Marci et al., 2021; Matt, 2025; Russi, 2020). This “vigatese” variety intensifies Sicilian features at multiple levels—graphic, phonetic, morphological, and lexical—while preserving largely Italian syntactic macro-structures, thus producing a hybrid language that maps either onto codified Sicilian or onto standard Italian (Marci et al., 2021; Rushing, 2014; Russi, 2020). In the context of Italian post-unitary literature, critics have linked this hybrid code to broader trends of refunctionalisation of dialect and regional forms for narrative, expressive, and ideological purposes, situating Camilleri among authors such as Gadda, Pasolini, and Sciascia, but with an unprecedented degree of systematic code-mixing (Marci et al., 2021; Matt, 2025). From a computational perspective, such an idiolect poses non-trivial challenges: canonical tokenisation, lemmatisation, and POS-tagging tools that presuppose a stable lexicon aligned with standard languages, instead of author-specific orthographies, hybrid forms, and dialectal spellings such as those of Camilleri (Rushing, 2014; Russi, 2020). The literature focused

on Camilleri’s language has expanded significantly in the last decade, especially through the *Quaderni camilleriani* series and the *CamillerINDEX* project, which have provided detailed analyses of phonography, morphosyntax, lexicon, style, and interlingual aspects (Demontis, 2021). Sociolinguistic and stylistic studies have examined the relationship between Camilleri’s language and spoken Sicilian, the representation of regional Italian, and how shifts between Italian and Sicilian index social strata, institutional roles, and degrees of familiarity between characters, highlighting the systematic stratification of registers and voices (Marci et al., 2021; Matt, 2025). Authors such as Villa (2021) and Rushing (2014) have discussed the difficulties of translating Camilleri’s non-standard features—dialect, regionalisms, code-switching, eye dialect—into English and other languages, as well as the range of strategies from domestication to foreignization adopted by translators. These studies focus on the role of language mixing and dialectal variation in Camilleri’s work, which offers a qualitative analysis of linguistic forms. However, they rarely provide machine-readable lexica that can be directly incorporated into computational processing, nor consistently isolate and encode “Camilleri’s Sicilian” as an autonomous lexical system. Indeed, to the best of our knowledge, no linguistic resources are available for the “vigatese” language that can be used in the context of topic modelling and sentiment analysis. Likewise, current Italian sentiment lexica do not cover the specific Sicilian or Sicilianised items that are distinctive of Camilleri’s works, and existing topic-modelling applications to Italian literary texts presuppose standard Italian pre-processing without dedicated treatment of dialectal stopwords or variety-specific function words (Basile and Nissim, 2013; Heibi et al., 2020; Monachini, 2014). This double gap—on the one hand, the absence of a Camilleri-specific machine-readable lexicon; on the other hand, the lack of non-standard resource for topic-modelling—motivates the present study. Specifically, starting from a corpus of 70 Camilleri’s novels and story collections, we operationalise “Camilleri’s Sicilian” by systematically identifying and classifying dialectal and hybrid forms at the token and lemma level, and by deriving from them a manually constructed list of dialectal stopwords that are not captured by standard Italian stopword lists and would otherwise introduce noise into textual analyses and topic models. In this paper we focus, as a first application, on a structural topic model fitted to the Montalbano subset of the corpus (34 novels and short-story collections), in order to explore how themes and the use of Sicilian evolve within this series.

The paper is structured as follows: Section 2. provides an overview of the corpora used and describe the construction of lexicon of Camilleri’s Sicilian; Section 3. outlines the structural topic model used; Section 4. presents the findings and results of the analysis. Finally, in Section 5., we will offer some conclusions.

2. Corpus and Stop-Words List Construction

The corpus used in this study consists of 70 works by Andrea Camilleri, provided by the publisher Sellerio for research purposes. The selection covers both the *Montalbano* series (novels and short stories, including the “young Montalbano” cycle) and stand-alone works of different types (e.g. crime fiction, historical novels, the “metamorphosis” trilogy). Each work was converted into plain text, and paratextual materials such as prefaces, presentations, interviews and appendices were systematically removed, so as to focus on Camilleri’s narrative prose. Within this corpus, we aimed to operationalise “Camilleri’s Sicilian” by manually identifying Sicilian and Sicilianised forms. This work was carried out by one of the authors, a native Sicilian speaker, who inspected the corpus and marked non-standard Italian items. In order to reduce subjectivity, ambiguous cases were resolved by systematically consulting the Sicilian vocabulary of the *Centro di Studi Filologici e Linguistici Siciliani* (Piccitto et al., 2002), the *Dizionario*

Camilleriano/Italiano by Genco (2001), and the CamillerINDEX (2018) project. The resulting set includes both “pure” Sicilian forms and hybrid Italian–Sicilian spellings characteristic of Camilleri’s idiolect. For each word identified as belonging to this layer, a lemma was assigned and, when appropriate, a standard Italian equivalent was recorded. The resulting lexicon is stored in a simple tabular format including, for each entry, the dialectal form, its lemma, an optional Italian equivalent and frequency counts in the 70-book corpus. Thus, on this basis, we compiled a preliminary stop-words list of Camilleri’s Sicilian, listing all dialectal and hybrid lemmas present in the 70 texts. At this stage the focus is lexical: we do not attempt a full morphosyntactic analysis, but we provide a systematic inventory of the non-standard forms that constitute the backbone of Camilleri’s mixed code. Furthermore, starting from the manually compiled Sicilian lexicon, we isolated those stopwords in Camilleri’s mixed code that are not captured by standard Italian stopword lists. This dialect-specific stopword list is designed as a list for preprocessing Camilleri’s works (and, potentially, similar Sicilian-influenced corpora), and is used in the topic modelling analysis presented in this paper. The lexicon and the dialect-specific stopword list are made available as an open resource in the public GitHub repository (https://github.com/vincenzogioseppegenova/camilleri_project), so that other researchers can inspect, reuse and extend our annotations for further studies on Camilleri and other Sicilian-influenced corpora.

3. Structural Topic Modelling

The Structural Topic Model (STM) is a generative model for word counts that can be used to identify topics and estimate their relationship to external data, i.e. data available on texts. A topic is defined as a mixture of words; each word in the vocabulary has a probability of being selected, and the sum of the word probabilities for a given topic is equal to 1. Similarly, a text is a mixture of topics; a single speech can comprise multiple topics, and the sum of topic proportions in each text is equal to 1. The STM can be divided into three conceptual components:

(1) A topic prevalence model, specified by Equations (1) and (2), which controls how words are allocated to topics as a function of covariates:

$$\gamma_k \sim \text{Normal}_p(0, \sigma_k^2 I_p), \quad \text{for } k = 1 \dots K-1, \quad (1)$$

$$\theta_d \sim \text{LogisticNormal}_{K-1}(\Gamma' x_d, \Sigma), \quad (2)$$

where Γ is the matrix of coefficients for the topic prevalence model specified by Equations (1) and (2), d stands for text, k is the number of topics, X_d is the matrix for topic prevalence, and σ is a k -dimensional hyper-parameter vector;

(2) A topical content model, which controls the frequency of the terms in each topic as a function of covariates:

$$\beta_{d,k,v} = \frac{\exp(m_v + k_{k,v}^{(t)} + k_{y_d,v}^{(c)} + k_{y_d,k,v}^{(i)})}{\sum_v \exp(m_v + k_{k,v}^{(t)} + k_{y_d,v}^{(c)} + k_{y_d,k,v}^{(i)})} \quad \text{for } v = 1 \dots V \text{ and } k = 1 \dots K, \quad (3)$$

where $k_{k,v}^{(t)} + k_{y_d,v}^{(c)} + k_{y_d,k,v}^{(i)}$ is a collection of coefficients for the topical content model, and m_v is the marginal log-transformed rate of term v ;

(3) A core observation model, which combines models (1), (2), and (3):

$$\vec{z}_{d,n} \sim \text{Multinomial}_k(\vec{\theta}_d), \quad \text{for } n = 1 \dots N_d, \quad (4)$$

$$\vec{w}_{d,n} \sim \text{Multinomial}_v(\vec{B} \vec{z}_{d,n}), \quad \text{for } n = 1 \dots N_d, \quad (5)$$

The core observation model uses the logistic normal distribution to allow for correlations in topic proportions (Roberts et al., 2016). Topic prevalence describes a text’s association with a topic, while topical content describes how words are used within a topic. These components allow the expected text–topic proportion and the topic–word probability, respectively, to vary as a function of observed text-level covariates, rather than arising from a single prior shared by all texts, as is the basic assumption of Latent Dirichlet Allocation (LDA) topic modelling.

In our application, there are two covariates: year and typology of book (novels or collection of stories). To account for the effect of time, we also used splines, applying a smoothed B-spline function (with 10 degrees of freedom) to the “year” variable to allow for a non-linear relationship.

The corpus pre-processing phase initially involved tokenising the terms, converting them all to lower case, and removing punctuation, symbols and numbers. The books were then divided into chunks of 500 words, and only terms with a minimum frequency of 50 occurrences were selected. Upon completion of these operations, the corpus for topic modelling consisted of 3,741 chunks and 2,519 word types. Further information on the overall corpus and the subcorpus relating to the works in the Montalbano cycle can be found in Table 1 and Figure 3 of Arjuna Tuzzi’s paper (Tuzzi, 2026).

4. Results

As mentioned in Section 2., the STM was fitted on a subset of our corpus comprising 34 Montalbano novels and short-story collections, with $K = 3$ topics modelled as a function of year and book type (novel *vs* short stories). We used the Consistency and Differentiation score (CD score) to determine the ideal number of topics K (Sciandra et al., 2025). CD score synthesises the trade-off between semantic coherence and exclusivity by calculating the L_2 norm of the two min–max normalized measures. In particular, with $K=3$, both the CD score and, specifically, its semantic coherence component are maximised, whereas measures such as the held-out log-likelihood and model residuals would favour a higher number of topics (10). This result is frequently encountered when using such diagnostics (Sciandra et al., 2023), so we chose to favour the more parsimonious solution with higher semantic coherence. Furthermore, increasing the number of topics makes it more likely to find correlations and thus potential redundancy among topics, whereas in the case of $K=3$ no significant correlations were detected. Finally, by calculating the distinctiveness (Chuang et al., 2012) of the terms, we measured the degree to which words are shared across topics. Distinctiveness corresponds to the Kullback–Leibler (KL) divergence between the topic distribution given a word and the marginal topic distribution. On average, distinctiveness is 0.477, whilst a value close to 0 would indicate that the words occur in all topics. Specifically, Topic 3 contributes more than the others to distinctiveness (0.19635).

Table 1 summarises, for each topic, the main lexical items ranked according to Highest Probability and FREX index—that balances exclusivity and word frequency, using the weighted harmonic mean (Bischof and Airolidi, 2012)—for the model in Eq. (3).

Looking at Table 1, Topic 1 is represented by words referring to domestic spaces (*casa, porta, macchina*), temporal markers (*anni, tempo, giorno, volta*), and bodily perception (*mano, testa, occhi*), together with the name *Livia*. This cluster suggests a thematic focus on Montalbano’s private life and interactions in non-institutional settings. Indeed, within this topic, *Livia* is particularly important. Her name appears among the most probable words, which means that a large part of the text where Livia is mentioned is associated with Topic 1. This is consistent with her narrative function. Livia does not work at the police station and does not belong to

Table 1: Top words for the topic model presented in Eq. (3), ranked according to Highest probability and FREX index.

Topic	Label	Highest Probability	FREX index
1	Montalbano's private sphere	casa, signora, anni, volta, porta, macchina, mano, livia, testa, occhi, tempo, giorno, parte, andare, momento	pareva, ingegnere, padre, luce, michela, poteva, rispose, voce, teneva, vide, preside, bocca, veniva, andava, giacomo
2	Police station and institutions	fazio, dottore, dottori, catarella, dottor, augello, signor, questore, commissariato, ufficio, pasquano, tempo, montelusa, elena, avvocato	pasquano, tommaseo, elena, dottor, catarella, dottore, questore, furto, alderighi, bonetti, sinagra, balduccio, spitaleri, bonfiglio
3	Dialect and popular speech	fazio, mano, casa, livia, porta, testa, tempo, momento, occhi, meglio, voci, vinni, parti, picciotta, augello	fusse, mari, vinni, vota, liliana, farisi, sintiva, annare, cellulari, fini, autri, enzo, viduto, buttiglia, vabbeni

the group of institutional characters that dominate Topic 2. She is, instead, the main reference point for Montalbano's life outside work. Scenes with Livia typically involve phone calls, visits to Marinella, arguments and reconciliations that characterise Montalbano's private sphere with Livia. In contrast, Topic 2 captures the institutional and procedural side of the series. Indeed, among the highest-probability words there are names of colleagues and institutional actors (*fazio, catarella, augello, questore, pasquano*), professional titles (*dottore, dottori, avvocato*) and institutional spaces (*commissariato, ufficio, montelusa*). These terms identify the topic inside the world of the police and the justice system, rather than in domestic or informal settings. FREX terms such as *pasquano, tommaseo, furto, alderighi, sinagra, balduccio, spitaleri, bonfiglio* mostly refer either to crimes (e.g. *furto*) or to actors who appear in judicial and investigative contexts (magistrates, suspects, members of criminal networks). Taken together, these elements suggest that Topic 2 focuses on passages describing how cases are handled: interactions within the police station, exchanges with the *questore* and magistrates, technical discussions with the coroner, and references to suspects and criminal networks. Regarding Topic 3, it highlights a linguistic dimension. Indeed, Topic 3 partly overlaps lexically with Topic 1 (for example *casa, mano, livia*), showing that it can also appear in familiar spaces and involving core characters. The peculiarity of this topic, however, is the strong presence of Sicilian dialect forms among the FREX words (*fusse, mari, vinni, vota, sintiva, annare, viduto, vabbeni*). These forms are typical of the hybrid Italian–Sicilian register that critics have identified as one of Camilleri's main stylistic signatures, involving code-mixing between standard Italian, regional Italian and dialect, used to mark characters, social groups and local settings (Marci et al., 2021; Matt, 2025; Russi, 2020). Thus this topic seems to be tied to the Vigata context suggesting it is not defined by a either private or institutional storyline, but by a way of speaking. In other words, this topic characterised those scene in which the narrator or the characters make heavy use of Sicilian forms.

Table 2 reports the estimated coefficients for the effects of year and text type on topic preva-

lence. Regarding Topic 1, the large and positive intercept suggests that for the short stories, at the beginning of The Montalbano series, the “Montalbano’s private sphere” topic has high prevalence. This prevalence shows a reduction in the subsequent years of publication. Indeed, almost all spline coefficients are negative and significant, which implies a decreasing trend in Topic 1 prevalence over time. The coefficient for the novel type is not significant, suggesting no systematic difference between novels and short-stories.

Table 2: Estimated effects of year (spline) and text type on topic prevalence for the model in Eq.s (1) and (2)

Covariate	Topic 1	Topic 2	Topic 3
Intercept	0.95203***	0.02752	0.02018
s(year)1	−0.11219**	0.08723*	0.02502
s(year)2	−0.03309	0.07069 [†]	−0.03743
s(year)3	−0.24179***	0.21817***	0.02356
s(year)4	−0.28601***	0.11191***	0.17481***
s(year)5	−0.78255***	0.48218***	0.30021***
s(year)6	−0.75969***	0.07135*	0.68820***
s(year)7	−0.81747***	0.34267***	0.47484***
s(year)8	−0.72508***	0.00873	0.71664***
s(year)9	−0.74391***	0.20798***	0.53543***
s(year)10	−0.77692***	0.17342***	0.60434***
text type = novel	−0.01412	0.03070*	−0.01622

Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; [†] $p < 0.10$.

For Topic 2, most spline coefficients are positive and statistically significant, pointing to an increasing prevalence of the “police station and institutions” topic over the publication timeline. In other words, as the series progresses, a larger proportion of the text is devoted to procedural aspects, institutional actors and investigative routines. Moreover, the text-type coefficient is positive and significant: novels exhibit a higher prevalence of Topic 2 than short-story collections. This is consistent with the publication type. Indeed, longer forms allow for more extensive development of institutional settings and recurring procedural patterns (Taffarel, 2012). Finally, the dialect/popular-speech topic shows a temporal pattern, with positive and highly significant spline coefficients from the mid-range onward. This suggests a growth in the use of the hybrid Italian–Sicilian register in later works. Regarding the text-type, it is not significant, suggesting that this trend is shared by both novels and short-story collections. This finding aligns with linguistic analyses of Camilleri that describe a progressive use of regional language and code-mixing over time (Longhitano Piazza, 2001; Russi, 2020; Sottile et al., 2019; Tomaiuolo, 2009).

5. Conclusions

In this work we presented a corpus-based study of Andrea Camilleri’s language that combines the manual construction of a Sicilian-based stop-words list with the application of structural topic modelling to the *Montalbano* series. Starting from a corpus of 70 works, we operationalised “Camilleri’s Sicilian” by identifying non-standard items and isolating a set of dialectal stop-words. This resource, while still preliminary, provides a machine-readable lexicon of the hybrid Sicilian that characterised Camilleri’s idiolect. As a first application, we estimated a Structural Topic Model incorporating this dialect-specific information obtaining interpretable

topics that align with the series. One topic captures Montalbano's private sphere, centred on domestic spaces, temporal routines and the character of Livia; a second topic concentrates on institutional actors and procedural settings, highlighting the centrality of the police station and judicial machinery; a third topic isolates the dialect/popular-speech register, characterised by Sicilian and Sicilianised forms associated with the Vigàta context (Marci et al., 2021; Matt, 2025). Overall, the use of the temporal covariate in the topic estimation process made it possible to observe patterns of increase and decrease in the prevalence of topics within the texts. The estimated parameters of the covariate effects suggest a rebalancing over time: the "private sphere" topic is more prevalent in early short-story collections and declines over the publication timeline, while the "police station and institutions" topic becomes more prominent, particularly in the novels. At the same time, the dialect/popular-speech topic displays an overall increasing trend that is shared by both novels and short-story collections, consistent with qualitative analyses that describe a progressive intensification and normalisation of Camilleri's hybrid Italian-Sicilian code (Ala-Risku, 2024; Matt, 2025; Russi, 2020). Furthermore, our results show that a targeted dialectal lexicon and stopword list can support topic modelling on literary texts with non-standard language by reducing noise and clarifying thematic structures and patterns of linguistic variation. Combining expert-driven, author-specific dialectal annotation with structural topic models and document-level covariates offers a useful starting point for studying other Italian regional and dialectal literary varieties where standard-language tools are inadequate. It is important to note that this work has several limitations. The lexicon was compiled by a single annotator, which is a limitation in terms of inter-annotator validation. However, the annotation was performed by a native speaker of Sicilian who systematically consulted the Sicilian vocabulary of the *Centro di Studi Filologici e Linguistici Siciliani* (Piccitto et al., 2002), the *Dizionario Camilleriano/Italiano* (Genco, 2001), and the CamillerINDEX (2018) project to resolve ambiguous cases. Moreover, our lexicon is still incomplete in terms of morphosyntactic information, and the STM analysis is restricted to the *Montalbano* corpus. In future research, we plan to enrich the annotation with sentiment and emotion labels and to extend the analysis to non-*Montalbano* works. Finally, by focusing solely on the Montalbano series, we were able to narrow down the themes, given Camilleri's extensive body of work and the variety of genres he covers (playwriting, letters, dialogues, etc.).

Acknowledgement

The authors are grateful to Sellerio Editore for granting access to the digitised archives of Andrea Camilleri's works used in this study.

References

- Ala-Risku R. (2024). The metalinguistic dimension of literary multilingualism: Linguistic biographies in Italian fiction. *Forum for Modern Language Studies*, 60(1):129–144.
- Basile V. and Nissim M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 100–107.
- Bischof J. and Airoldi E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th international conference on machine learning (icml-12)*, pp. 201–208.
- CamillerINDEX (2018). Progetto di glossario e banca dati dell'opera di Andrea Camilleri. A cura del progetto CamillerINDEX.it.
- Chuang J., Manning C. D., and Heer J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces*,

pp. 74–77.

- Demontis S. (2021). *Quaderni camilleriani 2021. Bibliografia tematica sull'opera di Andrea Camilleri*. Grafiche Ghiani.
- Genco M. (2001). Dizionario camilleriano-italiano. Pubblicato sul «Giornale di Sicilia», gennaio 2001.
- Heibi I., Peroni S., Pareschi L., and Ferri P. (2020). Mitao: a tool for enabling scholars in the humanities to use topic modelling in their studies. *arXiv preprint arXiv:2011.13886*.
- Longhitano Piazza S. (2001). Parole proprie e parole di altri: la lingua di Andrea Camilleri.
- Marci G., Ruggerini M. E., Szőke V., and Ghiani G. (2021). *Quaderni camilleriani 15. Sotto la lente: il linguaggio di Camilleri*. Grafiche Ghiani.
- Matt L. (2025). *Quaderni camilleriani 26. Lingue e stili di Camilleri*. Grafiche Ghiani.
- Monachini M. (2014). Italian sentiment lexicon. <https://www.cnr.it/it/banche-dati-istituti/banca-dati/1029/italian-sentiment-lexicon>. Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR). Italian polarity lexicon derived from ItalWordNet.
- Piccitto G., Tropea G., and Trovato S. C. (1977–2002). *Vocabolario siciliano*. Centro di studi filologici e linguistici siciliani, Palermo–Catania. Opera in 5 volumi.
- Roberts M. E., Stewart B. M., and Airolidi E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111(515):988–1003.
- Rushing R. A. (2014). The Language of Andrea Camilleri. <http://faculty.las.illinois.edu/rrushing/440c/camilleri.html>.
- Russi C. (2020). *Sicilian Elements in Andrea Camilleri's Narrative Language: A Linguistic Analysis*. Bloomsbury Publishing PLC.
- Sciandra A., Trevisani M., and Tuzzi A. (2023). Diagnostics for topic modelling. the dubious joys of making quantitative decisions in a qualitative environment. In *Proceedings of the Statistics and Data Science Conference*, pp. 61–66. Pavia University Press.
- Sciandra A., Trevisani M., and Tuzzi A. (2025). Temporal trends and presidential traits in the italian end-of-year addresses: comparing and contrasting kbs and stm results: A. Sciandra et al. *Quality & Quantity*, 59(1):905–935.
- Sottile R. et al. (2019). La lingua “inventata” di Andrea Camilleri: il peso della parola dialettale. *Dialoghi mediterranei*, 39:1–10.
- Taffarel M. (2012). Un'analisi descrittiva della traduzione dei dialoghi dei personaggi di Andrea Camilleri in castigliano.
- Tomaiuolo S. (2009). “I am Montalbano/Montalbano sono”: fluency and cultural difference in translating Andrea Camilleri's fiction.
- Tuzzi A. (2026). A new investigation for Sellerio. In *JADT2026 proceedings*.
- Villa M. (2021). Translating non-standard language: Andrea Camilleri in english. Master's thesis, Università degli Studi di Padova.

Assunti teorico-metodologici su linguaggio letterario esclusivo e dati testuali

Pietro Mazzarisi¹

¹University of Trieste – pietro.mazzarisi@units.it

Abstract

This study investigates the relationship between discriminatory literary language and authorial agency through a theoretical and methodological framework termed Diegetic-Mimetic Axis Analysis (DM^A). The aim is to determine whether, and to what extent, the use of denigratory or discriminatory words in literary texts can be attributed to the author or traced back to characters' voices and narrative dynamics. The analysis focuses on the linguistic representation of migrant figures across ten words, examining their distribution, frequency, and diachronic variation. The results reveal a shift from widespread use of denigratory words toward the progressive adoption of more neutral and inclusive lexical forms. The framework rests on three theoretical assumptions: *transversality* (discriminatory language appears at both diegetic and mimetic levels), *coincidence* (linguistic forms overlap across these levels), and *diachrony* (usage evolves consistently over time). Integration with non-fiction texts further supports the interpretation of authorial choices, offering a replicable model for the study of discriminatory language in literary and cultural contexts. The methodology combines corpus linguistics, distant reading, and differential analysis applied to the 102 texts in the *Commissario Montalbano* series.

Keywords: exclusive language, authorial agency, migrants, diachronic analysis.

Abstract 2

Questo studio indaga il rapporto tra linguaggio letterario discriminatorio e agency autoriale attraverso un quadro teorico e metodologico denominato *Analisi dell'asse diegetico-mimetico* (DM^A). L'obiettivo è determinare se, e in che misura, l'uso di parole denigratorie o discriminatorie nei testi letterari possa essere attribuito all'autore o ricondotto alle voci dei personaggi e alle dinamiche narrative. L'analisi si concentra sulla rappresentazione linguistica delle figure dei migranti attraverso dieci parole, esaminandone la distribuzione, la frequenza e la variazione diacronica. I risultati rivelano un passaggio dall'uso diffuso di parole denigratorie verso la progressiva adozione di forme lessicali più neutre e inclusive. Il quadro si basa su tre presupposti teorici: *trasversalità* (il linguaggio discriminatorio appare sia a livello diegetico che mimetico), *coincidenza* (le forme linguistiche si sovrappongono tra questi livelli) e *diacronia* (l'uso si evolve in modo coerente nel tempo). L'integrazione con testi di saggistica supporta ulteriormente l'interpretazione delle scelte autoriali, offrendo un modello replicabile per lo studio del linguaggio discriminatorio in contesti letterari e culturali. La metodologia combina la linguistica dei corpora, il distant reading e l'analisi differenziale applicati ai 102 testi della serie del Commissario Montalbano.

Parole chiave: linguaggio esclusivo, agency autoriale, migranti, analisi diacronica.

1. Introduzione

Il *politically correct* (Bernstein, 1990) ha mutato la sensibilità linguistica rispetto a ciò che viene detto, e come viene detto, sulle alterità e sulle minoranze sociali. Anche gli studi letterari hanno iniziato a interrogarsi sulla funzione degli stereotipi nelle narrazioni chiedendosi se, come e quanto chi scrive condivide il punto di vista espresso da personaggi e dalle voci narranti che ha creato oppure se l'impiego di determinate parole piuttosto che altre sia atto a riportare aspetti della realtà a chi legge. Ma quando è effettivamente possibile associare una figura

autoriale ai dati testuali del linguaggio esclusivo, ovvero alle espressioni denigratorie dei suoi personaggi e/o delle sue voci narranti? Ovvero, quando è possibile dunque inferirne l'*agency* autoriale? Le parole ricorrenti per le rappresentazioni delle figure migranti e BIPOC (Black, Indigenous and People of Color) - e in uno spettro più ampio per le figure di chi non è uomo, *cisgender*, bianco, occidentale, normodotato e longilineo - rimangono uno degli accessi privilegiati per la conduzione di studi sul linguaggio esclusivo poiché offrono i dati testuali necessari per indagare tali fenomeni. Ma è possibile stabilire se e quando gli usi di parole denigratorie in tali sfere siano fattivamente imputabili a chi ha scritto i testi? Nei testi letterari, infatti, tali usi possono occorrere nei personaggi e/o possono essere ascrivibili a una o più voci narranti e, quando attestati ai soli personaggi, gli usi possono essere sistemici e atti a una connotazione culturale e sociolinguistica oppure possono essere episodici e andrebbero diversamente interpretati. Per indagare queste tematiche, la ricerca mostra gli assunti teorici e la metodologia computazionale denominata *Analisi dell'asse diegetico-mimetico* (*Diegetic-Mimetic Axis Analysis, DM^A*, Mazzarisi, 2025) utile per fissare l'*agency* autoriale rispetto a tali aspetti delle rappresentazioni artistico-letterarie. Qui si fa riferimento al linguaggio esclusivo all'interno dei testi letterari di Andrea Camilleri in cui sono presenti figure di migranti, ma il quadro teorico-metodologico risulta utile anche per indagare il linguaggio esclusivo riguardo al gender, all'antisemitismo, alla disabilità e in altri ambiti in cui la discriminazione linguistica riflette le discriminazioni sociali. La metodologia è utile anche per stabilire se la scrittura letteraria è concretamente critica verso le strutture ideologiche e linguistiche massmediali, oppure si trova in linea e influenzata da queste. Il campione testuale camilleriano scelto per la ricerca sono tutti i testi che compongono una delle serialità letterarie di maggior successo in Italia negli ultimi trent'anni: *Le indagini del commissario Montalbano*, un *corpus* di 102 testi suddivisi in 28 romanzi e 74 racconti pubblicati tra il 1994 e il 2022 (si esclude il romanzo *Acqua in bocca*, l'unico con Montalbano ma scritto insieme a un altro autore, Carlo Lucarelli).

2. Teoria, metodologia, corpus

Prima di additare un testo come esclusivo, è base di partenza lo stabilire se le parole dispregiative o esclusive, ovvero i dati testuali, ricorrano limitatamente ai personaggi o se siano espresse dalla voce narrante e, in ambo i casi, in che misura e in che modo vengano impiegate. È necessario dunque prendere in considerazione diverse definizioni e assunti teorici prima di condurre un'inferenza sui dati testuali autoriali che avalli la presenza di linguaggio esclusivo nell'opera artistico-letteraria. Le definizioni riguardano i concetti letterari chiave di diegesi e mimesi. Secondo la tradizione classica e narratologica, semplificando per diegesi si intende l'universo del narrato mentre per mimesi si intendono le voci dei personaggi tramite il discorso diretto (Platone, 1990; Aristotele 1964; Genette, 1969). Gli assunti teorici includono tre concetti chiave: "trasversalità", "coincidenza" e "diacronia". Nello specifico: (A) l'assunto teorico della transversalità si verifica quando l'uso di stereotipi, linguaggio dispregiativo e parole negative in un'opera è trasversale, ovvero compare sia nei dati testuali diegetici che nei dati testuali mimetici; (B) l'assunto teorico della coincidenza si verifica quando questi dati testuali sono coincidenti — ovvero le parole utilizzate nelle porzioni testuali mimetiche sono altresì presenti nelle porzioni testuali diegetiche; (C) infine, l'assunto testuale della diacronia si verifica quando vi è una coerenza diacronica, dunque in uno studio longitudinale, nell'uso di queste forme, consentendo inferenze circa tendenze, asintoti e intersezioni tra fattori personali e culturali rilevanti per il periodo storico in cui le opere sono state create. Quali critiche è possibile muovere contro questo quadro teorico per la fissazione dell'*agency* autoriale nel linguaggio letterario esclusivo? Una critica potrebbe essere che l'attenzione ai dati testuali potrebbe rendere difficile cogliere i modi in cui un autore può usare un linguaggio esclusivo in modo

ironico o sovversivo. Anche se il linguaggio esclusivo appare nella diegesi, potrebbe essere usato per criticare, piuttosto che per avallare, tale atteggiamento esclusivo. Sebbene il linguaggio possa sembrare esclusivo, la narrazione più ampia potrebbe in definitiva riguardare la critica al pregiudizio o alla xenofobia. Pertanto, una critica sarebbe che il quadro teorico-metodologico può trascurare sfumature quali l'ironia, la parodia o la satira nella voce del narratore (diegesi), specialmente negli autori in cui l'umorismo e la sottile critica sociale sono spesso integrati nella narrazione. Per ovviare a possibili debolezze di questo tipo, l'assunto teorico sulla diacronia (C) accanto all'analisi longitudinale prevede l'analisi comparativa con gli eventuali testi di non-fiction di una data autorialità, ovvero con i testi di saggistica. L'aspetto longitudinale dell'analisi, che consiste nel tracciare i cambiamenti nell'uso del linguaggio nel corso del tempo, aiuta a individuare con precisione quando e come muta la sensibilità autoriale. La dimensione comparativa risulta necessaria per comprendere lo sviluppo dell'autonomia autoriale. Se il linguaggio esclusivo autoriale riflette le tendenze, gli asintoti e le intersezioni tra i rapporti di potere, le narrazioni prevalenti nei mass media e i fattori culturali rilevanti del periodo in cui i testi sono stati licenziati, l'ipotesi d'un impiego del linguaggio esclusivo nei testi finalizzato all'ironia, alla parodia, alla satira o alla sovversione perde consistenza.

La metodologia prevede la suddivisione di ogni testo letterario del corpus in due distinte porzioni testuali, quella diegetica e quella mimetica. In presenza di un basso numero di output dei dati testuali ricercati si procede con una taggatura manuale degli stessi in diegetici e mimetici. Con un alto numero questa operazione prevede tecniche di text mining, per esempio con Sketch Engine (con il comando `[word="\\""]{0,}[word="\\""]`) o Python e librerie specializzate come spaCy o NLTK per eseguire il parsing del testo, identificando le virgolette. In questo studio sono stati usati Voyant Tools (per il conteggio delle parole e per le Figure dei prossimi paragrafi) e taggatura manuale. Infine, l'analisi differenziale ha mirato a evincere i rapporti tra i dati testuali estratti e differentemente taggati in diegetici e mimetici. Infatti, il testo diegetico e il testo mimetico rivestono diverse salienze all'interno della stessa opera e, parimenti, i dati testuali, ovvero gli aspetti linguistici oggetto di indagine, acquisiscono una diversa salienza per la successiva attività inferenziale caratterizzando un'indagine nella duplice veste diegetica-mimetica come maggiormente granulare, una granularità dei dati appunto utile per le inferenze successive e persa diversamente. Quali critiche è possibile muovere contro questo approccio metodologico? Una critica potrebbe essere che gli assunti teorico-metodologici, operando solamente su diegesi e mimesi, non tengano conto del discorso indiretto libero, ovvero il discorso diretto dei personaggi inserito nella diegesi del narratore. La scelta di non ricercare sul discorso indiretto è in realtà deliberata, poiché scaturisce da problematiche sull'operazionalizzazione del discorso indiretto libero già rilevate in passato (Burrows, 1987) e tutt'ora in essere (Taivalkoski-Shilov, 2019; Piper and Bagga, 2024; Argenzio, Ciotti and Corradino, 2025). Il discorso indiretto libero, linguisticamente, rientra nella categoria del discorso riportato insieme al discorso diretto, al discorso indiretto ecc., e i principali studi linguistici italiani su questo aspetto (adottato come espediente narrativo in modo crescente a partire dall'Ottocento) non ne presentano delimitazioni univoche. Diversi linguisti e lingviste (Herczeg, 1963; Cane, 1969; Mortara Garavelli, 1985, 1995; Calaresu, 2000, 2004) non ne individuano delimitatori formalmente circostanziati e utili a operazionalizzarlo (Moretti, 2013). Per Mortara Garavelli, per esempio, il discorso indiretto libero è una condizione pragmatica data dalla «mescolanza di diegesi e di mimesi» (1985, 105). Tale condizione è rilevabile soprattutto laddove si incontrino i due centri deittici all'origine della citazione - il centro deittico originario di un enunciato e il centro deittico del contesto citante - e possono essere assunti come suoi segnali di riconoscimento i fenomeni intonativi, morfosintattici, lessicali, testuali e contestuali che origina - ma «l'impressione che se ne ritrae è che nessuno

basti singolarmente a definire la specificità del fenomeno [...]: né, d'altra parte, sembra sufficiente mettere insieme i tratti rilevabili sui vari piani per avere un identikit completo - e omogeneo - di questo tipo di discorso» (1985, 120). Inoltre, il livello lessicale può concorrere a complicare un quadro già di per sé sfumato, sfocato e privo, formalmente, di indicatori univoci, siano essi delimitanti, implicanti o escludenti, poiché quanta più similarità intercorre tra la lingua della diegesi e la lingua della mimesi, tanta maggiore sarà la difficoltà nel riconoscere la mimesi all'interno del cotesto diegetico. L'assenza di segnali linguistici finiti e definiti, al momento, rende impossibile operazionalizzare il discorso indiretto libero, poiché intraducibile nel linguaggio della programmazione informatica. Con l'AGI (Artificial General Intelligence) sarà magari possibile, in futuro, contemplare tutti i parametri che si possono intersecare nella resa del discorso indiretto per la sua individuazione automatizzata.

Il corpus soddisfa gli aspetti e i fini della ricerca: (1) ambientazione e periodo storico della finzione letteraria ne caratterizzano evidenti implicazioni geopolitiche rispetto alla tematica dei migranti e al diverso da sé; (2) l'autore è, tra i pochi scrittori italiani, ad aver passato in rassegna, quasi integralmente, tutte le più recenti tematiche della migrazione straniera in Italia; (3) il successo editoriale e l'eco culturale che ne è seguita pongono quesiti proprio sul come sia stata trattata tale tematica e quali rappresentazioni ne siano state date; (4) gli oltre venticinque anni della serialità rientrano nel periodo storico della suddetta mutazione di sensibilità linguistica e permettono analisi, in prospettiva diacronica, sulle eventuali evoluzioni della figura del migrante e con quali parole e caratteristiche si è eventualmente evoluta. Per rintracciare un insieme di parole che si riferisca alla figura di chi migra si è ricorso a liste tratte da altre ricerche e pubblicazioni (Faloppa, 2011; Cherubini, 2016; Milazzo, 2018, 2022). Il risultato è un insieme di dieci parole, dalle maggiormente esclusive all'ultima ritenuta oggi maggiormente inclusiva: “negro/nìvuro”, “vocomprà”, “arabo”, “extracomunitario”, “magrebino”, “clandestino”, “profugo”, “emigrato”, “immigrato”, “migrante”.

3. Non-fiction

Andrea Camilleri vanta impegno di lungo corso sui temi politici e sociali. Alla morte la nota rivista italiana di cultura, scienza, politica e filosofia *MicroMega* ha pubblicato la collezione integrale degli scritti apparsi sulle proprie pagine con il titolo *Tutto Camilleri 1999-2018* con una prefazione del direttore in cui si legge: «Non ha mai lesinato il suo impegno, pronto a spendere e rischiare la sua enorme popolarità per cause che andavano contro lo spirito dei tempi. Con un linguaggio senza accomodamenti, all'altezza della gravità delle nefandezze che denunciava e combatteva» (Flores D'Arcais, 2019, 4). Sul fenomeno migratorio Camilleri si schiera, per esempio, firmando la prefazione del primo rapporto sui richiedenti asilo da parte del Consorzio italiano di solidarietà (Camilleri, 2005, 6). In linea con il linguaggio impiegato nel rapporto, “uomini”, “donne” e “bambini” cinque volte sono definiti “extracomunitari” e, in chiusura, li definisce anche “cittadini africani”, criticando l'azione governativa di «respingere indiscriminatamente profughi, clandestini, richiedenti asilo, immigrati» (Camilleri, 2005, 8). Nel 2009, sulla rivista di geopolitica *Limes*, l'autore si interroga sull'identità italiana, vedendo in chi migra un “extracomunitario” che subisce aggressioni xenofobe ed è mal sopportato da chi, avendo a propria volta un passato migrante, dovrebbe piuttosto solidarizzare (Camilleri, 2009). Nel 2011, dopo le primavere arabe esce un libro-intervista in cui prevede cambiamenti radicali circa le proporzioni del fenomeno migratorio (in effetti verificatesi dopo quelle proteste): «Io credo che debbono avvenire mutamenti cosmici [...]. La prima cosa è l'immigrazione. L'immigrazione che noi consideriamo fino a questo momento, fin quando è assorbibile, come una forza lavoro non indifferente per esempio per lo sviluppo dell'Italia. E la tolleriamo. Mi chiedo: come reagiremmo se di queste persone non ne avessimo bisogno?»

(Camilleri, 2011, 11). Dal 2015 inizia a soffrire di ipovisione, l'anno seguente in un messaggio video per *Le giornate di solidarietà del servizio sociale per i rifugiati* organizzate dal Consiglio nazionale ordine assistenti sociali, ribadisce l'esperienza epocale del fenomeno migratorio (Camilleri, 2016) e presta altri propri interventi sul tema all'interno del documentario *Lontano dagli occhi* (Iannacone and Cambi, 2016) per la *Prima giornata nazionale delle vittime dell'immigrazione*. Ormai del tutto cieco, nel 2018 interviene su uno spartiacque delle politiche italiane sulla migrazione, il caso Aquarios, la nave ONG con 629 naufraghi migranti a cui viene rifiutato l'attracco presso i porti italiani nel giugno di quell'anno. In un'intervista rilasciata al quotidiano Repubblica, li definisce "profughi" e si definisce incapace di differenziarli in "clandestini" e "richiedenti asilo" né di trovare più ammissibile la distinzione con i "migranti economici" (Camilleri, 2018). Infine, in un video messaggio inviato al sito fanpage.it pochi mesi prima della morte sullo sgombero di una comunità di 540 migranti vicino Roma, si dice indignato per la politica migratoria persecutoria e nazista anche contro chi è già in terraferma italiana che segue la chiusura dei porti alle ONG operanti nel canale di Sicilia (Camilleri, 2019). In questo breve *excursus* nella non-fiction risultano assenti parole denigratorie e l'allineamento allo spirito del tempo che ha visto la diffusione della parola "extracomunitario" all'inizio del nuovo millennio (Orrù, 2019) gradualmente, negli anni successivi e con la spinta del *politically correct*, essere sostituita dalla parola più neutra "migrante" nei canali di informazione.

4. Fiction

Passando alla fiction, la parola "extracomunitario" (74 occorrenze, 29 in diegesi e 45 in mimesi) fa la prima comparsa all'interno della raccolta di racconti *La paura di Montalbano* del 2002, a seguito della nascita della moneta unica europea, e in prospettiva diacronica per sostituire altre parole su cui, intanto, la sensibilità linguistica dei parlanti stava mutando. All'inizio la parola avrebbe dovuto avere valenza neutra, in realtà l'impiego ha quasi da subito iniziato ad acquisire gradazione negativa e un uso pressoché circoscritto a migranti e BIPOC, poiché spesso accompagnata da modificatori di natura negativa (per esempio "clandestino" in funzione aggettivale). L'anno successivo, nel 2003, nel primo romanzo che tratta specificamente le recenti migrazioni attraverso il Mediterraneo, *Il giro di boa*, la parola compare per 35 volte, diventa maggioritaria e ricorre all'interno del corpus, fino all'ultimo romanzo, senza alcuna distinzione di impiego tra narratore e personaggi, racconti e romanzi.

La parola "migrante" (43 occorrenze, 13 in diegesi e 30 in mimesi) inizia a comparire negli ultimi testi del corpus. La sensibilità dell'autore, infatti, sembra mutare nel 2016 con la pubblicazione del romanzo *L'altro capo del filo*, dove, come già ai tempi del romanzo *Il giro di boa*, il testo riporta l'attualità dei movimenti migratori in Europa e Italia attraverso la Sicilia. Sembrano significative le polemiche che anticipano, accompagnano e seguono la messa in onda nel 2019 dell'adattamento filmico che ne viene tratto, considerato pro-immigrazione (Lijoi, 2019). Sembra essere anche significativo il fatto che sparisce la parola "extracomunitario", in questo testo sostituita dalla sola parola "migrante" (42 volte); migranti che vengono anche definiti "poverazzi" ("poveretti", 6 volte). Un altro aspetto che appare significativo è la ricomparsa della parola "nivuro", ma non più riferito a un essere umano in carne e ossa come avveniva nei primi testi del corpus, piuttosto usata una sola volta nei confronti della statua di San Calogero, un santo africano di cui il narratore ipotizza l'antico arrivo in Sicilia con gli stessi mezzi degli odierni migranti. Questo testo indica il cambio di sensibilità linguistica verso la figura del migrante, confermato anche in un successivo romanzo del 2018, *Il metodo Catanalotti*, dove la tematica non è centrale come nel testo del 2016, ma l'impiego è utile

all'autore per aggiungere, in sottofondo, le contraddizioni europee sulle redistribuzioni delle quote e sulla costruzione di barriere, soprattutto nell'Europa centrale, tra i paesi dell'Unione.

La parola italiano “negro” e il suo corrispettivo in siciliano “nìvuro”, nelle loro declinazioni di genere e numero, sono impiegate nel corpus 41 volte (26 in diegesi e 15 in mimesi) e non mostrano sull'asse diegetico-mimetico una differente trattazione significativa, poiché vi ricorrono sia il narratore che i personaggi (principali o secondari), escludendo dunque un possibile uso atto a marcare differenti *milieu* sociolinguistici. I primi segnali recessivi appaiono nel 2003 con il romanzo *Il giro di boa*, dove la parola “nìvuro” continua a essere impiegata, ma come minoritaria, poiché i migranti iniziano a essere definiti con la nuova parola di origine burocratica diffusa dalla piena formazione economica della Comunità Europea, “extracomunitario”, la quale diventa maggioritaria (appare qui 35 volte). L'affiorare di una diversa sensibilità non sopprime comunque del tutto l'uso della parola dispregiativa “negro”; va notato, infatti, che la parola si ripresenta nel 2013 nel romanzo *Un covo di vipere* e che ancora nel 2019 ne *Il cuoco dell'Alcyon* il narratore la ripropone in funzione denigrante verso le origini e le competenze linguistiche di una collaboratrice domestica migrante.

Anche la parola “arabo” (39 occorrenze, 11 in diegesi e 28 in mimesi) mostra un percorso di simile evoluzione benché più ascrivibile ai personaggi. L'impiego della parola è spesso improprio, poiché non è riferita a sauditi ma viene impiegata per riferirsi a persone di fede islamica con determinate caratteristiche somatiche più che linguistiche (anche in Sudan si parla arabo e professa l'Islam, ma difficilmente un sudanese è chiamato “arabo” piuttosto che “nero”) e quest'uso improprio è anche trasversale come per esempio nel romanzo del 2006 *La vampa d'agosto*, dove personaggi e narratore se ne servono per riferirsi a un muratore straniero morto lavorando in nero in una ditta edile. Infine, è interessante notare che la parola sparisce per più di un decennio per riapparire solo più tardi, nel 2017 nel romanzo *La rete di protezione*, con una diversa sensibilità: infatti, viene usata per parlare di un matematico alla corte di Federico II e dunque in riferimento ai veri arabi che nei secoli precedenti hanno vissuto in Sicilia.

La parola “clandestino” conta 34 occorrenze (13 in diegesi e 21 in mimesi), metà delle quali (17) si trovano nel romanzo *Il giro di Boa*, che appunto con l'associazione alla parola lì maggiormente usata, “extracomunitario”, conferma una sensibilità verso i migranti meno sviluppata rispetto al romanzo del 2016 *L'altro capo del filo*. La presenza della parola “clandestino” è inoltre costante in tutto il corpus, comparando nel primo romanzo del 1994 e ricorrendo poi nel 1996, 1999, 2002, 2003, 2004, 2005, 2006, 2008, 2009, 2016 e 2020.

La parola “magrebino” (12 occorrenze, 4 in diegesi e 8 in mimesi) viene impiegata soprattutto nel romanzo *L'età del dubbio* del 2008 per non specificare la nazionalità di Ahmed Chaikri, un infiltrato della polizia all'interno di una organizzazione criminale e ricompare solo un'altra volta in tutto il corpus all'interno del romanzo *Un covo di vipere* per riferirsi più a un indeterminato lignaggio di lontane origini magrebine di una persona. La dicotomia “immigrato/emigrato” compare 10 volte (4 nella diegesi e 6 nella mimesi) e può essere fatta risalire già alla raccolta di racconti *La paura di Montalbano*. Chiudono il conteggio le parole a basso peso statistico e basso potenziale interpretativo “profugo” (compare 4 volte, 2 in diegesi e 2 in mimesi) e “vocomprà” (4 occorrenze, 1 in diegesi e 3 in mimesi).

5. Risultati

La comparazione fiction e non-fiction ravvisa un andamento simile con la cruciale differenza per quel che riguarda il linguaggio esclusivo che è invece ricorrente nella fiction almeno fino al romanzo *L'altro capo del filo* del 2016 come viene sintetizzato nella Figura 1.

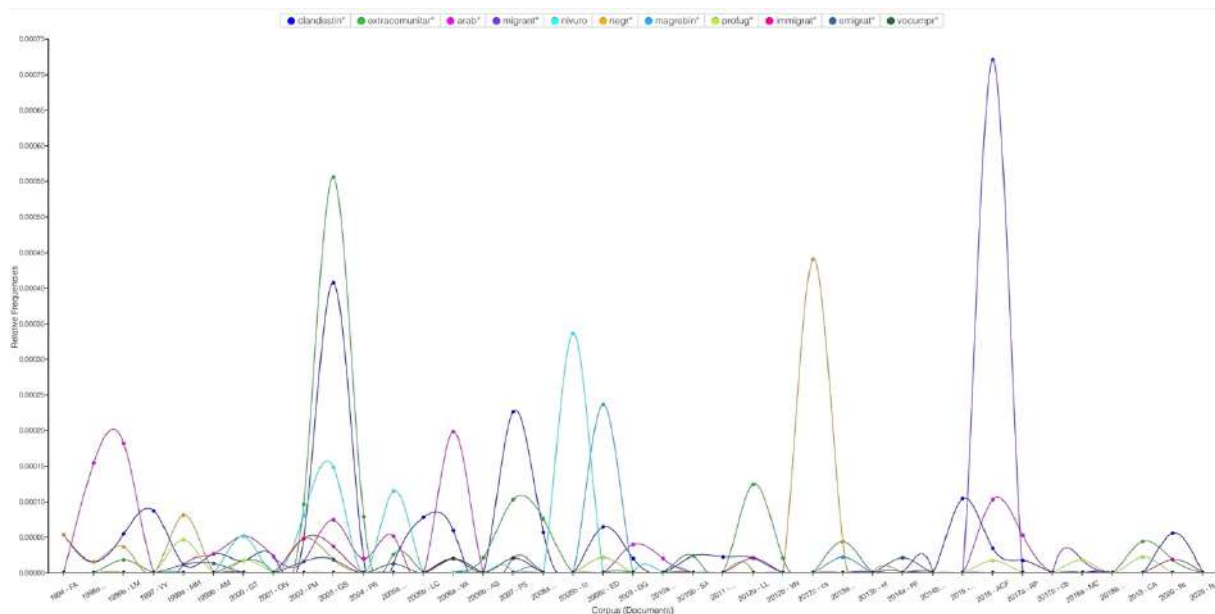


Figura 1. Impiego diacronico del linguaggio inclusivo/esclusivo

Nella Figura 1 notiamo due picchi. Il primo picco, da sinistra a destra, si riferisce al romanzo *Il giro di boa* del 2003 e il secondo picco appunto a *L'altro capo del filo* (rispettivamente abbreviati in “GB” e “ACF” nelle Figure e nei materiali supplementari). Va ricordato che all’interno del corpus i due testi che hanno la migrazione quale tematica principale sono appunto i due romanzi *Il giro di boa* e *L'altro capo del filo*. La parola più inclusiva “migrante” appare solo a partire da quest’ultimo romanzo, quindi nel 2016, ed è preceduta da tutte le altre parole maggiormente esclusive. Le dieci parole indagate nel corpus contano complessivamente 261 occorrenze e l’aspetto sottolineato per i principali due picchi della Figura 1 risulta più chiaro nella Figura 2, dove viene riportato l’andamento diacronico delle quattro parole che possono segnare un graduale passaggio da un linguaggio più esclusivo a uno maggiormente inclusivo: “negro/nivuro”, “arabo”, “extracomunitario”, “migrante”.

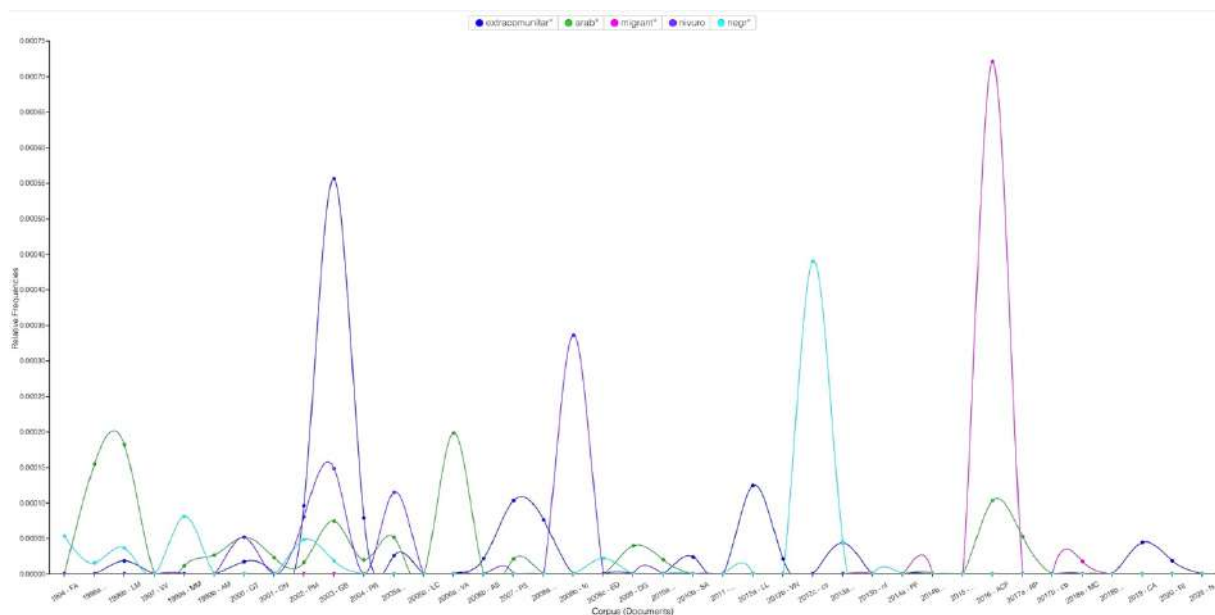


Figura 2. Diacronia delle parole “negro/nivuro”, “arabo”, “extracomunitario”, “migrante”

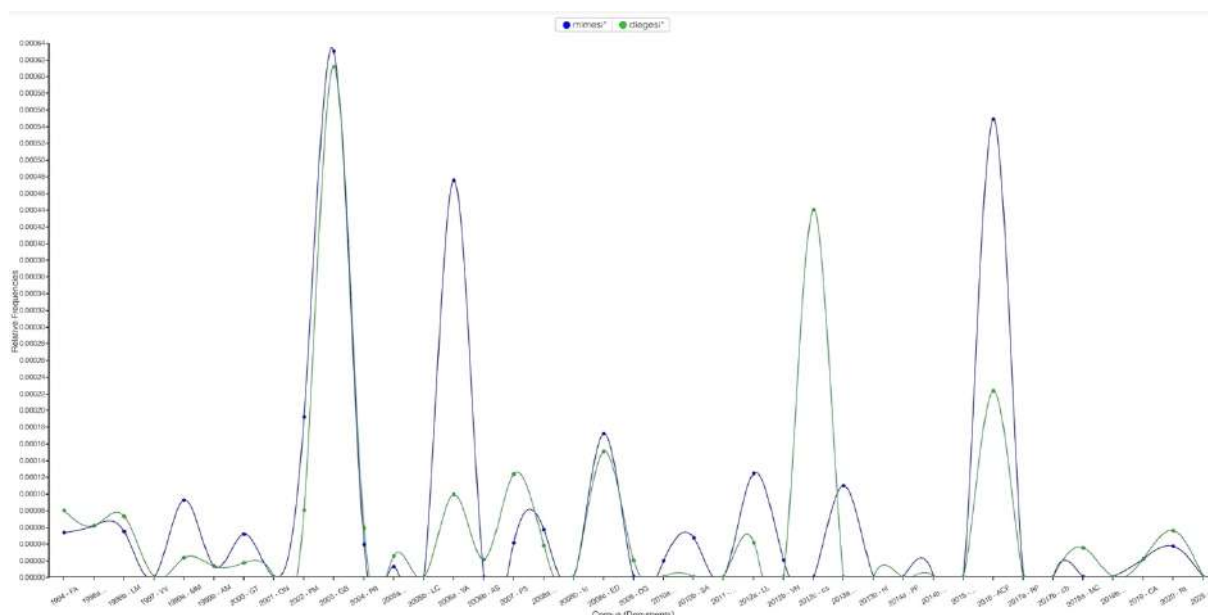


Figura 3. Differenziale dell'impiego diacronico del linguaggio inclusivo/esclusivo

La Figura 3 mostra l'analisi differenziale delle 261 occorrenze registrate per le dieci parole indagate, riportandone il differenziale d'impiego diegetico/mimetico. È interessante notare per i principali due picchi che, di nuovo da sinistra a destra, il primo picco del romanzo *Il giro di boa* del 2003 attesta il verificarsi degli assunti teorici della trasversalità e della coincidenza nell'impiego del linguaggio esclusivo. Le parole maggiormente qui presenti, “extracomunitario” e “clandestino” (come si è visto nella Figura 1), la compresenza e il fatto che alla trattazione della tematica non sia associato un correlativo formale, ovvero l'impiego di parole meno esclusive né una differenziazione d'impiego tra la voce narrante e i suoi personaggi suggeriscono una scarsa consapevolezza autoriale, influenzata soprattutto dal linguaggio post-comunitario dei media. Nel secondo picco, quello che si riferisce al romanzo *L'altro capo del filo* del 2016, mentre viene meno il dato della trasversalità affiora invece l'impiego della parola inclusiva “migrante” e sembra possibile affermare che, nel 2016, la trattazione del tema della migrazione sia più matura e soprattutto più (linguisticamente) consapevole rispetto al 2003.



Figura 4. Word cloud delle parole in diegesi (a destra) e in mimesi (a sinistra)

La Figura 4, infine, raggruppa in nuvole le parole e i loro impieghi differenziali mostrando, in proporzione, la trasversalità e coincidenza delle stesse forme. Il risultato finale dell'analisi differenziale dà 107 occorrenze in diegesi, il 41%, e 154 occorrenze in mimesi, il 59%. Il maggior impiego in mimesi è in linea con la tipologia di linguaggio indagata che, appunto, ci si aspetta maggiormente impiegato nell'oralità. Il 41% in diegesi non è comunque una

percentuale bassa, al contrario conferma un impiego diegetico proporzionato all'impiego mimetico non schiacciante.

6. Conclusioni

La ricerca indaga l'*agency* autoriale rispetto al linguaggio esclusivo attraverso assunti teorici e tecniche computazionali mirando a ricostruire una base empirica per l'analisi delle scelte linguistiche nei testi narrativi che superi approcci puramente interpretativi e soggettivi. I risultati evidenziano la rilevanza della dimensione diacronica nell'evoluzione del lessico relativo alla figura del migrante, riflettendo il mutamento della sensibilità linguistica e culturale nel contesto italiano tra gli anni Novanta e il secondo decennio del Duemila. Parole inizialmente diffuse e socialmente accettate come "extracomunitario" mostrano un progressivo slittamento semantico verso connotazioni negative, mentre emergono forme più neutre come "migrante", in linea con l'affermarsi del *politically correct*. Un elemento significativo riguarda la distribuzione delle parole lungo l'asse diegetico-mimetico. La presenza consistente di linguaggio esclusivo anche nella diegesi, e non soltanto nella mimesi, costituisce un indicatore cruciale per l'inferenza dell'*agency* autoriale quando le stesse forme compaiono sia nella voce narrante sia nelle voci dei personaggi (trasversalità e coincidenza), escludendo l'ipotesi che tali usi siano funzionali alla caratterizzazione sociolinguistica dei personaggi. Il linguaggio sembra riflettere una più ampia adesione a modelli discorsivi diffusi nel contesto culturale di riferimento. La comparazione con la produzione non-fiction aiuta a raffinare l'interpretazione, infatti l'assenza di parole denigratorie negli scritti saggistici e giornalistici, unita alla coerenza etico-politica autoriale rispetto ai temi della migrazione, suggerisce che l'uso di linguaggio esclusivo nella fiction non sia stato consapevole fin dall'inizio, esso piuttosto appare come una progressiva rielaborazione critica di modelli linguistici. In questo senso, la dimensione diacronica consente di cogliere un'evoluzione della sensibilità autoriale che, infine, si orienta verso scelte linguistiche inclusive. Permangono criticità legate all'attuale impossibilità di operationalizzare il discorso indiretto libero, uno spazio testuale che potrebbe incidere sulla distribuzione dei dati. In conclusione, la metodologia appare promettente per lo studio del linguaggio letterario esclusivo, capace di integrare analisi quantitativa e interpretazione critica e di contribuire alla comprensione delle dinamiche tra linguaggio, ideologia e rappresentazione.

Materiale supplementare

La bibliografia del corpus è disponibile all'indirizzo <https://github.com/PietroMazzarisi/Le-indagini-di-Montalbano>.

Bibliografia

- Argenzio A., Ciotti B. and Corradino A. (2025). Usare i Large Language Model per l'analisi del testo narrativo: strategie di prompt engineering per il riconoscimento del discorso indiretto libero nella narrativa italiana 1830-1930. In Rebora S., Rospocher M. and Bazzaco S., curatori, *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, 11-13 giugno, a cura di, Università di Verona, Atti del XIV Convegno Annuale*, pp. 349-356.
- Aristotele. (1964). *Poetica*. Valgimigli M., curatore, Laterza.
- Bernstein R. (1990). The Rising Hegemony of the Politically Correct. *The New York Times* (28 October 1990).
- Burrows J. F. (1987). Word-Patterns and Story-Shapes: the Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2 (2): 61-70.

- Calaresu E. (2000). *Il discorso riportato. Una prospettiva testuale*. Il Fiorino.
- Calaresu E. (2004). *Testuali parole. La dimensione pragmatica e testuale del discorso riportato*. Franco Angeli.
- Camilleri A. (2005). Prefazione. In Olivieri M. S., curatore, *La protezione negata. Primo rapporto sul diritto di asilo in Italia*: pp. 5-8.
- Camilleri A. (2009). Cos'è un italiano? Appunti per una definizione. *Limes*, 24 Febbraio.
- Camilleri A. (2011). *Questo mondo un po' sgualcito. Conversazioni con Francesco De Filippo*. Infinito edizioni.
- Camilleri A. (2016). Andrea Camilleri con il CNOAS per le Giornate di solidarietà del servizio sociale per i rifugiati. *CNOAS*, 04 giugno.
- Camilleri A. (2018). Come italiano sento di aver fallito. Intervista di Simonetta Fiori. *La Repubblica*, 07 luglio.
- Camilleri A. (2019). Tenere i porti aperti, non sarò complice di questa volgarità nazista. *Fanpage.it*, 24 gennaio.
- Cane, E. (1969). *Il discorso indiretto libero nella narrativa italiana del Novecento*. Silva.
- Cherubini M. (2016). *Glossario sull'asilo e la migrazione*. Commissione europea.
- Falloppe F. (2011). *Razzisti a parole (per tacer dei fatti)*. Laterza.
- Flores D'Arcais P. (2019). Andrea. Un amico, un compagno, un grande scrittore. *MicroMega*. Tutto Camilleri 1999-2018, 20 July: 3-4.
- Genette G. (1969). *Figures II*. Seuil.
- Herczeg G. (1963). *Lo stile indiretto libero in italiano*. Sansoni.
- Iannaccone D. e Cambi L. (2016). *Lontano dagli occhi*. RAI 3.
- Lijoi L. (2019). Naufragio senza telespettatore. Qualche riflessione su *L'altro capo del filo* di Andrea Camilleri. *Trasparenze* 4 (5): 183-200.
- Mazzarisi, P. (2025). *Teoria e framework DM^A per il Distant Reading. Radici classiche e studi su corpora italiani*. EUT Edizioni Università di Trieste.
- Milazzo G. (2018). *Notizie di chiusura. Sesto rapporto Carta di Roma 2018*. Associazione Carta di Roma.
- Milazzo G. (2022). *X rapporto Carta di Roma 2013/2022. Notizie dal fronte*. Associazione Carta di Roma.
- Moretti F. (2013). 'Operationalizing': Or, the Function of Measurement in Modern Literary Theory. *Pamphlets of the Stanford Literary Lab*, 6 (December).
- Mortara Garavelli B. (1985). *La parola d'altri*. Sellerio.
- Mortara Garavelli B. (1995). Il discorso riportato. In Renzi L., Salvi G. and Cardinaletti A., curatori, *Grande grammatica italiana di consultazione*, vol. 3, pp. 427-468.
- Orrù P. (2019). Riflessioni sul lessico delle migrazioni degli ultimi anni. *Nuova antologia*, 620(2291): 324-33.
- Piper A. and Bagga S. (2024). Using Large Language Models for Understanding Narrative Discourse. In Kumar Lal Y. et al., curatori, *Proceedings of the The 6th Workshop on Narrative Understanding*: pp. 37-46.
- Platone (1990). *La Repubblica*. Lozza G., curatore, Mondadori.
- Taivalkoski-Shilov K. (2019). Free Indirect Discourse: An Insurmountable Challenge for Literary MT Systems? In Hadley J., et al., curator, *Proceedings of the Qualities of Literary Machine Translation*: pp. 35-39.

Classification of Andrea Camilleri's works using bottom-up and top-down approaches

Andrea Sciandra¹, Michele A. Cortelazzo²

¹University of Padova – andrea.sciandra@unipd.it

²University of Padova – cortmic@unipd.it

Abstract

This study aims to classify Andrea Camilleri's works into groups based on their stylistic similarities, employing various approaches. The available corpus consists of 70 works published by Sellerio spanning approximately 40 years (1978–2024). First, an exploratory approach was taken, using unsupervised methods (bottom-up classification), such as cluster analysis and multidimensional scaling. This revealed some notable trends, particularly in a collection of short stories and some standalone works. For example, the standalone novel 'La rizzagliata' bears similarities to the Montalbano novels, probably because it mentions the Inspector, despite not being part of the series. Thus, while the reasons for the anomalies found in the exploratory analyses are clear for some works, others require further investigation. To extrapolate additional clues from these results, we performed a supervised classification using information from the special issue of Quaderni Camilleriani published in 2021. This publication identified three main categories: 'Inspector Montalbano'; 'Historical and Civic Works'; and 'Fiction of Various Kinds'. We added an additional residual class to this categorisation, consisting of plays, dialogues and letters — writings that were not included in the three main categories of Quaderni Camilleriani. Supervised (top-down) classification using a Support Vector Machine with a Radial Basis Kernel Function produced highly accurate results, particularly through the most frequent words and above all through the axes of a Correspondence Analysis. Exclusive use of functional or 'empty' words does not achieve the best result, as might be expected given the presence of several neologisms (idiolect) typical of Camilleri. Overall, the results of the analysis demonstrate the potential of distant reading and classification methods, while leaving ample room for qualitative analysis of the writing style of this prolific and complex author.

Keywords: Andrea Camilleri, Exploratory Methods, Supervised Classification, Qualitative Analysis.

1. Introduction

2025 marked the centenary of Andrea Camilleri's birth. He was an important and prolific author, best known for his Inspector Montalbano series of novels. With over 120 works translated into more than 30 languages spanning a period from 1978 to 2024, his output is difficult to categorise. This prolific opus reflects a gradual broadening of Camilleri's range of interests, which makes it somewhat difficult to classify his works (Demontis, 2021). Although this expansion of themes coincided with growing critical interest, the literature on Camilleri's production does not appear to have specifically sought to group his works into homogeneous categories. Indeed, at an academic level, many analyses of Camilleri's works focused primarily on his distinctive language ('Camillerese' or 'Vigatese'), as highlighted by Russi (2020). Instead, Douthwaite (2007) explored how Camilleri's speech and thought presentation can be

used to analyse the nature of a text and identify its genre. He distinguished between two main sub-genres: a lowbrow form, primarily focused on entertainment and subtly promoting conservative ideological values, and a highbrow form, which is less common and aims to critique society, challenge traditional detective fiction conventions, and introduce innovation into the novel as a genre. Demontis (2021), on the other hand, provided an analytical description of the thematic bibliography of Andrea Camilleri's works, which allows us to identify the themes that have primarily attracted the interest of critics and which suggest an initial hypothesis for classification.

This research uses stylometric methods to analyse Camilleri's works, trying to identify unexpected elements and confirming or refuting existing attempts at classification. This type of quantitative analysis can be divided into two approaches: bottom-up, which uses unsupervised methods of an exploratory nature, and top-down, which uses statistical modelling to confirm the quality of the subdivision of works that has emerged from the literature.

The available corpus comprises 70 works published by Sellerio Editore (Tuzzi, 2026) and covers the entirety of Camilleri's output, totalling over three million words (3,315,184) and featuring almost 100,000 unique words (96,758).

The paper is structured as follows: Section 2 presents the explorative analyses, while Section 3 focuses on the supervised classification. Section 4 contains a discussion of the results and conclusions.

2. Explorative analysis

According to the approach described by Eder et al. (2016), exploratory analyses within the bottom-up approach include various techniques, such as cluster analysis (CLA) and multidimensional scaling (MDS). The analysis setup using the R `stylo` package involves importing .txt files in UTF-8 encoding and indicating the Italian language. This could have an impact on the recognition of terms belonging to certain grammatical categories, such as pronouns. Our analysis focuses on single words and initially uses a limited set of the most frequent words (MFWs), totalling 1000. The exploratory analyses (CLA and MDS) are based on calculating intertextual distances within this set of words, particularly using the Delta distance (Burrows, 2002). This distance has the advantage of providing a standardised measure using the mean and standard deviation of the occurrence of each word in the available texts. Given the large number of works, to simplify the results and graphs reading, the novels in the Montalbano cycle are indicated with the initial letter 'M', while the standalone works are indicated with 'sa', the short stories with 'racc', the trilogy of metamorphoses with 'tri', the dialogues with 'dia', the letters with 'let', the monologues with 'mono', and the plays with 'tea'.

The CLA (Fig. 1) obtained with these settings shows consistent recurrences alongside results that cannot be easily explained. Initially identifying either two or four main clusters, the works in the Montalbano cycle are placed in two large, disconnected groups in the dendrogram (top and down, in blue), raising doubts about the effectiveness of this classification. Conversely, the increasing presence (Russi, 2020) of neologisms (idiolect) in the Inspector's works could make possible this kind of clustering, given the peculiarity and frequency of the terms used.

On the other hand, grouping the 'Trilogy of Metamorphoses', most of the short stories, dialogues, monologues, letters and plays is consistent. Of particular note are the placements of

a standalone work (*La rizzagliata*, 2009¹ – sa_09c, marked with a red arrow in Fig. 1) and a volume of short stories (*Il giudice Surra e altre indagini in Sicilia*, 2023 – racc_23a, marked with a red arrow in Fig. 1), which will be revisited following further analysis.

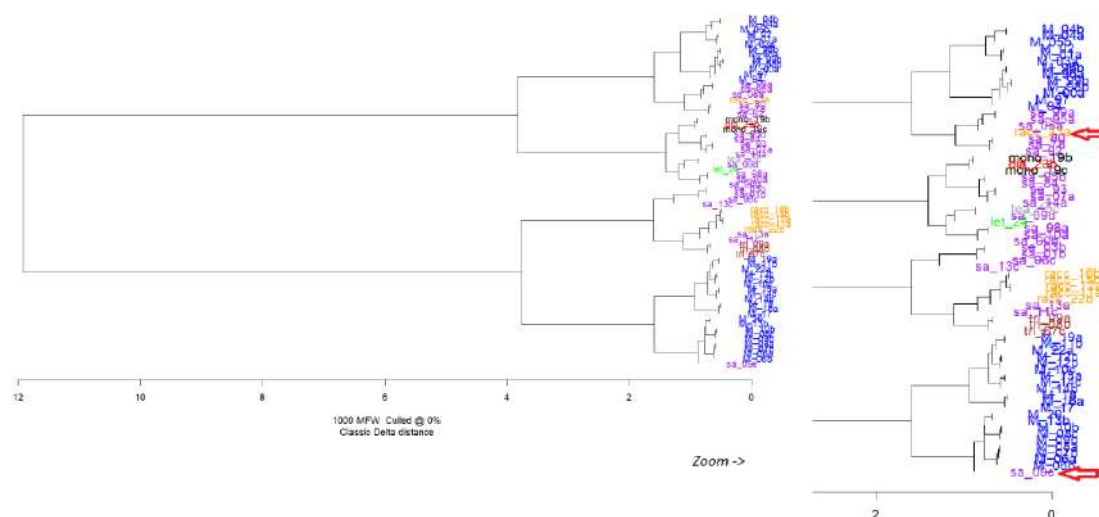


Fig. 1. Cluster Analysis based on 1000 MFWs of Camilleri's works (zoom on the right).

Multidimensional scaling, applied once again through the Delta distance of texts on the basis of the 1,000 most frequent words, yields even clearer results. The lower part of the graph (Fig. 2), following the second dimension in the Cartesian plane, contains all the works in the Montalbano cycle (together with *La rizzagliata* – sa_09c), while the top left contains monologues, dialogues, plays, and letters. The 'Metamorphoses Trilogy' and all the short stories except *Il giudice Surra* (racc_23a) are located at the top right. Interestingly, some standalone works are located close to the short stories: *La rivoluzione della luna* (2013 – sa_13a), *La presa di Macallè* (2003 – sa_03b) and *La setta degli angeli* (2011b – sa_11c).

In relation to the clustering results, which divided the works in the Montalbano cycle into two groups, it should be noted that MDS confirms this trend to some extent. While all the Montalbano works are located in the lower part of the Cartesian plane (Fig. 2), a break appears to exist (at the x-axis value of 0) between Montalbano novels published up to 2004 and those published afterwards. This could indicate Camilleri's increasing use of idiolect over time. MDS therefore confirms some of the trends that emerged in the CLA, while also pointing to other possible peculiarities in this type of unsupervised analysis.

¹ The year of publication indicated in the bibliographical references may not correspond to the copyright year held by Sellerio Editore. Given the type of analysis conducted, we chose to indicate the year of the work's first publication as the definitive year for its temporal classification.

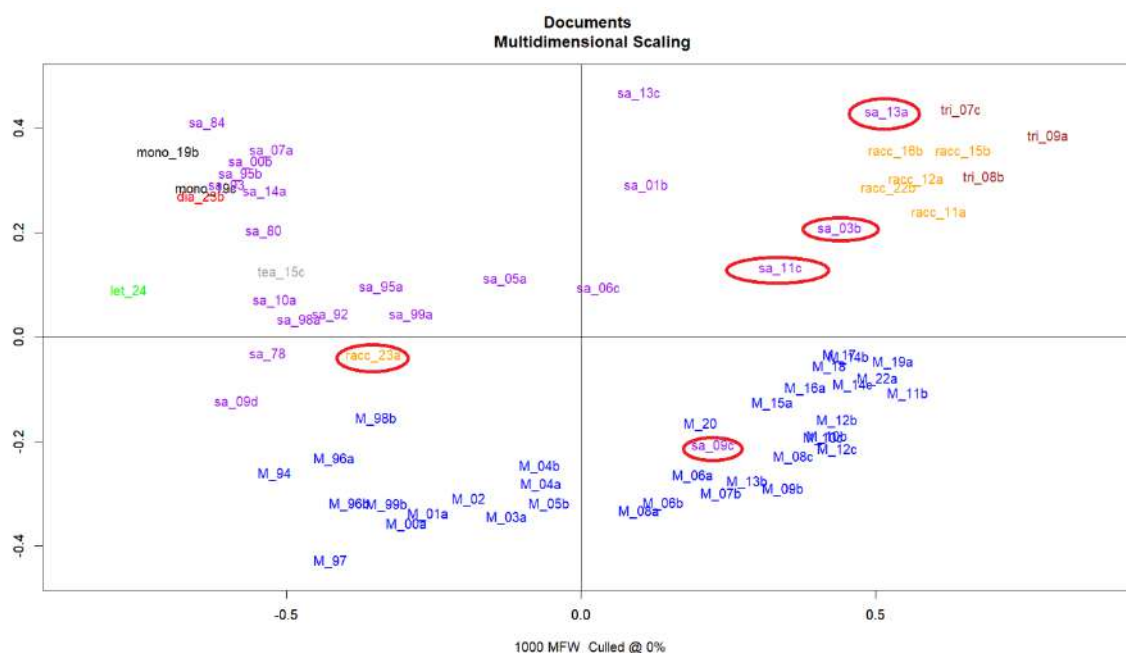


Fig. 2. Multidimensional Scaling based on 1000 MFWs of Camilleri's works.

Some anomalous results can be explained quite easily. For instance, the standalone novel *La rizzagliata* (2009) bears similarities to the Montalbano novels, probably because it mentions the Inspector despite not being part of the series. In the MDS, *La presa di Macallè* (2003 – sa_03b) is closed to works such as *La guerra privata di Samuele e altre storie di Vigàta* (2022 - racc_22b) and *Gran Circo Taddei e altre storie di Vigàta* (2011a - racc_11a). Examining these works for commonalities, it is evident that *La presa di Macallè* is set during the period of Fascism, a theme also explicitly referenced in the other two short story collections.

Il giudice Surra (2023 – racc_23a), on the other hand, represents a special case. Although it was published in 2023, it contains stories written at different times (2005, 2011), which are quite distant from the other collections of stories. However, this temporal difference alone does not fully explain the similarities with other works. Even for *La rivoluzione della luna* (2013 – sa_13a) and *La setta degli angeli* (2011b – sa_11c), no elements have emerged that would justify their location in the MDS.

3. Supervised classification

To extrapolate additional clues from these results, we performed a supervised classification using information from the special issue of *Quaderni Camilleriani* published in 2021 (Demontis, 2021). This publication is a bibliographic collection with a taxonomic structure that contains references to the analysis of individual works. As Demontis (2021) highlights, establishing consistent scientific criteria in a subjective field such as literature is challenging. We extracted from Demontis' publication the following three main categories: 'Inspector Montalbano' (34 works); 'Historical and Civic Works' (24 works); and 'Fiction of Various Kinds' (7 works). We added an additional residual class to this categorisation, consisting of plays, dialogues and letters (5 works)—writings of our corpus that were not included in the three main categories of *Quaderni Camilleriani* special issue. Therefore, the starting point was

to identify the themes that attracted the most interest from critics and suggest an initial classification. These macro-topics are also divided into more specific categories (Demontis, 2021), which include titles dealing with specific aspects. Notably, the category 'Historical and Civic Works' encompasses historical works situated from the 17th and 18th centuries, the post-*Risorgimento* era, and the early 20th century, as well as fictionalised and non-fictionalised essays. The 'Fiction of Various Kinds' category contains fantasy, art- and entertainment-related fiction, along with novels and short stories of various genres. Another category that was not considered was 'Language', as most of the works were already included in previous categories.

In order to classify Camilleri's works using a supervised (top-down) approach, we divided the texts into 1,000-word chunks (3,262 in total). The training set for the model included a random sample of 80% of the texts (2,610), while the remaining chunks (652) were used for out-of-sample verification. First, we selected a subset of 2,284 types from the 96,000-word vocabulary that appeared at least 100 times in the created chunks.

Among the various machine learning models, we selected the Support Vector Machine with a Radial Basis Kernel Function (SVM-RBKF) to maximise separation between classes and capture nonlinear relationships between features and classes (Claesen et al., 2014). This model is particularly effective at capturing subtle differences between classes, as it allows complex relationships between word frequencies to be modelled while maintaining good generalisation capabilities. We applied a 5-fold cross validation to optimise the hyper-parameters of the model. In this machine learning approach, we also decided to use and compare the classification accuracies of different methods of text vectorisation:

- The most frequent words (MFWs).
- The exclusive use of function words.
- The dimensions derived from multidimensional scaling (MDS).
- The dimensions derived from correspondence analysis (CA).

For MDS and CA, it is important to note that the out-of-sample texts used in the test set do not contribute to determining the axes derived from reducing the dimensionality of the document-term matrix (DTM). In fact, to obtain the coordinates of the test set texts, they were projected into the reduced dimensionality system produced by the training set. This ensures that the final classification accuracy is based only on data that was not observed during training. As dimensionality reduction methods were employed in these two cases, different numbers of derived representations were tested in the classification models (e.g. 50, 100, 200 dimensions). Table 1 shows the best result for each text representation method and specifies the number of dimensions used. F1 macro is computed as the unweighted average of F1-scores of all the categories, to give equal weight to each class, regardless of the number of chunks in them.

Results obtained through selecting the MFWs show that the chunks are classified correctly in 92.64% of cases (F1-macro: 0.78067). Focusing the error analysis only on the texts that showed peculiarities in the exploratory analyses revealed that *La rizzagliata* ('Fiction of Various Kinds') was the only text misclassified, being confused again with the 'Montalbano' category. In fact, the 'Fiction of Various Kinds' category achieves the worst result in terms of F1-score (F1=0.33333), while all the other categories exceed an F1 value of 0.9.

Exclusive use of function (stop)words does not achieve the best result (accuracy: 82.98%, F1-macro: 0.67223), as might be expected given the presence of several neologisms (idiolect) typical of Camilleri. For this reason, it would be useful to compile a list of Sicilian function words (Genova and Sciandra, 2026) that can also be used as predictors of the type of work.

Using the MDS axes (200 dimensions), we achieved an accuracy rate of 92.94% (F1-macro: 0,83486), with an improved F1 score for the ‘Fiction of Various Kinds’ category (F1 = 0.54545). In addition to *La rizzagliata*, among the misclassified chunks we also find an excerpt from *Il giudice Surra*, which is mistakenly attributed to the ‘Montalbano’ category — perhaps because it is also a crime novel.

The 50-dimensional axes of the correspondence analysis produced the best results, achieving an accuracy of 98.16% and significantly improving the F1 score for the ‘Fiction of Various Kinds’ category (F1 = 0.85714).

Table 1. Best results of different features for classifying Camilleri’s works in the test set (model: SVM-RBKF).

Features (n)	Accuracy (%)	F1-macro	Misclassifications (chunks)
MFWs (2284)	92.64	0,78067	48
Function words (295)	82.98	0,6722295	111
MDS (200 dimensions)	92.94	0,83486	46
CA (50 dimensions)	98.16	0,9534525	12

Interestingly, the incorrect classifications of the model with CA features do not concern the works that emerged in the exploratory analyses, but rather the novel *Il corso delle cose* (1978), which was mistakenly categorised as ‘Historical and Civic Works’ in six out of a total of thirty-two chunks in the test set. However, despite being set in the post-war period, this novel does not seem to correspond chronologically to the other works included in the ‘Historical and Civic Works’ category.

Analysing the words (Fig. 3) that contribute most to the main axes of the correspondence analysis (which also correspond to the most important variables in the classification model) reveals the importance of Sicilian dialect terms, which are particularly relevant to the first dimension. Examples include ‘*na*’, ‘*aviva*’ and ‘*fici*’. These dialect words are most likely useful markers for distinguishing the works in the Montalbano cycle.

When analysing the most important variables for classification in the MFWs model, which is the easiest to interpret, we find a combination of function and content words (see Fig. 4). In particular, the presence or absence (depending on the different classes) of terms such as ‘*Commissario*’ and ‘*Montalbano*’ helps to distinguish the relative class easily, together with some dialectal terms (‘*aviva*’, among the twenty most important words) that were already highlighted in the previous analysis of the terms with the greatest contribution to the first dimension of the correspondence analysis. Also for this reason—avoiding content words—it should be improved the list of function words with Sicilian dialect ‘stopwords’ in order to better generalise the set of predicting features.

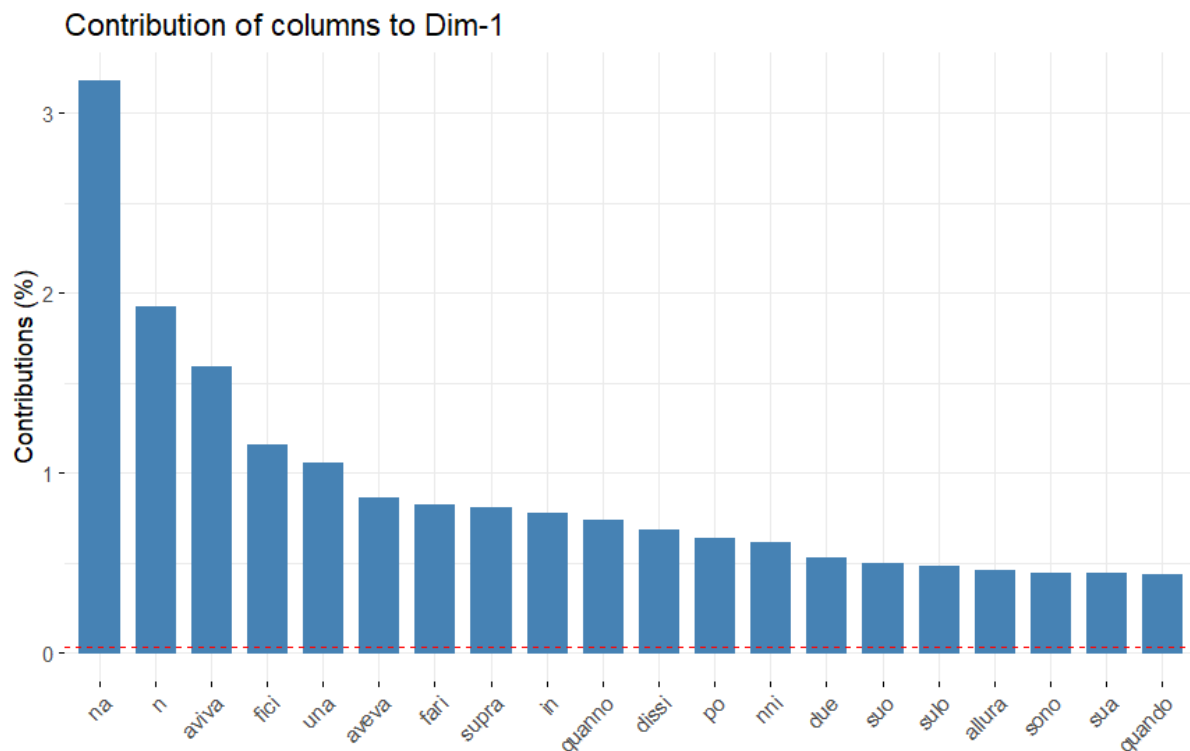


Fig. 3. Most contributing words for the first dimension of the correspondence analysis.

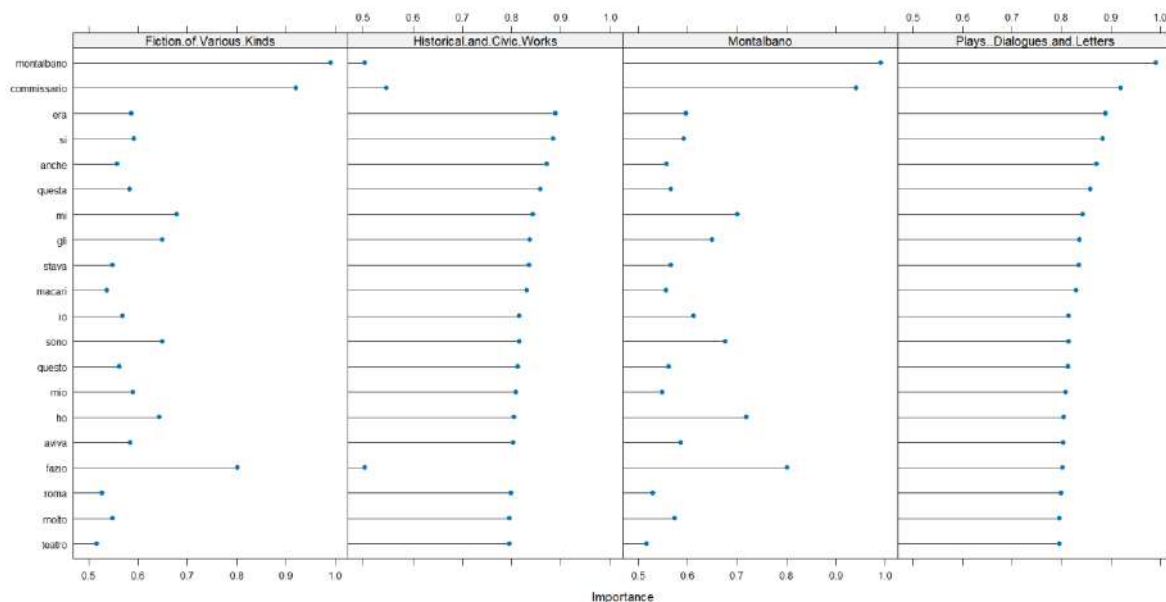


Fig. 4. Most important variables for each class – SVM-RBKF model with 50 dimensions of correspondence analysis.

4. Discussion and conclusions

Some interesting trends emerged from exploratory (bottom-up) analyses, such as the potential presence of a break in the Montalbano cycle, which may be linked to an increased use of dialectal terms. These terms are important for two reasons. First, they help to determine the main axes of correspondence analysis. Second, they emerge as important predictors for classification through machine learning. Therefore, an in-depth analysis of these terms would be an important starting point for classifying Camilleri's works.

Overall, supervised (top-down) classification showed highly accurate results, starting from the most frequent words and above all through the axes of a Correspondence Analysis. These results confirm the overall consistency of the classification displayed in the *Quaderni Camilleriani*, which proved to be a valuable resource for researchers in this field.

Using a statistical model for classification reduced or eliminated several ambiguities that emerged in the exploratory analyses, but also highlighted some special cases, such as *Il corso delle cose* (1978). Generally speaking, the fact remains that a category such as 'Fiction of Various Kinds' can be rather vague and heterogeneous for the purposes of systematic classification. Furthermore, Camilleri's distinctive style (idiolect) is very prominent in some of the works in this category, such as the 'Metamorphoses Trilogy', making it more difficult to distinguish them from other works, such as those in the Montalbano series.

However, some insights are closely linked to a linguistic analysis through close reading and can reveal additional elements that statistical models do not consider or cannot explain, as in the case of *Il corso delle cose* (1978). Therefore, the results of the analysis demonstrate the potential of distant reading (Moretti, 2013) and classification methods, while leaving ample room for qualitative analysis of the writing style of this prolific and complex author.

Acknowledgement

The authors are grateful to Sellerio Publisher for granting permission to access and use the digitized archives of Camilleri's works for this study.

References

- Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3), 267-287.
- Camilleri, A. (1978). *Il corso delle cose*. Sellerio Editore srl.
- Camilleri, A. (2003). *La presa di Macallè*. Sellerio Editore srl.
- Camilleri, A. (2009). *La rizzagliata*. Sellerio Editore srl.
- Camilleri, A. (2011a). *Gran Circo Taddei: e altre storie di Vigàta*. Sellerio Editore srl.
- Camilleri, A. (2011b). *La setta degli angeli*. Sellerio Editore srl.
- Camilleri, A. (2013). *La rivoluzione della luna*. Sellerio Editore srl.
- Camilleri, A. (2022). *La guerra privata di Samuele: e altre storie di Vigàta*. Sellerio Editore srl.
- Camilleri, A. (2023). *Il giudice Surra e altre indagini in Sicilia*. Sellerio Editore srl.

- Claesen, M., De Smet, F., Suykens, J. A., and De Moor, B. (2014). Fast prediction with SVM models containing RBF kernels. *arXiv preprint arXiv:1403.0736*
- Demontis, S. (2021). Bibliografia tematica sull'opera di Andrea Camilleri. *Quaderni camilleriani /Volume speciale, Cagliari, Ghiani*.
- Douthwaite, J. (2007). Using Speech and Thought Presentation to Validate Hypotheses Regarding the Nature of the Crime Novels of Andrea Camilleri. In *Stylistics* (pp. 143-167). Brill.
- Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*. Vol. 8(1): 107-121.
- Genova, V. G., and Sciandra, A. (2026). Camilleri's Sicilian: From Stopwords To Topic Modelling. *JADT2026 proceedings*.
- Moretti, F. (2013). *Distant reading*. Verso Books.
- Russi, C. (2020). *Sicilian Elements in Andrea Camilleri's Narrative Language: A Linguistic Analysis*. Bloomsbury Publishing PLC.
- Tuzzi, A. (2026). A New Investigation for Sellerio. *JADT2026 proceedings*.

Distinctive Lexical Traits in Andrea Camilleri: A Keyness Measures Analysis of a Corpus of Sellerio Crime Novels

Michele A. Cortelazzo¹, Matilde Trevisani², Arjuna Tuzzi²

¹Accademia della Crusca – cortmic@unipd.it

²University of Trieste – matilde.trevisani@deams.units.it

³University of Padova – arjuna.tuzzi@unipd.it

Abstract

The umbrella term *keyness* is widely used in text analysis to denote a range of quantitative measures aimed at identifying distinctive words of a text or a class of texts. Over time, numerous approaches to *keyness* have been developed, reflecting different theoretical perspectives and methodological needs. Drawing on a corpus of 78 Italian crime novels written by 16 authors and published by Sellerio, this study investigates the concept of *keyness* with a twofold objective. First, it assesses the strengths and limitations of different *keyness* measures by examining their consistency, degree of overlap, and practical applicability in large-corpus analysis. Second, it uses these measures to identify and characterize the lexical profile that distinguishes Andrea Camilleri within the broader landscape of Sellerio crime fiction. Starting from a set of words that consistently and robustly emerge across *keyness* measures as distinctive of Camilleri, the analysis explores their role from a qualitative perspective. Particular attention is devoted to the interaction between standard Italian, the Sicilian dialect, and Camilleri-style language in order to highlight his distinctive linguistic traits.

Keywords: *keyness* measures; distinctiveness; Camilleri's lexicon

1. Introduction

The umbrella term *keyness analysis* is widely used in corpus linguistics to denote the identification of distinctive words—so-called "keywords"—of a text or collection of texts. A commonly cited procedural definition says that a *keyword* is "a word which occurs with unusual frequency in a given text, by comparison with a reference corpus of some kind" (Scott, 1997). While this definition suggests an approach to extract keywords, it says little about their nature or the functions they serve. A concise synthesis of the various definitions found in the literature is that keywords function as pointers, markers or—metaphorically—access-points to the style, terminology, contents ("aboutness"), and discourse of a text or a corpus (Moreno-Ortiz, 2024). In any case, the Scott's methodological definition presupposes a correspondence between a statistical property (e.g., a *keyness* score produced by a given metric) and the underlying notion of a word's qualitative outstandingness in a corpus, even allowing for the ranking of key items according to their relative degree of *keyness*. Several issues arise in the implementation of such a definition. The most important concerns the choice of the reference corpus, which largely determines what will be identified as a keyword. Different configurations of the target corpus (*tc*) and the reference corpus (*rc*) are possible: e.g., a single text vs. a text collection; two texts or corpora in a contrastive framework; or texts from a particular domain or genre compared against a large general-language *rc*. Then of course there is the issue of the choice of statistical metric to apply. *Keyness* measures (KMs) are generally grouped into three types of indicators: *effect size* (ES) (magnitude: how large is the difference?); *statistical significance* (SS) (reliability: is

the difference real or due to chance?); and *dispersion/variability* (distribution across texts: how evenly a word is distributed across the corpus). A fourth type consists of *distance/divergence* (DIV) measures, which treat keyness as a difference between distributions rather than simply a ratio or a hypothesis test. Modern approaches often combine the first three for more robust results, whereas the 4th-way view arguably offers a more comprehensive and up-to-date understanding of keyness. A further issue is the selection of the unit of analysis, which is typically the word, although it may be any lexical or morphosyntactic unit, or particular semantic or pragmatic category. At this stage, the analyst may also apply frequency thresholds in order to focus on a frequency range of interest. Finally, the selection of keywords is particularly important, especially in real-world large corpus studies: beyond being dependent on all previous choices, this step cannot be reduced to a routine list of a few top-ranked items. If anything, they are *keyword candidates* (KCs), whose status as actual keywords must be validated using additional, often somewhat subjective criteria. This is because the concept of "keyword" is rather subjective and depends on the objectives that are being pursued in the analysis.

Drawing on a corpus of 78 Italian crime novels written by 16 authors and published by Sellerio (Tuzzi, 2026), this study investigates the concept of keyness with a twofold objective. First, it assesses the strengths and limitations of different KMs by examining their consistency, degree of overlap, and practical applicability in large-corpus analysis. In particular, we analyze the behaviour of eleven KMs, only one of which—the *Log Ratio Conservative* (LRC) (Evert, 2022)—is purely of the SS type. The others, although formally classifiable as ES, SS, or DIV measures, can in practice be treated as measures yielding a magnitude score to which a significance filter may optionally be applied. Second, the study uses these measures to identify and characterize the lexical profile that distinguishes Andrea Camilleri within the broader landscape of Sellerio crime fiction.

2. Keyness for Italian contemporary popular mystery novels

The data consist of a corpus of 78 contemporary Italian mystery novels (*gialli italiani*) written by 16 different authors. After simple tokenization based on alphabetic sequences only, and normalization through lowercasing, the corpus was reduced to a term-document matrix with 111,066 rows (word types) and 16 columns (authors), for a total of 5,022,784 word tokens.

In the keyness analysis of Andrea Camilleri, we retained word forms as the units of analysis, without applying any pre-filtering based on either frequency (see comments below) or POS tagging. With regard to the *rc*, we reduced the polytomous classification problem to a binary comparison: Camilleri as the *tc* versus the other authors as the *rc* (i.e., the remainder of the corpus after excluding the *tc*). Note, however, that this is not a strictly binary case, since the *rc* is defined by *negation*, so that the analyzed words are taken exclusively from the *tc*. The resulting matrix comparing the *tc* and the *rc* consists of 26,542 rows (words for which Camilleri's frequency, f_1 , is non-zero) and two columns containing the frequencies in the target and reference corpora (f_1, f_2).

At this point, it is crucial to decide whether to consider all the words in the *tc* or to analyse a reduced subset obtained by excluding either (i) *absences* in the *rc* ($f_2 = 0$) or, more restrictively, (ii) words whose frequency falls below a given threshold (e.g., $f_i > 5$). As for (ii), this is a subjective decision and, from a theoretical point of view, it should be avoided, as it amounts to a form of cherry-picking (Gabrielatos, 2018), since it may exclude not only important absences but also large frequency differences, by removing items with very low frequency in one corpus but very high frequency in the other. It may, however, be applied after the keyness analysis.

As for (i), it is worth noting that the proportion of exclusive words of the *tc* amounts to 33,3% of Camilleri's vocabulary, that is the highest observed and considerably higher than that of the other authors (ranging from 8% to 16%). In addition, many such *exclusives* fall within the medium-frequency linguistic zone of the *tc*. On this basis, we decided to begin with a full analysis of all Camilleri items in order to obtain a comprehensive overview of the KCs, while potentially adjusting the choice of metrics (as they are differentially sensitive to absences) and filtering out non-(statistically) significant items, so as to reduce low-frequency terms that may generate high spurious scores. Subsequently, we repeated the analysis after removing absences in the *rc*.

We examined several KMs, ranging from standard measures used within either ES-type approaches or SS tests to less common ones that quantify DIV between distributions. In detail, letting f_i denote a given word frequency in corpus i , n_i the size of corpus i , and $i = 1/2$ the *tc/rc*,

1. log-(Relative Frequency) Ratio (*IR*): $IR = \log_2 \left(\frac{f_1/n_1}{f_2/n_2} \right)$

We add 0.5 to f_i to avoid undefined results (Walter's adjusted estimator, 1975)

2. *Simple Maths* (*IRFR_N*): $\log_2 \left(\frac{p_1 \cdot 10^6 + N}{p_2 \cdot 10^6 + N} \right)$ with $p_i = \frac{f_i}{n_i}$

They are calculated using the commercial corpus analysis platform Sketch Engine (Kilgariff, 2009): first, each word's (relative) frequency is multiplied by 10^6 to obtain frequencies per million (*wpm*, words per million), second, a constant N is added to all normalized frequencies in order to emphasize more frequent words.

3. log-Odds Ratio (*IOR*): $IOR = \log \left(\frac{f_1/n_1}{f_2/n_2} \right) = \left(\frac{f_1 \cdot (n_2 - f_2)}{(n_1 - f_1) \cdot f_2} \right)$

The same adjustment as for *IR* needs to be applied here.

4. Difference Coefficient (*DC*): $DC = \frac{f_1 - f_2}{f_1 + f_2}$

5. log-Likelihood Ratio (LLR) or $G^2 = -2LLR$:

$$G^2 = 2 \cdot \left(f_1 \log \frac{f_1}{f_1^*} + (n_1 - f_1) \log \frac{n_1 - f_1}{(n_1 - f_1)^*} + f_2 \log \frac{f_2}{f_2^*} + (n_2 - f_2) \log \frac{n_2 - f_2}{(n_2 - f_2)^*} \right)$$

where a starred symbol indicates the expected frequency under the null hypothesis that the distribution of the word in the two corpora have the same underlying parameter (represented by the subset $\Omega_0 = \{(\pi_1, \pi_2) | \pi_1 = \pi_2\}$). G^2 is asymptotically χ^2 -distributed with 1 degree of freedom.

6. Pearson's X^2 : $X^2 = \frac{(f_1 - f_1^*)^2}{f_1^*} + \frac{((n_1 - f_1) - (n_1 - f_1)^*)^2}{(n_1 - f_1)^*} + \frac{(f_2 - f_2^*)^2}{f_2^*} + \frac{((n_2 - f_2) - (n_2 - f_2)^*)^2}{(n_2 - f_2)^*}$

X^2 is a good approximation of G^2 provided that expectations are more than roughly 5.

7. *phi* coefficient (ϕ or r_ϕ) or Matthews correlation coefficient *phi*: $\phi = \frac{f_1 \cdot (n_2 - f_2) - f_2 \cdot (n_1 - f_1)}{\sqrt{(f_1 + f_2) \cdot (n_1 - f_1 + n_2 - f_2) \cdot n_1 \cdot n_2}}$

where $\phi = \sqrt{\phi^2}$ with *inertia* $\phi^2 = \frac{X^2}{n}$ and $n = n_1 + n_2$.

8. Hypergeometric-based one-sided test for over-representation (*hyp*):

Suppose that the total number of tokens in a urn is $n = n_1 + n_2$, of which $m = f_1 + f_2$ are associated with a certain word of interest w . If we select n_1 tokens randomly from the urn without replacement, the probability of seeing f_1 or less tokens of type w in the n_1 drawn can be calculated as the cumulative probability of a hypergeometric distribution,

$$hyp = Prob(X \leq f_1; m, n, n_1) = \sum_{x=0}^{f_1} \frac{\binom{m}{x} \binom{n-m}{n_1-x}}{\binom{n}{n_1}}$$

hyp provides a ES measure. In an SS-based interpretation, a word-type w is considered (statistically) significantly enriched if its p -value (calculated as the retro-cumulative probability) is less than a given α after adjusting for multiple hypothesis testing.

9. Distinctiveness (*dstnct*) of word w for a target corpus: $dstnct(w) = \sum_i P(c_i|w) \log \frac{P(c_i|w)}{P(c_i)} = (f_1 + f_2)^{-1} \left(f_1 \log \frac{f_1}{f_1^*} + f_2 \log \frac{f_2}{f_2^*} \right)$ where c_i is any corpus i at comparison.

dstnct of word w for a c_i is defined as the Kullback-Leibler divergence between $P(c_i|w)$ —the

likelihood that word w was drawn from c_i , and $P(c_i)$ —the likelihood that any randomly-selected word w' was extracted from c_i . The formula has been reinterpreted from the original set up for topic quality evaluation (Chuang et al., 2012). This formulation describes how much *key* the specific term w is for identifying the target corpus, versus a randomly-selected term w' .

10. Saliency (*slnc*) of word w for a target corpus: $slnc(w) = P(w) \cdot dstnct(w) = \frac{f_1+f_2}{n}dstnct(w)$. It is used to prevent very rare but super distinctive words from dominating.

11. Log Ratio conservative (LRC) is a conservative estimate of log-Ratio based on a confidence interval estimate, which is 0 if the score is not (statistically) significant (Evert, 2022).

The above metrics can be formally distinguished into three groups: ES (*lR*, *lRFR_N*, *lOR*, *DC*), SS (G^2 , X^2 , ϕ , *hyp*, *LRC*), and DIV (*dstnct*, *slnc*) metrics. Although initially treated as ES-type measures (except for LRC), they are subsequently combined with a *significance filter* to weed out candidate items lacking (statistically) significant evidence against the null hypothesis $H_0 : \pi_1 = \pi_2$.

Note that G^2 , X^2 , *dstnct* (and *slnc*) are ≥ 0 . However, after multiplying them by $sign(f_1 - f_1^*)$, the resulting scores take either a $+/-$ sign, thereby becoming *directional* KMs. All the other measures are naturally directional, with the exception of *hyp*, which takes values in $[0, 1]$.

2.1. Comparing keyness measures

We examined each measure distribution and, in particular, compared the keyword rankings they produced to determine whether any measure was effectively a “copy” of another or otherwise *redundant*. From both a direct comparison of the normalized scores of each measure (figure available on request) and an analysis of the correlation between keyword rankings (Figure 1) we found that: *lR* = *lOR* and G^2 = *slcn* once normalized; X^2 and ϕ produce exactly the same ranking, which is very close to that of G^2 ; *DC* and *dstnct* yield the same ranking, close both to that of *lR* and to that of *hyp*; *lRFR₁* is close to X^2 and lies at equal distance from G^2 and *lR*; *lRFR₁₀₀* and *lRFR₁₀₀₀* differ more substantially from the other measures, particularly from the cluster organized around *lR*, while remaining relatively close to G^2 (recall that they are more strongly driven by commoner words). In summary, three clusters organized around *lR*, G^2 and *lRFR_{N(>>1)}* can be identified. In the following analyses all redundant measures are omitted.

We then investigate how candidate items are distributed across the frequency zones of both the *tc* and *rc*, and whether different KMs produce different distributions of KCs across these zones. Topographic maps representing the (f_1, f_2) coordinate space provide useful insights into the distinct mathematical properties of the KMs. It is apparent that some measures generate more KCs in the low-frequency zone (e.g., *dstnct* and *lR*), whereas others produce more KCs in the medium/high-frequency zone (e.g., G^2 and *lRFR_{N(>>1)}*) (Figures 2 and 3).

2.2. Adding Significance

Pure ES measures ignore statistical uncertainty in the form of sampling variation and are prone to yielding spuriously high scores for low-frequency terms. If a significance filter is applied *appropriately* to the ES measures, 24,214 are deemed non-significant for G^2 (and likewise for *lR* and *dstnct*, to which the G^2 filter is applied), 17,064 for *hyp*, and 24,463 for LRC (out of a total of 26542 items), leaving 1,766 significant KCs for G^2 (as well as for *lR* and *dstnct*), 9,478 for *hyp*, and 1,562 for LRC (right-bottom panel of Figure 2).

As expected for a conservative version of *lR*, LRC-based KC frequencies occupy intermediate frequency zones between the lower-frequency ones (e.g., *lR*) and the higher-frequency ones (e.g., G^2 , *lRFR_{N(>>1)}*). However, a further overlap analysis across the different measures showed that LRC is much closer to *lR* than to G^2 , and that *hyp* identifies essentially the same KCs as *lR*

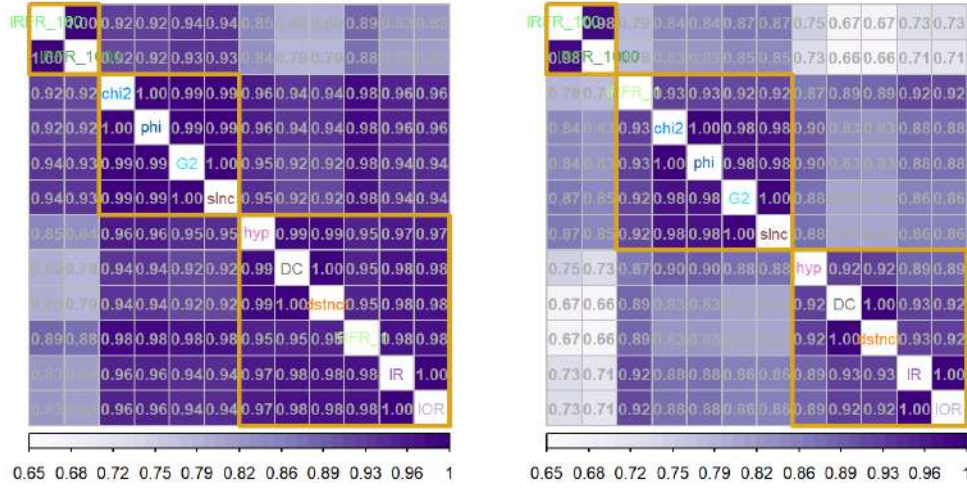


Figure 1: Spearman (left) and Kendall correlation of measure rankings with more relevant clusters highlighted.

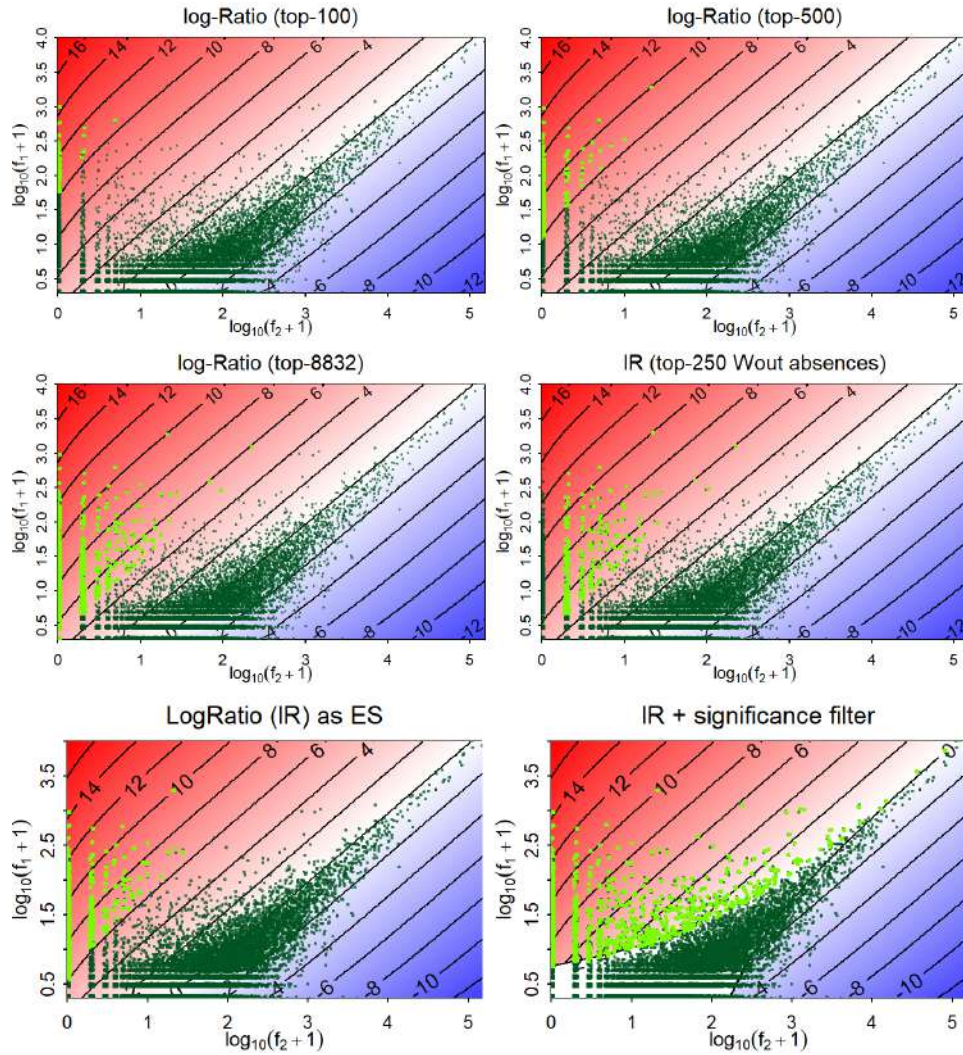


Figure 2: Topographic maps of IR scores: contour plots, overall words (dark green), top- n KCs (light green). Top four panels show top- n KCs without significance filter and include absences except for top-250 panel; bottom panels show top-1766 KCs without (left) and with (right) significance filter.

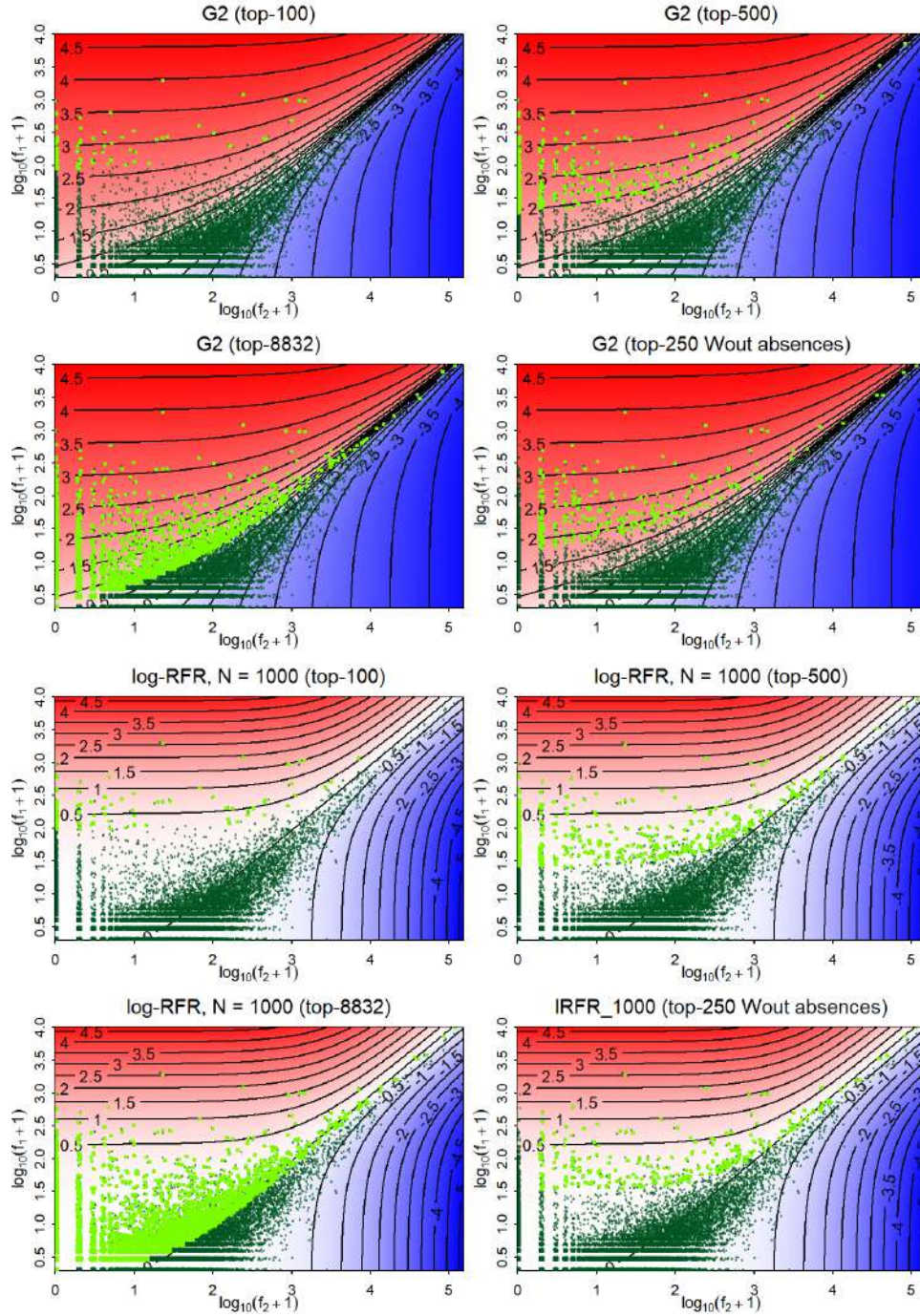


Figure 3: Topographic maps of G^2 (top four panels) and $IRFR_{1000}$ (bottom four panels) scores: contour plots, overall words (dark green), top- n KCs (light green). All include absences except for top-250 panels.

for any chosen top- n list, possibly with slight differences in ranking order.

Finally, rankings produced by lR , LRC , G^2 , and $IRFR_{N(>>1)}$, with and without the significance filter, are presented in tabular form for a qualitative investigation of the KCs.

2.3. Keyness without Camilleri's Exclusives

We repeated the same flow of analyses after removing the absences in the rc , i.e., the exclusives of the tc (8,832 words). In this case, we were left with 649 significant KCs for G^2 (as well as for

dstnct and *lR*), 736 for *hyp*, and 611 for *LRC* (out of 17,710 total words). A quick comparison with the results obtained from the complete target corpus shows that, after removing 1/3 of Camilleri's vocabulary, only about 1/3 of the significant KCs remain (e.g., for *LRC*, from 1,562 to just 611). Such a drastic reduction indicates that it is primarily the exclusive vocabulary that characterizes Camilleri, and confirms the need to begin with a complete analysis before moving to a more focused investigation of the non-exclusive vocabulary. Nevertheless, non-exclusive words were already top-ranked in the keyness analysis of the complete target corpus, especially under KMs less biased toward low-frequency zones (such as G^2 and $IRFR_{N(>>1)}$; Figure 4).

3. A qualitative reading of Camilleri's keywords

Lists of words ranked according to keyness measure values, as presented in the preceding sections, can guide research on the writing style of Camilleri. They may corroborate interpretative hypotheses advanced in literary criticism or stylistic analysis, or suggest new lines of inquiry.

A first clear result emerges from the analysis of exclusive lexicon, defined as the set of forms occurring in Camilleri's texts (target corpus) and never in other authors' texts (reference corpus). Although all such forms are, by definition, distinctive, only a subset qualifies as statistically significant keywords (depending on the measure adopted, as some measures are more sensitive to low-frequency items, while others privilege medium- or high-frequency differences).

Exclusive forms amount to 8,832 word types, corresponding to 38,545 word tokens. They consist almost entirely of Sicilian-origin terms, representing the most distinctive feature of Camilleri's vocabulary. This language is a hybrid Sicilian (drawing on different Sicilian varieties), partly invented and generally morphologically adapted to Italian. Only 1,549 types, corresponding to 1,766 tokens, are not Sicilianisms. These include Italian forms, as well as a limited number of loanwords (Anglicisms such as *Anonymous*, *bypass*, *playmen*; Latinisms such as *prosit*-cheers, *sursum*-upwards, *vulnus*-wound; a Portuguese borrowing such as *viados*; and Gallicisms such as *décollage*), alongside proper names (personal names, e.g. *Augello*, *Catarella*, *Eisenhower*, *Fazio*, *Jacomuzzi*, *Mistretta*, *Rahman*, *Liliana*, *Ernestino*, with the already mentioned *Anonymous* also functioning as a proper name); place names such as *Acquapendente*, *Copacabana*, *Eraclea*, *Montereale*; company names, real or fictional, such as *Fininvest* or *Televigata*, and onomatopoeic forms (*ciaf*, *driin*, *ih*, *iih*, *tintin*). As can be inferred from the number of tokens, they are mostly low-frequency forms: 135 words occur twice, and 1,410 are hapaxes. Only four forms display higher frequencies: *principiò*-began, which may function as either Sicilian or Italian depending on the context of use, occurs 49 times (with additional isolated forms from the same verbal paradigm of *principiare*); *Anonymous* occurs 27 times; *portolano*-portolan 7 times; and *tirabaci*-kiss-curl 3 times. The rest of non-Sicilian forms are largely hapaxes, which are difficult to classify into homogeneous categories. However, some recurring categories can be identified: verbal forms accompanied by a clitic pronoun (e.g., *continuarlo*-continue it, *denunziarlo*-denounce him/it, *estrargli*-extract from him, *partirsene*-leave, *pigliarli*-take them, *precedendolo*-preceding him/it, *reggerlo*-hold it, *scassarle*-break them, *sciuparle*-waste them, *scordarseli*-forget them) or adverbs ending in *-mente* (e.g., *testitevasivamente*-evasively, *interminabilmente*-endlessly, *lealmente*-faithfully, *animalescamente*-animal-like, *bellicosamente*-belligerently, *caparbiamente*-stubbornly, *cerimoniosamente*-ceremoniously, *convenzionalmente*-conventionally, *doverosamente*-dutifully, *geometricamente*-geometrically, *gustosamente*-tastefully, *idealisticamente*-idealistically, *illogicamente*-illogically, *imparzialmente*-impartially, *indolentemente*-lazily, *inguaribilmente*-incurably, *leggerissimamente*-very slightly, *linearmente*-linearly, *lussuosamente*-luxuriously, *matematicamente*-mathematically, *minutissimamente*-extremely minutely, *nobilmente*-nobly, *operosa-*

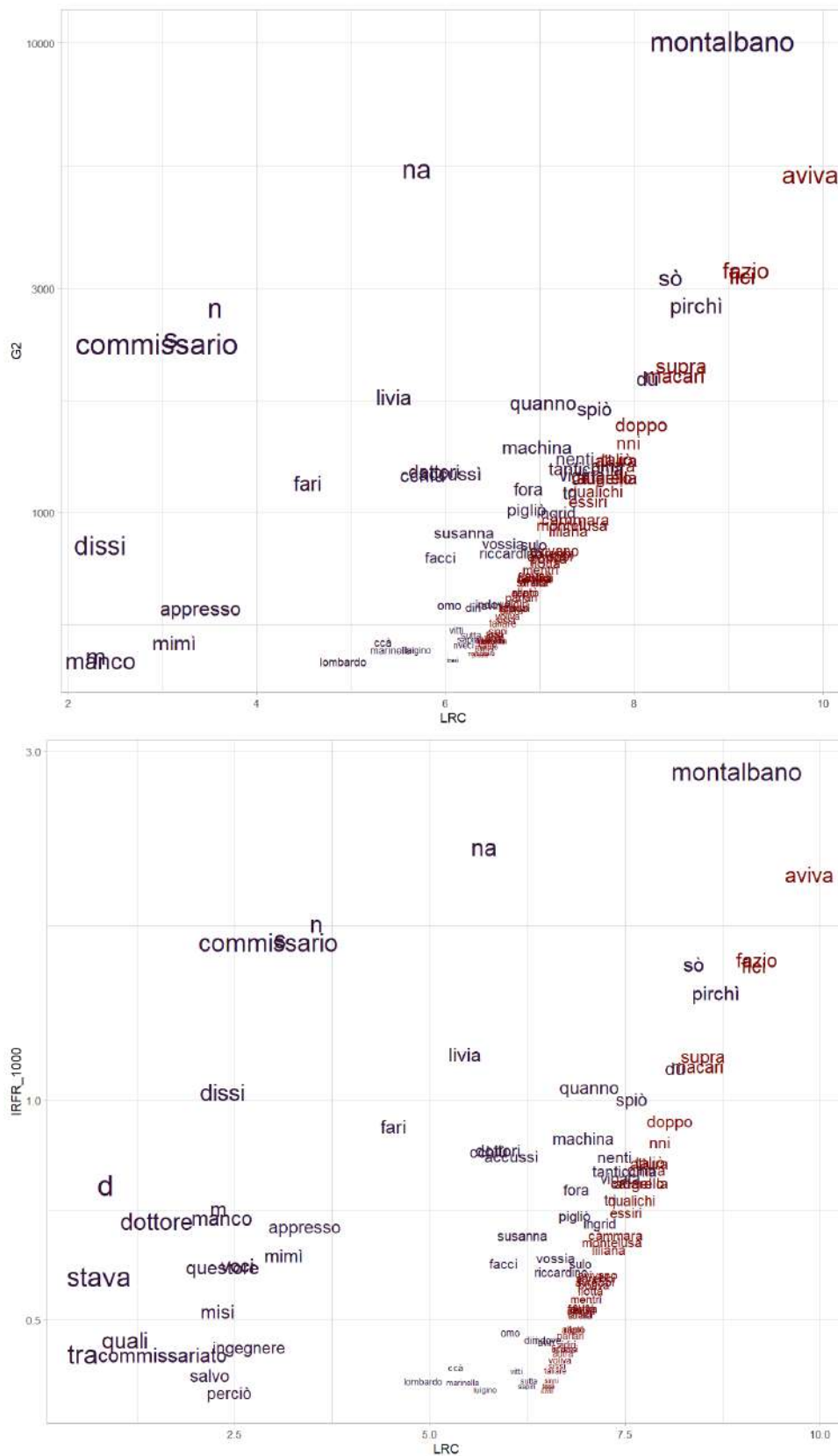


Figure 4: Bidimensional ranking: G^2 (top) and $IRRF_{1000}$ (bottom) on y-axis versus LRC on x-axis. Top-100 significant KCs. Exclusives (red) are 54 (top) and 44 (bottom).

mente-laboriously, *pappagallescamente*-parrot-like, *perigliosamente*-dangerously, *piattamente*-flatly, *rarissimamente*-extremely rarely, *reiteratamente*-repeatedly, *sacrilegamente*-sacrilegiously, *scarsissimamente*-extremely scarcely, *sconciamente*-indecently, *sibillinamente*-sphinx-like).

Among Camilleri's exclusive words, many attest to the author's preference for a high-level register: abstract words (*anonimità*-anonymity, *improcrastinabilità*-improcrastinability, *irreperibilità*-untraceability, *lievità*-lightness, *remissività*-submissiveness, *sinteticità*-conciseness, *ubiquità*-ubiquity), words from specialized sublanguages (*abigeato*-cattle rustling, *autismo*-autism, *biplano*-biplane, *bisarca*-car transporter, *bisurata*-short note, *cardiochirurgo*-cardiac surgeon, *corazzatura*-armoring, *controfirmato*-countersigned, *ecoscandaglio*-echo sounder, *epilessia*-epilepsy, *radiali*-radials, *tirannicida*-tyrannicide, *tirannicidio*-tyrannicide act), and other infrequently used words (*abietta*-abject, *abiurare*-abjure, *acclarate*-ascertained, *acclarerò*-I will verify, *allibiscono*-they are astonished, *allignava*-it was taking root, *centellinandolo*-sipping it, *codicillo*-codicil, *condolersi*-to express condolences, *condolgo*-I offer my condolences, *conficcavano*-they were embedding, *corrugate*-wrinkled, *dedali*-labyrinths, *deprecò*-he/she deprecated, *ebbri*-inebriated, *efebica*-ephebic, *esaudimento*-granting [a wish], *esorcizzata*-exorcised, *giallore*-yellowness, *imberbi*-beardless, *inappetenza*-loss of appetite, *inscena*-he/she stages, *labirintiche*-labyrinthine, *leonino*-lion-like, *maggiorenti*-notables, *malversazioni*-embezzlements, *osteologo*-osteologist, *panoramizando*-scanning panoramically, *pappagallescamente*-parrot-like, *periglio*-danger, *perigliosamente*-dangerously, *perigliose*-dangerous, *reiteratamente*-repeatedly, *sbrilluccicava*-it sparkled, *sciacquatura*-rinsing, *sciacquettio*-splashing, *scialava*-he/she squandered, *squittii*-squeaks, *squittire*-squeak, *sragionando*-ranting, *stortignaccolo*-twisted little neck, *vociferava*-he/she vociferated).

3.1. Camilleri's non-exclusive words

Significant non-exclusive keywords can be useful for interpreting Camilleri's style, as they also appear in the reference corpus, even if only once or twice. Different measures yield different results. In particular, the IR measure ranks almost exclusively Sicilian forms among the top 100 keywords. For Camilleri's style, this adds little beyond the exclusive forms; what is more revealing comes from the reference corpus perspective. After removing true homographs through concordances (e.g., *dottori* as Camilleri's Sicilianized form of *dottore*-doctor vs. standard Italian plural *dottori*-doctors), it was possible to identify the authors who sometimes use Sicilian words that have been assessed as Camilleri's keywords. These include Gian Mauro Costa and Gaetano Savatteri, followed by Santo Piazzesi, Roberto Alajmo (all Sicilian writers), and Antonio Manzini (born in Rome). In Manzini's case, the words that match Camilleri's forms are not Sicilianisms but Romanesco words homographic with Sicilian terms (e.g., *quanno*-when, *macchina*-car). In contrast, the Sicilian authors clearly draw on Camilleri's model to incorporate Sicilian words into their novels, often citing him in their works or prefaces. Differences are nonetheless evident: focusing on the two authors who use Camilleri's keyword Sicilianisms most, Savatteri integrates them almost always within Italian sentence contexts, whereas Costa uses them more frequently in fully Sicilian contexts (i.e., lines spoken by characters in Sicilian). Citations to Camilleri's work, especially in Savatteri's novels, are further evidenced by the presence of Camilleri's protagonist, Montalbano, in the reference corpus (with particular reference to Savatteri). In the ranking of the Camilleri keywords which occur also in the reference corpus according to G^2 , LRC , RFR_{100} , and RFR_{1000} metrics, *Montalbano* ranks first (contrasting with 11th place under the IR metric). *Montalbano* occurs 1,953 times in Camilleri (target corpus) and 21 in the reference corpus, 12 of which are in Savatteri.

A final observation emerges by extracting non-Sicilian words from the top 100 Camilleri key-

words identified by G^2 , *LRC*, *RFR*₁₀₀, and *RFR*₁₀₀₀ metrics. Some keywords appear in all rankings: excluding proper names, examples are *pigliò*-he/she took, *picciotto*-young man (a Sicilianism, also adopted as a regionalism in Italian), *sacchetta*-small bag, *verandina*-small veranda. G^2 , *RFR*₁₀₀, and *RFR*₁₀₀₀ agree on more keywords if *LRC* is disregarded: *commissariato*-police station, *commissario*-commissioner, *dottore*-doctor (the Sicilian form *dottori* is indicated by all metrics), *grotta*-cave, *ingegnere*-engineer, *perciò*-therefore, *misi*-I put, *picciotta*-young woman (feminine of *picciotto*), *pigliato*-taken (past participle, alongside the remote past *pigliò* identified by all measures), *questore*-police chief (alternating with the Sicilian *questori*, also ranked by *LRC*), *villino*-small villa.

Among Camilleri's Italian keywords that are most significant according to all or most metrics, two main semantic clusters emerge: the protagonists' work environment (*commissariato*-police station, *commissario*-commissioner, *dottore*-doctor, *questore*-police chief) and the external environment in which they conduct their investigations (*grotta*-cave, *verandina*-small veranda, *villino*-small villa). These results provide an overview of the picture emerging from the most significant keywords identified by the KMs, and forming the basis for an evaluation of Camilleri's style. From a confirmatory standpoint, one of the clearest findings in Camilleri studies (La Fauci, 2000; Matt, 2020) is confirmed: the rejection of a neutral register, achieved partly through Sicilian resources and partly through Italian words of specialized or elevated register. From an exploratory standpoint, Italian keywords function as the primary means to contextualize events within the narrative environment, both the internal setting of the protagonists' activities and the external setting of the events. These findings remain to be further scrutinized in Camilleri corpus, to assess whether they are consistent across all works or vary by theme or period. At the level of the Crime novels corpus, Camilleri's prose emerges as a model for several Sicilian crime writers published by Sellerio. While clearly influenced by him, these authors follow different paths (less radical than Camilleri's) and also diverge among themselves.

Acknowledgements. The research group gratefully acknowledges Sellerio Editore for providing the corpus and valuable guidance in both the selection of works and access to the digital texts.

References

- Chuang J., Manning C. D., and Heer J. (2012). Termite: visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, p. 74–77. Association for Computing Machinery.
- Evert S. (2022). Measuring keyness. In *Digital Humanities 2022: Conference Abstracts*, pp. 202–205.
- Gabrielatos C. (2018). Keyness analysis: Nature, metrics and techniques. In *Corpus Approaches To Discourse: A critical review*, pp. 225–258. Routledge, Oxford.
- Kilgarriff A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*.
- La Fauci N. (2000). L'italiano perenne e Andrea Camilleri. In N. M. and T. P. S., editors, *Italia linguistica anno mille Italia linguistica anno duemila. Atti del XXXIV Congresso Internazionale di Studi della Società di Linguistica Italiana (Firenze, 19-21 ottobre 2000)*, pp. 331–340. Bulzoni, Roma.
- Matt L. (2020). Lingua e stile della narrativa camilleriana. In Caocci D. M. G. and M.E. R., editors, *Quaderni camilleriani 12. Parole, musica (e immagini)*, pp. 39–93. Grafiche Ghiani, Cagliari.
- Moreno-Ortiz A. (2024). *Making Sense of Large Social Media Corpora: keywords, topics, sentiment, and hashtags in the Coronavirus Twitter Corpus*. Springer Nature Switzerland.
- Scott M. (1997). Pc analysis of key words - and key key words. *System*, 25(2):233–245.
- Tuzzi A. (2026). A new investigation for Sellerio. In *JADT2026 proceedings*.

A New Investigation for Sellerio

Arjuna Tuzzi

University of Padova – arjuna.tuzzi@unipd.it

Abstract

This study presents selected results from an analysis of works published by Sellerio Editore, based on two distinct corpora. The first corpus (78 crime novels by 16 authors) provided an overall mapping of texts and writers in terms of similarities and differences, and situated the six novels by Andrea Camilleri within the broader landscape of contemporary crime fiction. Although the corpus includes both authors with strong individual stylistic traits and others who embody the most canonical traits of the genre, Camilleri's works emerge as markedly distinct from all other texts examined in the study. The second corpus (70 works by Camilleri) allows for a more in-depth investigation of both continuity and discontinuity in his narrative trajectory, as well as a clear separation between the Montalbano series and the rest of his production. Further analysis of the Montalbano series reveals a well-defined diachronic evolution and a structured set of recurring topics, alongside a distinctive use of linguistic registers ranging from standard Italian to Sicilian dialect and beyond. Overall, this first exploratory analysis of the two corpora provides a methodological and empirical basis for several avenues of further research, opening the way to more detailed investigations into stylistic variation, thematic development, and the linguistic strategies that characterize both Camilleri's work and contemporary Italian crime fiction more broadly.

Keywords: Andrea Camilleri, *Camillerese*, diachronic corpora, correspondence analysis, topic extraction

1. Introduction

Andrea Camilleri is widely known to the general public for his crime novels about the investigations of Commissario Salvo Montalbano and for their successful television adaptations. However, his production is far broader, as he was a highly prolific author working across multiple genres and also active as a director, playwright, and screenwriter in both theatre and radio-television contexts.

Studying Camilleri through text data analysis methods allows for the identification of recurrent structures, relevant themes, and narrative dynamics that already vibrate strongly with thousands of contemporary readers and, therefore, deserve a systematic investigation. However, examining Camilleri's work using a scientific approach (whether qualitative, in linguistics and literary criticism, or computational, in statistics) is far from straightforward. Literary scholarship has historically privileged the so-called "high" literature (works recognized by the academic community as canonical and of elevated artistic status) as worthy of academic legitimacy, while marginalizing the so-called "low" literature (popular and genre fiction, including crime novels, noir, and detective narratives) as too commercial to be considered of scientific interest. With regard to language, Andrea Camilleri's work has attracted the interest of several linguists. Since the early 2000s, Nunzio La Fauci (2000; 2022) has devoted a number of studies to Camilleri's language; he was followed by Marina Castiglione (2013), Massimo Arcangeli (2004), Luigi Matt (2020), Roberto Sottile (2017), Gigliola Sulis (2007), and Carolina Tundo (2024). Distinguished observations have been provided by the literary scholars Salvatore Silvano Nigro (2004) and Pietro Mazzarisi (2025), and by Andrea Camilleri himself in dialogue with Tullio De Mauro (2013). However, no study has yet produced a comprehensive

account of Camilleri's language, and no scholar has examined his entire body of work in a systematic manner.

This paper briefly presents selected exploratory findings from a wider study conducted on a balanced corpus of 78 crime novels by 16 authors, which includes six novels by Camilleri, and on a second corpus of 70 works of different genres by Camilleri, which includes 34 works featuring Commissario Montalbano, all published and made available by Sellerio Editore. The aim is to investigate Camilleri's distinctive features within the landscape of contemporary crime writers and, with specific reference to the Montalbano series, to identify the main topics and relevant aspects of the temporal evolution of his lexicon. Sellerio Editore contributed to the development of the two corpora by providing editorial expertise in the selection of works and authors and by granting permission to use the electronic texts of the novels for research purposes. In order to preserve the integrity of the original texts, the preprocessing followed a conservative approach (Tuzzi, 2024), involving only plain tokenization (removing only punctuation and numbers) and lowercasing. The descriptive measures are reported in Table 1.

Table 1. Description of corpora involved in the study.

	n. of texts	n. of authors	N tokens	V types	TTR %	hapax %	min length	max length
Crime novels	78	16	5,022,784	111,066	2.2	38.0	20,133	123,451
subcorpus Camilleri	6	1	319,298	26,542	8,3	51,2	38,897	66,915
Camilleri's works	70	1	3,315,184	96,758	2.9	42.9	6,257	109,575
subcorpus Montalbano	34	1	1,862,335	63,408	3.4	43.2	38,897	89,464

This paper is connected to four contributions published in JADT2026 proceedings, which deal with Camilleri's works and address the issues of genre classification (Sciandra and Cortelazzo, 2026), the role of linguistics resources to manage the *Camillerese* in topic modelling tasks (Genova and Sciandra, 2026), the distinctiveness of Camilleri's language usage (Cortelazzo et al., 2026), and his peculiar narrative style and plot design (Mazzarisi, 2026).

2. The distinctiveness of Andrea Camilleri within a corpus of crime novels

The Crime novels corpus consists of 78 books written by 16 contemporary authors, two of whom are deceased (Andrea Camilleri and Carlo Flamigni). Only writers with at least three crime novels published by 2023 with Sellerio Editore were included. For prolific authors, up to six novels were selected from different time periods, including books featuring the same main characters as well as those with different main characters when available. Figure 1 shows the size of each crime novel in word tokens. The complete list of works and labels is available in the supplementary material.

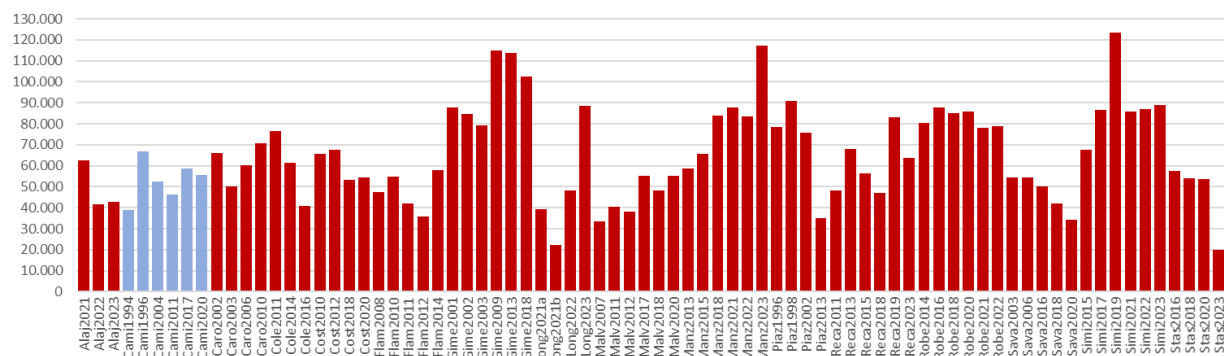


Figure 1. Crime novels corpus: word token counts for the 78 novels; in light blue the six novels by Camilleri.

Correspondence analysis (CA) based on the term-document matrix (TDM) including 4,554 most frequent words (MFW, number of occurrences ≥ 90 , corpus coverage rate 85%) and all 78 crime novels positions Andrea Camilleri's works as clearly distinct from the other 15 authors' ones (fig. 2, top). His works span the first horizontal axis, dominate the right half-plane, pushing all the rest into the left half-plane and, therefore, masking most authors' fine-grained distinctions. Further analyses can easily highlight that the authors of this corpus display very interesting individual traits and a clear-cut profile in authorship attribution tasks, however, none with the power and uniqueness of Camilleri (Sciandra and Tuzzi, 2024; 2025). CA shows that after the 1990s Camilleri progressively moved away from the lexical space of the other authors, as if he had gradually created a language and a narrative world of his own, and then, in more recent years, seems to regress toward the common space (Riccardino, 2020; *La rete di protezione*, 2017). The most significant words are, as expected, names of places, actions, and characters involved in the investigations of *Commissario Montalbano*, such as *Vigàta*, *Marinella*, *Montelusa*, *Fazio*, *Mimì*, *Augello*, *Livia*, *Catarella*, *Gallo*, *Ingrid*, *Riccardino*, *Luigino*, *Susanna*, *Liliana*, *Lombardo*, *Licausi*, *Liotta*. However, many words are also part of both Sicilian dialect and Camilleri-style lexicon (*accussi*-like that, *allato*-beside, *allura*-then,

arrispunni-answer, *assà*-very much, *autra*-other, *autro*-other, *aviri*-have, *aviva*-had, *avivano*-they had, *avrebbe*-would have, *bono*-good, *ccà*-here, *cchiù*-more, *detti*-said, *dintra*-inside, *diri*-say, *doppo*-after, *doviva*-had to, *dottori*-doctor, *essiri*-be, *facenna*-did, *facci*-do, *faciva*-did, *fici*-did, *fimmina*-woman, *fora*-outside, *forsi*-perhaps, *indove*-where, *longo*-long, *macari*-also, *machina*-car, *mentri*-while, *minutolo*-minute, *nenti*-nothing, *niscì*-went out, *nni*-of them, *omo*-man, *pariva*-seemed, *parlari*-speak, *picciotta*-girl, *picciotto*-boy, *pirchì*-because, *pirsona*-person, *potiva*-could, *propio*-just, *punni*-punches, *qualichi*-some, *quanno*-when, *rapri*-opened, *sapiri*-know, *sarebbi*-would be, *sinni*-itself, *sissi*-yes, *stisso*-same, *strata*-street, *sulo*-only, *sunno*-are, *supra*-above, *sutta*-under, *taliare*-look, *taliò*-looked, *tanticchia*-a little, *telefono*-telephone, *trasi*-get in, *tri*-three, *vidiri*-see, *vinni*-came, *vitti*-saw, *voli*-want, *voliva*-wanted, *vossia*-your honor, *zito*-fidanzato, *vucca*-mouth). Distinctive syntactic features are also evident, for example the frequent use of the remote past tense. CA (Lebart et al., 1984) has been performed using FactoMiner R-package (Le et al., 2008).

2. Distant reading Camilleri's works at a glance

The Camilleri corpus includes 70 works published between 1978 and 2024; 34 works tell of the investigations of Commissario Montalbano and constitute a specific subcorpus. Figure 3 shows the size of each work in word tokens. The complete list of works and labels is available in the supplementary material, which also reports the year of first publication and the edition used for corpus compilation. In terms of size, two theatrical works are relatively short (*Conversazione su Tiresia* and *Autodifesa di Caino*, both 2019), and a collection of letters published after Camilleri's death is the largest text (*Vi scriverò ancora. Lettere alla famiglia 1949–1960*, 2024); all the rest fall within a range from a minimum of 14,741 (*La strage dimenticata*, 1985) to a maximum of 99,361 word tokens (*Il Re del Girgenti*, 2001). With reference to the Montalbano subcorpus (fig. 3, in red), the largest works are three short-story collections (*Un mese con Montalbano*, 1998; *La prima indagine di Montalbano*, 2004; *Gli arancini di Montalbano*, 1999); the shortest novel is the first in the series (*La forma dell'acqua*, 1994), and the largest is the second one (*Il cane di terracotta*, 1996).

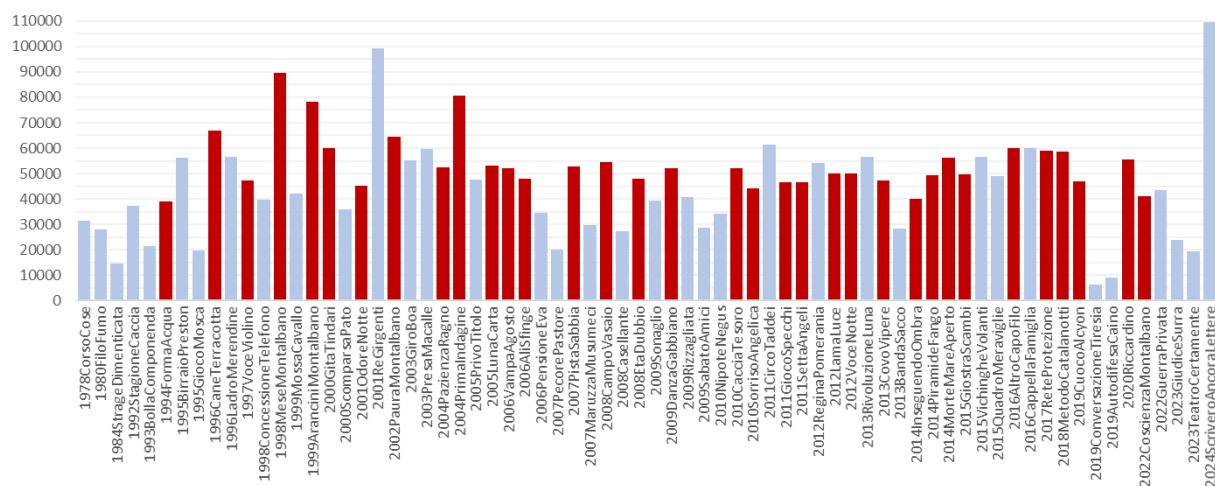


Figure 3. Camilleri corpus: word token counts for the 70 works; in red the Montalbano subcorpus.

CA based on the TDM including 3,079 MFW (occurrences ≥ 100 ; corpus coverage rate 82%) and all 70 works (fig. 4) reveals a clear distinction between the lower half-plane, largely occupied by texts from the Montalbano subcorpus, and the upper half-plane, which contains the remaining works. More recent works cluster in the right half-plane, while earlier texts are

mainly located on the left. As it involves both stopwords and content words, the analysis captures stylistic and thematic similarities among the texts and uncovers a temporal pattern.

CA based on the TDM including 1,828 MFW (occurrences ≥ 100 ; corpus coverage rate 79%) and the 34 works of the Montalbano series (fig. 5) reveals the existence of a clear-cut temporal evolution. The books align along the Cartesian plane quadrants and clearly depict an initial phase (1994-2002, top-left), a second phase in the early 2000s (2003-2005, bottom-left), a third phase (2006-2010, bottom-right), and a final phase (2011-2020, top-right). The typical horseshoe pattern highlights not only the chronological progression along the first axis but also significant elements of discontinuity. A shifted position is observed for several recent works (*Il cuoco dell'Alcyon*, 2019; *Riccardino*, 2020; and, to some extent, the collection *La coscienza di Montalbano*, 2022), which move backward along the temporal trajectory, with *Il metodo Catalanotti* (2018) as its extreme (top-right). Further investigation of the different versions of *Riccardino* and of the genesis of the stories in the collection *La coscienza di Montalbano* may explain this bias, as well as the possible influence of Valentina Alferj, who acted as Camilleri's assistant and transcribed the books he dictated after he lost his sight. Also noteworthy is the role of several pivotal crime novels (*La vampa di Agosto*, 2006; *Il campo del vasaio*, 2008; and, partly, *La pista di sabbia*, 2007), which may reflect developments in the portrayal of female characters. On the one hand, dangerous, manipulative, and seductive female figures emerge; on the other, Montalbano, long depicted as the devoted partner of Livia Burlando, engages in occasional affairs and eventually finds a new love in Laura Belladonna (*L'età del dubbio*, 2008). However, these hypotheses require further verification through the traditional methods of philology, linguistics, and literary criticism.

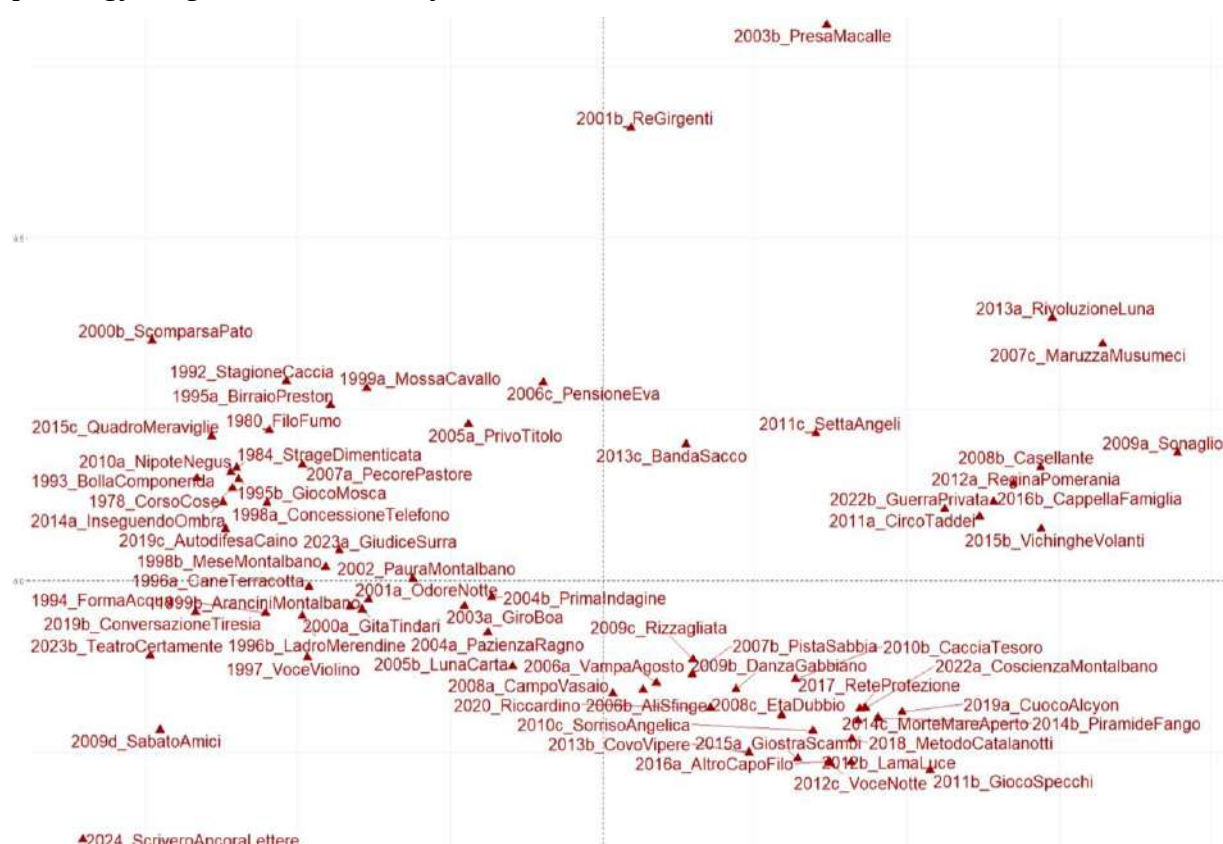


Figure 4. Camilleri corpus: projection of 70 works on CA first (13.1%) and second (5.4%) axis.

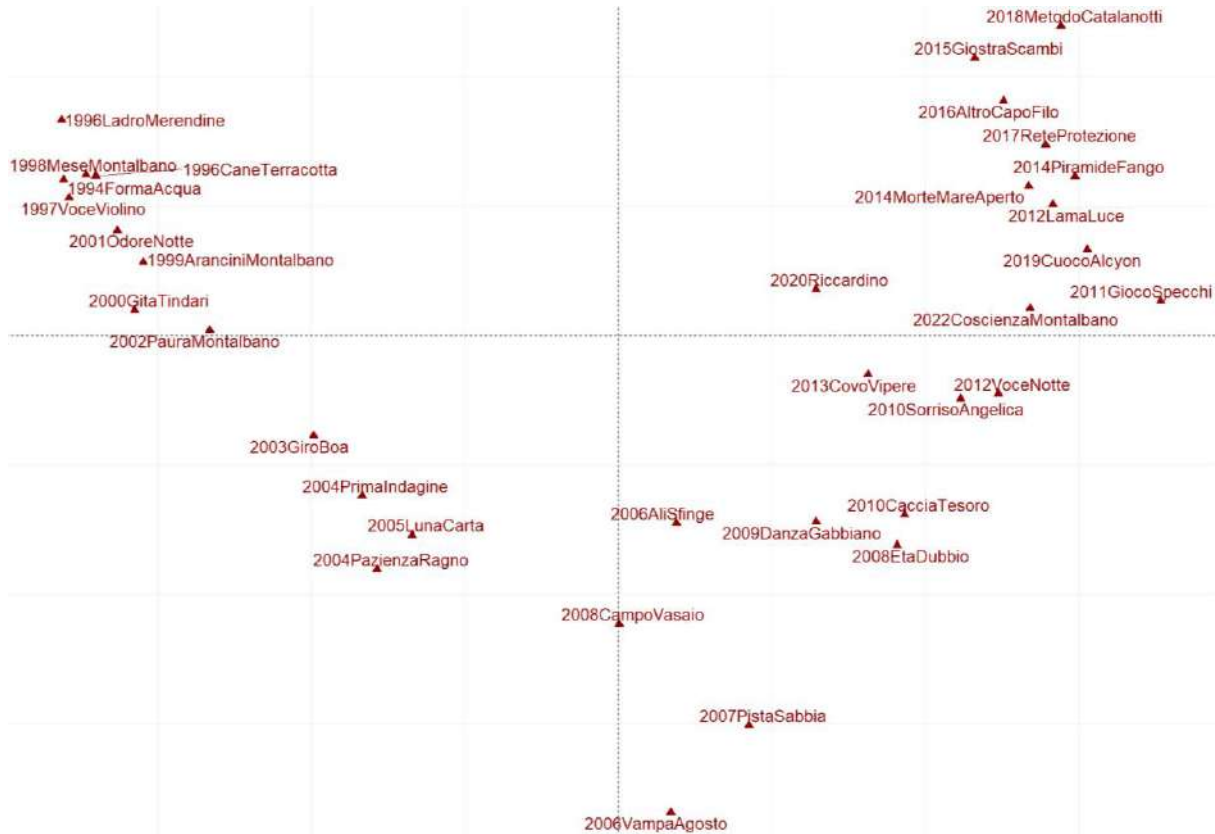


Figure 5. Montalbano subcorpus: projections of 34 works on CA first (25.7%) and second (5.6%) axis (left) and of 34 works of the Montalbano series on CA first (%) and second (%) axis (right).

3. Camillerese: the role of Camilleri-style lexicon

Camillerese (or *Vigatese*) is a language created by Andrea Camilleri to give voice to the characters of his stories and to shape the vivid atmosphere of the microcosm of Vigàta, the fictional setting of the Commissario Montalbano series. It is neither an existing Sicilian dialect nor an area-specific variety of Italian, but rather an authorial *idiolect*, formed through the combination of standard Italian, Sicilian dialect, archaic expressions, neologisms, and phonetically adapted forms, according to narrative, rhythmic, and theatrical needs. Although never spoken outside Camilleri's world, it has reached a wide readership, proving remarkably accessible and successful even among non-Sicilian readers.

The assessment of *Camillerese* across a large corpus of Camilleri's works faces substantial methodological challenges, i.e., full corpus lemmatization is not viable (available lemmatizers cannot process texts combining Italian and Sicilian with non-standard, idiosyncratic forms) and no fully integrated digital resources exist neither for Sicilian nor for *Camillerese*. For the purposes of this study, the occurrence of *Camillerese* was estimated using a non-academic lexical resource compiled by Mario Genco and available on the official website of Camilleri's fandom, which has been accredited by Andrea Camilleri himself as an official source (www.vigata.org). The number of occurrences of *Camillerese* words out of 1,000 tokens shows the highest values in the *Trilogy of Metamorphoses* (*Maruzza Musumeci*, 2007; *Il casellante*, 2008; *Il sonaglio*, 2009), *La presa di Macallè* (2003) and *Il re di Girgenti* (2001), while in theatrical, essayistic, and epistolary works (*Autodifesa di Caino*, 2019; *Il teatro certamente*, 2023; *Conversazione su Tiresia*, 2019; *Vi scriverò ancora. Lettere alla famiglia 1949–1960*, 2024) shows low values as well as in some civic-oriented works, with particular reference to

first editions published in origin with publishing houses different from Sellerio Editore. Within the Montalbano subcorpus (fig. 6, left), an increasing temporal trend can be observed: *Camillerese* represented a rare event in late-1990s novels and increased from the 2000s ongoing. Peaks occur in *Il giro di boa* (2003), *La pista di sabbia* (2007), *Il gioco degli specchi* (2011), *La voce del violino* (2012), *La piramide di fango* (2014), and *Il metodo Catalanotti* (2018). A controlled analysis of *Camillerese* stopwords (Genova & Sciandra, 2026) corroborates these patterns, showing a clearer upward trajectory within the Montalbano subcorpus (fig. 6, right).

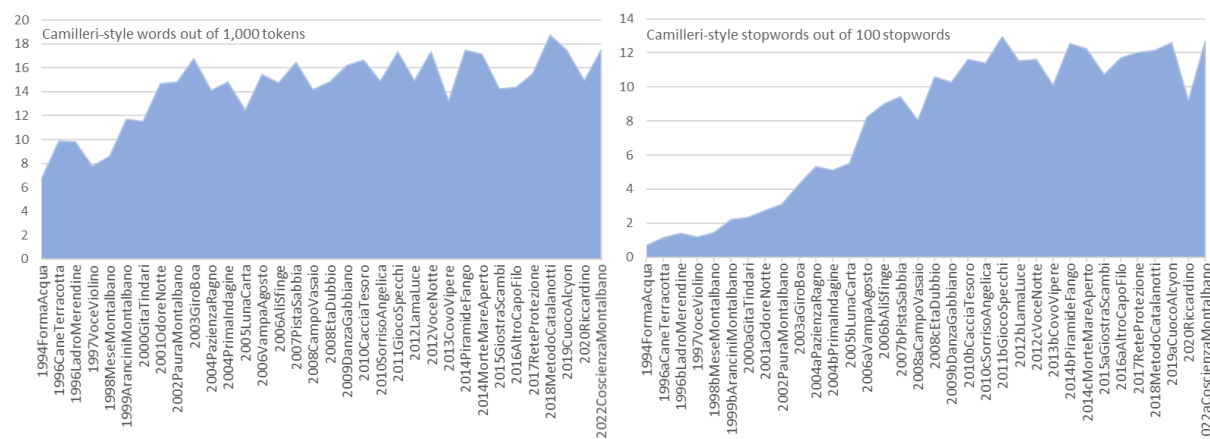


Figure 6. Montalbano subcorpus: occurrences of *Camillerese* words per 1,000 tokens (left) and *Camillerese* stopwords per 100 stopwords (right).

Despite methodological limitations, our data indicate a progressively consolidated and recurrent linguistic strategy, while allowing for recurring cases of standardization, variation, and experimentation. Results suggest the existence of a temporal pattern and a combined effect of publication date, genre, and publisher. However, again, these hypotheses require further verification through the traditional methods of philology, linguistics, and literary criticism.

4. Main topics in Commissario Montalbano's series

As shown in the paper by Genova and Sciandra (2026), topic extraction from the works of the Camilleri corpus requires the construction of *ad hoc* linguistic resources. The results presented here constitute a first exploratory analysis of the Montalbano subcorpus. In these analyses, stopwords (from Italian, Sicilian, and *Camillerese*) were removed in the preprocessing phase ($N_{\text{red}} = 801,759$ word tokens). Topics (lexical worlds) were generated using Reinert's divisive hierarchical algorithm for text clustering (Reinert, 1983;1990) performed using the *rainette* R-package (Barnier, 2023), on 10,668 MFW (occurrences ≥ 10) and 20,045 text chunks of 40 tokens in size. From an interpretative perspective, the optimal eight-class structure appears well balanced in terms of the proportion of attributed text chunks (fig. 7). It reduces the complexity of Montalbano's world to eight core topics and highlights recurrent topics as well as several relevant dichotomies, such as bodily vs. environmental descriptions, domestic vs. work situations, and *Camillerese*-informal vs. institutional-formal language. In brief, clusters 1 and 2 illustrate the physical and environmental descriptions of the stories; clusters 3 and 4 the investigative domain; clusters 5 and 6 everyday life, and clusters 7 and 8 two complementary dimensions of activities and relationships within the police station.

Cluster 1 represents the body (*spalle*-shoulders, *voce*-voice, *faccia*-face, *bocca*-mouth, *braccio*-arm, *occhi*-eyes, *gambe*-legs, *mani*-hands, *braccia*-arms, *piedi*-feet, *capelli*-hair) and perception verbs (*sentì*-heard, *teneva*-held, *pareva*-seemed). The topic describes physical

appearance and bodily attitudes, associated with intimate scenes, emotions, and close observation (*luce*-light, *letto*-bed, *poltrona*-armchair, *vestito*-dress, *mutande*-underwear). In his books, Camilleri often illustrates Montalbano's daily bodily routines (waking up, getting out of bed, dressing). Cluster 2 concerns landscapes, external places, and natural settings in which the narratives unfold (*macchina*-car, *mare*-sea, *auto*-car, *cane*-dog, *scogli*-sea rocks, *strada*-road, *metri*-meters, *collina*-hill, *villa*-villa, *rena*-sand, *grotta*-cave). It groups natural and geographical settings, local toponyms, and landscapes that provide the background for Montalbano's investigations across the Sicilian territory.

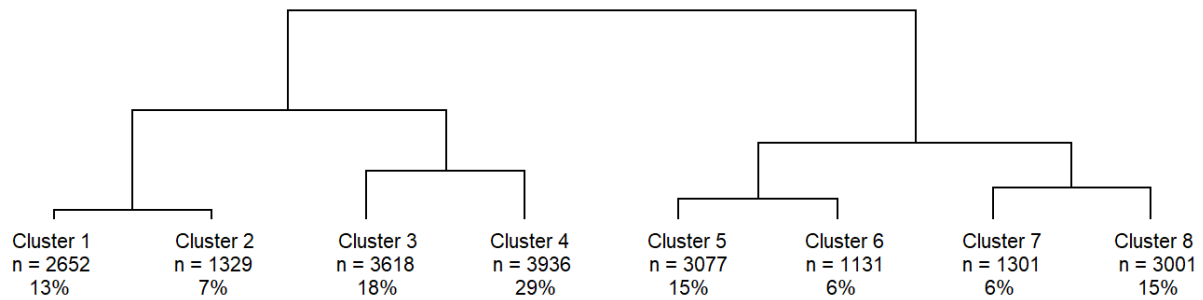


Figure 7. *Montalbano subcorpus*. Dendrogram of the 8 lexical worlds identified. Number and percentage of chunks attributed to each cluster.

Cluster 3 represents family relations (including ties within mafia families) and the life histories of the characters involved in the investigations (*padre*-father, *fratello*-brother, *amante*-lover, *famiglia*-family, *marito*-husband, *moglie*-wife, *Balduccio*, *Sinagra*, *onorevole*-member of Parliament, *ingegnere*-engineer). The topic evokes dynamics often connected to evidence, profiling, investigative cases, information collection, clues and suspicions (*nome*-name, *soldi*-money, *denaro*-money, *milioni*-millions, *lettera*-letter, *lettere*-letters), with emphasis on long-term temporal references (*anni*-years, *mesi*-months) and the characters' biography. Cluster 4 focuses on institutional dialogues and interrogations (*questore*-police superintendent, *dottor*-doctor, *signor*-sir, *numero*-number, *telefono*-telephone, *telefonato*-called, *telefonata*-phone call, *senta*-listen, *vorrei*-I would like, *volevo*-I wanted, *vuole*-you want, *parlare*-speak, *capito*-understood, *capisco*-I understand, *domani*-tomorrow, *ieri*-yesterday, *sera*-evening, *stamattina*-this morning, *mattina*-morning, *pomeriggio*-afternoon, *giorni*-days). The topic reveals a formal register used in official interactions such as phone calls, interviews, and interrogations, characterized by institutional titles and short-term temporal references.

Cluster 5 focuses on body, movement, space, and domestic environment (*mano*-hand, *spalli*-shoulders, *pedi*-feet, *macchina*-car, *cammara*-room, *portoni*-gate/s, *finestra*-window, *dritta*-right, *addritta*-right, *raprì*-opened, *scinnì*-went down). Words are associated with house and movement in everyday situations but in this topic the *Camillerese* emerge stronger than in cluster 1, enriching descriptions and making dialogues more vivid and realistic. Cluster 6 concerns food and everyday rituals (*verandina*-small veranda, *buttiglia*-bottle, *piatto*-plate, *bicchieri*-glasses, *tavola*-table, *cucina*-kitchen, *pasta*, *forno*-oven, *spaghetti*-spaghetti, *passiata*-walk, *conzò*-seasoned, *mangiari*-eat, *whisky*, *vino*-wine, *caffè*-coffee). The topic captures Montalbano's gastronomic passion, including typical dishes, food rituals, domestic settings, and two key characters such as *Adelina* and *Enzo*. The register remains informal and domestic, rich in *Camillerese* words and often associated with moments of pause and pleasure.

Cluster 7 represents daily life in the police station, characterized by gags and dialogs (*Catarella*, *Catarè*, *dottori*-doctor, *signori*-gentlemen, *vossia*-your honor, *telefono*-telephone, *linia*-line, *nummaro*-number, *passamillo*-pass it to me, *sissi*-yes, *nonsi*-no, *pirsona*-person,

pirsonalmente-personally, *subitissimo*-immediately). The topic expresses a comic and surreal linguistic register marked by errors, humorous situations, and phonetic distortions. The author employs agent *Catarella* to create humor, again with a strong use of *Camillerese*. Cluster 8 concerns discussion, thoughts and reflections (*sunno*-I am, *sapiri*-know, *ditto*-said, *secunno*-according to, *sugno*-I am, *dinaro*-money, *dimanna*-question, *vota*-time, *conclusioni*-conclusions, *saccio*-I know, *raggiuni*-reason). The topic contains words associated with investigative reasoning, deduction, and the formulation of conclusions. Two key collaborators, *Fazio* and *Augello*, play a crucial role in the process of case solving.

5. Conclusions

Grounded in a wider study on two corpora, the analysis of a first corpus of crime novels published by Sellerio Editore consistently situates Andrea Camilleri as clearly distinct from all other crime fiction writers, whose profiles do not reach his singular uniqueness. Within a second corpus composed of his works from different genres, the analysis captures both stylistic and thematic patterns and reveals interesting temporal trends. A relevant distinction emerges between the Montalbano series and other works, as well as between older and more recent works and between Italian and *Camillerese* language usage. Focusing on the Montalbano series, the analysis identifies temporal patterns and main topics. The *Camillerese* usage increased steadily from the 2000s, a trend reinforced by more controlled analysis on stopwords. Montalbano's world can be reduced to eight core topics, highlighting recurrent themes and key dichotomies, such as bodily vs. environmental description, domestic vs. work situations, and informal vs. formal register.

Although Andrea Camilleri is widely recognized as a major and eclectic contemporary Italian writer and one of the most beloved *cantastorie* (storyteller) by the general public, both in Italy and internationally, his literary and artistic work spans multiple genres and media, rendering him a rich subject for further systematic scientific analysis. Text analysis uncovers recurring structures, thematic patterns, and distinctive narrative strategies that deserve further investigation through qualitative and quantitative approaches.

Aknowledgements

The research group gratefully acknowledges Sellerio Editore for providing the two corpora and for their valuable guidance in both the selection of works and access to the digital texts.

Supplementary Material

The supplementary material is available at <https://github.com/arjtuzzi/2ListeSellerio>

References

- Arcangeli M. (2004). Andrea Camilleri tra espressionismo giocoso e sicilianità straniata. Il ciclo di Montalbano. In: Marci G. editor, *Lingua, storia, gioco e moralità nel mondo di Andrea Camilleri. Atti del seminario (Cagliari, 9 marzo 2004)*. CUEC, Cagliari, pp. 203-232.
- Barnier J (2023). rainette: The Reinert Method for Textual Data Clustering. R package.
- Camilleri A. and De Mauro T. (2013). *La lingua batte dove il dente duole*. Laterza, Roma-Bari.
- Castiglione M. (2013). Intorno al nome camilleriano, tra motore e scioglimento dell'azione narrativa. *Il nome nel testo*, 15: 191-204.
- Cortelazzo M.A., Trevisani M. and Tuzzi A. (2026). Distinctive Lexical Traits in Andrea Camilleri: A Keyness Measures Analysis of a Corpus of Sellerio Crime Novels. *JADT2026 proceedings*.

- Genova V.G. and Sciandra A. (2026). Camilleri's Sicilian: From Stopwords To Topic Modelling. *JADT2026 proceedings*.
- La Fauci N. (2000), L'italiano perenne e Andrea Camilleri. in Maraschio N. and Poggi Salani T. editors, *Italia linguistica anno mille Italia linguistica anno duemila. Atti del XXXIV Congresso Internazionale di Studi della Società di Linguistica Italiana (Firenze, 19-21 ottobre 2000)*, Bulzoni, Roma, pp. 331-340.
- La Fauci N. (2022). Postilla all'opera di Andrea Camilleri. *Letteratura e dialetti*, 15: 45-50.
- Le S., Josse J. and Husson F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1): 1-18.
- Lebart L., Morineau A. and Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley and sons, New York.
- Matt L. (2020). Lingua e stile della narrativa camilleriana. In: Caocci D., Marci G., Ruggerini M.E. editors, *Quaderni camilleriani 12. Parole, musica (e immagini)*, Grafiche Ghiani, Cagliari, pp. 39-93.
- Mazzarisi P. (2026). Assunti teorico-metodologici su linguaggio letterario esclusivo e dati testuali. *JADT2026 proceedings*.
- Mazzarisi, P. (2025). *Teoria e framework per il Distant Reading: Radici classiche e studi su corpora letterari italiani*. EUT Edizioni Università di Trieste, Trieste.
- Nigro S.S. (2004). *Le "Croniche" di uno scrittore maltese (saggio introduttivo)*. In: Camilleri A., *Romanzi storici e civili*, Arnoldo Mondadori, Milano, pp. XXV-XXVI.
- Reinert M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par context. *Les cahiers de l'analyse des données*, 8(2): 187-198.
- Reinert M. (1990). Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. *Les cahiers de l'analyse des données*, 15(1): 21-36
- Sciandra A. and Cortelazzo M.A. (2026). Classification of Andrea Camilleri's works using bottom-up and top-down approaches. *JADT2026 proceedings*.
- Sciandra A. and Tuzzi A. (2024). Can Correspondence Analysis Challenge Transformers in Authorship Attribution Tasks? In: Plaia A., Egidi L., Abbruzzo A. editors, *Proceedings of the Statistics and Data Science 2024 Conference – New perspectives on Statistics and Data Science*, Palermo University Press, Palermo, pp. 362-367.
- Sciandra A. and Tuzzi A. (2025). Competing Sets of Predictors in an Authorship Attribution Task: Most Frequent Words, Large Language Models and Correspondence Analysis. In: Pollice A., Mariani P. editors, *Methodological and Applied Statistics and Demography IV. SIS 2024. Italian Statistical Society Series on Advances in Statistics*, Springer, Cham, pp. 517-523.
- Sottile R. (2017). Dialetto e dialettalità nella scrittura di Andrea Camilleri. L'incidenza delle parole 'autoctone'. In: Marcato G. editor, *Dialetto. Uno nessuno centomila*, CLEUP, Padova, pp. 329-340.
- Sulis G. (2007). Un esempio controcorrente di plurilinguismo. Il caso Camilleri. In: Berger C., Capra A. and Nimis J. editors, *Les enjeux du plurilinguisme dans la littérature italienne*, Cirillis, Toulouse, pp. 213-230.
- Tundo C. (2024). *Andrea Camilleri e una lingua di cose. Lettura linguistica, lessicale e testuale dei primi romanzi di Montalbano*. Cesati, Firenze.
- Tuzzi A. (2024). *Fondamenti di analisi dei dati testuali*. Carocci, Roma.

C11

Science, technology and knowledge networks

09-lug
12:00 - 13:30
Aula Multimediale C

Language as a Predictor of Scholarly Citation in European Policy Documents

Francis Lareau¹²³, Jérémie Dion¹³, Louis Renaud-Desjardins²³, Malvina Barra²,
Luisa Ripoll-Alberola⁴, François Claveau¹³

¹Université de Sherbrooke – Francis.Lareau@usherbrooke.ca

²Université du Québec à Montréal

³Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST)

⁴Computational Humanities, Leipzig University

Abstract

This paper investigates whether the language of a policy document predicts its scholarly citation behaviour across a large multilingual corpus of European grey literature (N = 2,884,260), covering eight countries, eleven language varieties, and the period 2009–2024. Using logistic regression and Zero-Inflated Negative Binomial (ZINB) models for univariate and language-by-country analyses, and Binomial and Negative Binomial GLMs for a full multivariate analysis controlling for country, source type, topic, and year, we show that English-language documents have approximately twice the grand-mean odds of containing at least one scholarly citation (OR = 2.06) and accumulate citations at more than 2.5 times the average rate (IRR = 2.58). These two dimensions are partially independent: language-country combinations that rank highly on citation probability do not consistently rank highly on citation intensity. Both effects are strongly conditioned by national institutional context: English is a powerful citation gateway in Germany, the Netherlands, and Austria, but falls below the grand mean in Spain and Portugal. The multivariate models, including language–country interactions, reveal that language independently predicts both dimensions of citation behaviour (binary: OR = 1.890; count: IRR = 1.828), while institutional source type — above all IGO affiliation (IRR = 11.090) — remains the dominant driver of citation intensity.

Keywords: citation analysis, English language advantage, policy documents, Zero-Inflated Negative Binomial, logistic regression, multilingual corpus, predictive modelling, Overton database.

1. Introduction

Science-policy communication has become a defining feature of modern governance: from pandemic response to climate legislation, policy actors increasingly invoke scholarly evidence to legitimise their decisions. Yet the conditions under which policy documents actually cite academic work remain poorly understood, particularly outside English-language contexts. While the relationship between publication language and scholarly visibility has been studied extensively within academia (Ammon, 2007; Di Bitetti and Ferreras, 2017), whether language shapes science citation practices in grey literature — government reports, NGO publications, think tank outputs, legislative texts — remains an open question with implications for both the sociology of knowledge (Jasanoff, 2004) and science-policy communications (Pielke, 2007).

The evidence points to two intertwined factors. First, citation intensity varies with institutional infrastructure: policy documents in countries with stronger research funding and higher HDI tend to cite academic work more extensively (Torres-Salinas et al., 2025; Isett and Hicks, 2020). Second, differences across language groups may partly reflect the national environments in which those languages are embedded — making country and institutional context potential

confounders in any language-based analysis. That said, citation-based measures remain imperfect proxies for science uptake, since formal references account for only 72% of the ways scholarly work is invoked in policy documents, with the remainder appearing as footnotes, appendices, or uncontextualised hyperlinks (Yu et al., 2023). This paper addresses this confounding problem directly. Drawing on a corpus of 2,884,260 policy documents retrieved from Overton (Szomszor and Adie, 2022) — a specialised grey literature indexing platform tracking citation links between policy documents and academic publications — and covering eight European countries, eleven language varieties, and the period 2009–2024, we ask: *does the language of a policy document predict its citation behaviour, net of cultural and institutional confounds?* The limitation identified above defines the boundary of our claims: we model formal citation, not science uptake in its full sense.

Descriptive statistics are confined to corpus characterisation (Section 2); methodology (Section 3) and results (Section 4) rely entirely on predictive models estimating the effect of language on citation behaviour while controlling for key structural confounds. Section 5 discusses language as a structural barrier — or enabler — of science-policy knowledge transfer.

2. Corpus

The corpus comprises all policy documents published in the eight most prolific European countries, systematically retrieved from the Overton database. After deduplication, removal of multi-PDF documents, and exclusion of scholarly publications and rare-language documents (using a <2% threshold applied within each country individually) — the final cleaned corpus comprised $N = 2,884,260$ documents. Below is some data Overton provided:

Language: Spanish (27%), English (18%), German (16%), French (9%), Swedish (9%), Dutch (5%), Portuguese (5%), Catalan (4%), Italian (4%), Galician (2%), and Basque (1%).

Country: Spain (41%), Germany (14%), France (12%), Portugal (7%), Netherlands (6%), Italy (6%), and Austria (4%). Local language (non-English) represents 82.1% ($N = 2,368,789$).

Source type: It is mainly Government (84%), followed by Legislative Bodies (12%), Think Tanks (4%), NGOs (0.3%), Judicial bodies (0.1%), and IGOs (0.1%).

Topic: Overton's topic variable was recoded into ten broader categories — Economy (23%), Justice (17%), Culture (10%), Politics (8%), Science & Technology (6%), Education (6%), Society (4%), Health (3%), Environment (3%), and no label (Unknown, 20%).

Year: Coverage expanded substantially over time, with document counts following a strong positive linear trend between 2009 (1.8% of the corpus) and 2024 (11.4%), increasing by an estimated 20,860 documents per year on average (pseudo- $R^2 = 0.963$, LLR $p < .001$).

Citation: Links to scholarly work yielded 3,077,173 total citations across 179,793 citing documents (6.23% of the corpus). In accordance with Overton's documented limitations, citations without a resolvable DOI and non-standard reference formats may be underrepresented, particularly in local-language documents (Overton, 2025). The mean citation count across all documents is 1.07; among citing documents only, it rises to 17.12, reflecting a highly skewed distribution. English-language documents have a citation probability of 15.6% (vs. 4.2% for local languages) and a mean citation count of 3.64 (vs. 0.51). Among the citing documents, the mean number of citations is 23.25 for English versus 12.13 for others.

3. Methodology

3.1. Dependent Variables and Analytical Strategy

Two dependent variables operationalize citation behaviour: a binary indicator (citing/not citing) and a raw citation count, each modelled separately across three progressive stages — i) univariate language comparison (English vs. other languages), ii) language-by-country interaction yielding 19 language-country groups, and iii) a full multivariate model with language, country, institutional source type, topic, and year as simultaneous covariates. All categorical predictors use sum-to-zero (effect) coding, so the intercept estimates the grand mean and each coefficient represents a group's deviation from it.

3.2. Binary Models

For the *binary analysis*, all three stages use the same logistic model:

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \sum_{k=1}^K \beta_k x_{ik} \quad (1)$$

where Y_i is the binary citation indicator for the document i , β_0 is the grand-mean log-odds under effect coding, and β_k are the effect-coded deviations for the predictor k . Exponentiated coefficients yield odds ratios (ORs) relative to the grand mean.

The univariate and language-country stages estimate equation (1) by Maximum Likelihood (MLE); fit is assessed via McFadden's pseudo- R^2 and a likelihood-ratio test against an intercept-only null. The multivariate stage uses Iteratively Reweighted Least Squares (IRLS) with HC0 robust covariance to guard against heteroscedasticity across the large heterogeneous corpus; fit is reported as Cox-Snell pseudo- R^2 alongside a likelihood-ratio test, since log-likelihoods are not directly comparable across the two estimators.

3.3. Count Models

For the *count analysis*, the distribution presents two compounding challenges: 93.77% of documents have zero citations, and a Poisson dispersion statistic exceeding 6 in all subanalyses confirms severe overdispersion, ruling out standard Poisson regression.

For the univariate and language-country stages, we apply a Zero-Inflated Negative Binomial (ZINB) model, which separates structural zeros (documents that will never cite regardless of attributes) from sampling zeros via a two-component mixture:

$$P(Y_i = y) = \pi_i \cdot 1_{[0]} + (1 - \pi_i) \cdot f_{NB}(y; \mu_i, \alpha) \quad (2)$$

where π_i is the probability that the document i belongs to a structural-zero state (a document that will never cite regardless of its attributes), f_{NB} is the negative binomial probability mass function with conditional mean μ_i and dispersion parameter α , and $1_{[0]}$ is the indicator function for zero counts. The inflation component is modelled with an intercept-only logistic equation; the count component follows the same effect-coded linear predictor as equation (1), with a log link. Exponentiated count-component coefficients yield Incidence-Rate Ratios (IRRs) relative to the grand mean. Model fit is validated by comparing the observed to ZINB-predicted zero proportions; a negligible difference confirms correct specification.

For the multivariate stage, the ZINB specification failed to converge due to non-identification of the inflation component under high-dimensional effect coding with sparse cells (Long, 1997). We substitute a Negative Binomial GLM with a log link estimated by IRLS with HC0-robust covariance — fully consistent with the count component of equation (2), but treating all

zeros as sampling zeros. IRRs reflect each covariate's effect on the expected citation count, including non-citers. Fit is reported as Cox-Snell pseudo- R^2 alongside a likelihood-ratio test.

4. Results

4.1. Language Analysis (English vs. other languages)

A chi-square test confirms a significant association between language and citation status ($\chi^2 = 95,030.03$, $p < 0.001$; Cramér's $V = 0.182$). The logistic regression (pseudo- $R^2 = 0.056$) and the ZINB model (pseudo- $R^2 = 0.034$; both LLR $p < 0.001$) establish a strong English citation advantage across both dimensions, establishing a grand-mean citation odds of $OR = 0.09$ (95% CI [0.0896, 0.0905]) and a grand-mean citation rate of $IRR = 6.6$ (95% CI [6.528, 6.674]).

English-language documents deviate strongly above both grand means: their binary citation probability is approximately twice the average ($OR = 2.06$, 95% CI [2.050, 2.07]) and their citation rate is more than 2.5 times the average ($IRR = 2.576$, 95% CI [2.560, 2.593]). Non-English documents mirror these deviations symmetrically: their citation probability is roughly half the grand mean ($OR = 0.485$, 95% CI [0.483, 0.488]), and their citation rate is only 39% of it ($IRR = 0.388$, 95% CI [0.386, 0.391]), placing them approximately 6.6 times below English documents in citation accumulation. The extremely narrow confidence intervals across both models confirm the statistical robustness of these effects at this corpus scale.

The ZINB specification is further validated by the near-perfect reproduction of the observed zero proportion: the model predicts 93.78% zeros, compared with an observed 93.77% ($\Delta = -0.000095$), confirming that the two-component structure correctly captures the generative process. The highly significant inflation intercept ($\beta = 1.330$, $p < 0.001$) indicates that a substantial share of non-citing documents belongs to a structural-zero state — documents that are systematically unlikely to cite regardless of their language.

4.2. Language-by-Country Analysis

Grouping languages by country produces 19 language-country groups. The logistic regression (pseudo- $R^2 = 0.148$; LLR $p < 0.001$) and the ZINB model (pseudo- $R^2 = 0.084$; LLR $p < 0.001$) reveal substantial cross-national heterogeneity, with a grand-mean citation $OR = 0.064$ (95% CI [0.063, 0.064]) and a citation rate $IRR = 1.367$ (95% CI [1.348, 1.387]). The ZINB zero-inflation is again validated (observed: 93.766%; predicted: 93.773%; $\Delta = -0.000062$). Critically, the binary and count model (Table 1) do not consistently rank groups, revealing a dissociation between the probability of citing and the intensity of citation once citing occurs.

English-language groups: English-Germany shows the strongest positive deviation in both dimensions ($OR = 16.953$; $IRR = 22.096$), making it the dominant group across both models. English-Austria also ranks highly in both ($OR = 4.763$; $IRR = 4.389$). Other English-language groups show notable reversals between the two dimensions. English-Netherlands ranks second in the binary model ($OR = 8.032$) but shows the most extreme negative deviation in the count model ($IRR = 0.04$), indicating that English-language Dutch documents cite frequently but with very low per-document counts. English-France ($OR = 5.168$; $IRR = 0.951$) and English-Sweden ($OR = 3.963$; $IRR = 0.639$) both show strong binary advantages that do not translate into above-average citation intensity. Conversely, English-Spain falls sharply below the grand mean in the binary model ($OR = 0.327$) yet ranks second in citation intensity ($IRR = 7.822$), suggesting that the rare English-language documents produced in Spain are disproportionately citation-dense. English-Portugal shows a similar low-binary, high-intensity trend ($OR = 0.296$; $IRR = 1.240$). English-Italy is approaching the grand-mean in both ($OR = 1.217$; $IRR = 0.759$).

National-language groups: Most fall below the grand mean in both models. Two exceptions show below-average citation probability but substantially above-average intensity — Dutch-Netherlands (OR = 0.968; IRR = 5.116) and Catalan-Spain (OR = 0.649; IRR = 4.101) — suggesting that when these documents cite, they do so extensively. Swedish-Sweden is the only group above the mean on both dimensions (OR = 1.235; IRR = 1.203). All remaining groups are below the grand mean in both models, with Galician-Spain showing the most extreme negative deviation (OR = 0.04; IRR = 0.295), followed by Italian-Italy (OR = 0.454; IRR = 0.326), Portuguese-Portugal (OR = 0.493; IRR = 0.481), German-Austria (OR = 0.493; IRR = 0.486), French-France (OR = 0.775; IRR = 0.542), Basque-Spain (OR = 0.552; IRR = 0.511), Spanish-Spain (OR = 0.517; IRR = 0.828), and German-Germany (OR = 0.848; IRR = 0.448).

Those results show that language and country interact in ways reducible to neither language or country alone — English-Germany versus English-Spain makes that asymmetry concrete — and that citation probability and intensity are partially independent dimensions, each structured differently by language-country combinations.

Table 1. Effect-coded language-by-country model results: (A) logistic regression odds ratios on binary citation; (B) ZINB incidence-rate ratios on citation count. Intercept = grand mean; all other entries are group-level deviations. 95% CIs reported. Cell shading: red = above grand mean; blue = below; intensity scales with magnitude.

	(A) Binary Model			(B) Count Model		
	OR	CI_low	CI_high	IRR	CI_low	CI_high
<i>Intercept</i>	0.064	0.063	0.064	1.367	1.348	1.387
<i>English-Germany</i>	16.953	16.646	17.267	22.096	21.732	22.465
<i>English-Netherlands</i>	8.032	7.791	8.280	0.040	0.034	0.046
<i>English-France</i>	5.168	5.073	5.264	0.951	0.929	0.973
<i>English-Austria</i>	4.763	4.608	4.923	4.389	4.250	4.533
<i>English-Sweden</i>	3.963	3.861	4.067	0.639	0.621	0.658
<i>English-Italy</i>	1.217	1.177	1.259	0.759	0.744	0.775
<i>English-Spain</i>	0.327	0.317	0.338	7.822	7.593	8.059
<i>English-Portugal</i>	0.296	0.279	0.314	1.240	1.199	1.282
<i>Swedish-Sweden (inf.)</i>	1.235	1.213	1.257	1.203	1.182	1.225
<i>Dutch-Netherlands</i>	0.968	0.946	0.991	5.116	5.023	5.211
<i>German-Germany</i>	0.848	0.833	0.863	0.448	0.432	0.464
<i>French-France</i>	0.775	0.759	0.791	0.542	0.515	0.570
<i>Catalan-Spain</i>	0.649	0.630	0.668	4.101	3.996	4.208
<i>Basque-Spain</i>	0.552	0.525	0.582	0.511	0.503	0.519
<i>Spanish-Spain</i>	0.517	0.509	0.526	0.828	0.813	0.843
<i>German-Austria</i>	0.493	0.476	0.511	0.486	0.471	0.501
<i>Portuguese-Portugal</i>	0.493	0.477	0.509	0.481	0.465	0.499
<i>Italian-Italy</i>	0.454	0.439	0.471	0.326	0.316	0.337
<i>Galician-Spain</i>	0.040	0.034	0.046	0.295	0.278	0.313

4.3. Multivariable Analysis: Language, Country, Source Type, Topic, and Year

To isolate the independent contribution of each covariate, a Binomial GLM (binary dependent variable, pseudo- $R^2 = 0.116$; LLR $p < 0.001$) and a Negative Binomial GLM (count dependent variable, pseudo- $R^2 = 0.022$; LLR $p < 0.001$) are each estimated with the effects of language, country, source type, topic, year and language-country interaction as predictors, all entered jointly and coded by effects. The low count-value pseudo- R^2 is expected for population-scale citation data, where structural predictors capture systematic tendencies but residual document-

level variation — author practice, genre, sub-institutional norms — dominates. The observed effects are therefore best interpreted as robust directional estimates rather than measures of predictive precision. University source type was excluded due to insufficient representation ($N = 24$). The grand-mean citation probability across all groups is $OR = 0.193$ in the binary model (Table 2a), and the grand-mean citation rate is $IRR = 1.513$ in the count model (Table 2b).

Table 2. Effect-coded multivariate model results ($N = 2,884,234$): (A) binomial GLM odds ratios; (B) negative binomial GLM incidence-rate ratios. Predictors: language, country, source type, topic, year, and language–country interactions, entered simultaneously. Intercept = grand mean; inferred categories (inf.) derived by sum-to-zero constraint; year is continuous (per-year multiplicative change). 95% CIs report. Cell shading: red = above grand mean; blue = below; intensity scales with magnitude.

	A) Binary Model			B) Count Model		
	OR	CI_lower	CI_upper	IRR	CI_lower	CI_upper
Intercept	0.193	0.186	0.200	1.513	1.411	1.623
Language:English	1.890	1.870	1.910	1.828	1.768	1.891
Language:Other (inf.)	0.630	0.617	0.643	0.524	0.489	0.562
Country:Austria	1.083	1.050	1.116	0.876	0.804	0.954
Country:France	0.597	0.587	0.607	0.790	0.704	0.887
Country:Germany	1.645	1.621	1.670	1.345	1.291	1.401
Country:Italy	0.703	0.687	0.721	0.791	0.739	0.846
Country:Netherlands	1.619	1.583	1.656	1.989	1.889	2.094
Country:Portugal	0.671	0.649	0.693	0.421	0.384	0.460
Country:Spain	0.610	0.599	0.621	0.866	0.808	0.927
Country:Sweden (inf.)	2.017	1.981	2.054	1.877	1.790	1.967
Type:Government	0.351	0.339	0.364	0.420	0.384	0.460
Type:IGO	7.015	6.385	7.708	11.090	9.217	13.345
Type:Judicial Body	0.432	0.376	0.496	0.132	0.104	0.168
Type:Legislative Body	0.281	0.269	0.294	0.269	0.243	0.298
Type:NGO	1.143	1.072	1.218	2.233	1.977	2.522
Type:Think Tank (inf.)	2.925	2.819	3.035	2.696	2.504	2.902
Topic:Culture	0.803	0.787	0.819	0.687	0.637	0.741
Topic:Economy	1.777	1.758	1.796	1.831	1.762	1.903
Topic:Education	1.255	1.229	1.281	1.291	1.194	1.397
Topic:Environment	1.766	1.722	1.812	1.682	1.557	1.817
Topic:Health	3.777	3.704	3.851	6.045	5.716	6.393
Topic:Justice	0.557	0.547	0.567	0.305	0.287	0.323
Topic:Politics	0.520	0.508	0.532	0.412	0.380	0.445
Topic:Science & Tech.	2.080	2.044	2.115	2.661	2.509	2.823
Topic:Society	0.882	0.861	0.905	0.723	0.658	0.794
Topic:Unknown (inf.)	0.158	0.153	0.163	0.251	0.232	0.271
Year	0.969	0.967	0.970	0.961	0.955	0.966
Interact:Eng-Austria	1.030	1.003	1.058	1.169	1.081	1.264
Interact:Eng-France	0.687	0.677	0.698	0.817	0.729	0.914
Interact:Eng-Germ.	1.806	1.782	1.831	1.531	1.477	1.588
Interact:Eng-Italy	1.086	1.061	1.112	1.256	1.173	1.345
Interact:Eng-Netherl.	0.989	0.968	1.012	1.308	1.241	1.378
Interact:Eng-Portugal	0.909	0.881	0.938	0.820	0.749	0.897
Interact:Eng-Spain	0.672	0.661	0.684	0.530	0.500	0.562
Interact:Eng-Sw. (inf.)	1.190	1.171	1.211	0.959	0.916	1.003

Language: After controlling for all other predictors, English remains positively associated with both citation probability (OR = 1.89) and intensity (IRR = 1.828), confirming language as an independent gatekeeping variable at both dimensions. The inferred Other-language coefficient mirrors these patterns symmetrically (OR = 0.63; IRR = 0.524).

Language and country: In this interaction model, the language and country main effects quantify deviations from the grand mean averaged across all groups, while the language–country interaction terms capture country-specific amplifications or suppressions of the English effect. Country-level main effects elaborate the English advantage without displacing it: Sweden (OR = 2.017; IRR = 1.877), the Netherlands (OR = 1.619; IRR = 1.989), and Germany (OR = 1.645; IRR = 1.345) show the highest average citation propensity, while France (OR = 0.597; IRR = 0.790), Italy (OR = 0.703; IRR = 0.791), Spain (OR = 0.61; IRR = 0.866), and Portugal (OR = 0.671; IRR = 0.421) all fall below the grand mean in both models.

The interaction terms reveal that the English main effect is further amplified in Germany (OR = 1.806; IRR = 1.531), Austria (OR = 1.030; IRR = 1.169), and Italy (OR = 1.086; IRR = 1.256), and suppressed in Spain (OR = 0.672; IRR = 0.530), Portugal (OR = 0.909; IRR = 0.820), and France (OR = 0.687; IRR = 0.817). The Netherlands shows a near-neutral binary interaction (OR = 0.989) but a positive count interaction (IRR = 1.308).

Conversely, Sweden's inferred interaction is positive on the binary dimension (OR = 1.190) but negative on the count dimension (IRR = 0.959). These patterns are consistent with the saturated language-country effects observed in Section 4.2, confirming that country moderation of the English advantage is robust to controlling for source type, topic, and year.

Source type shows the largest deviations from the grand mean in both models. IGOs exhibit by far the strongest association with citation behaviour (OR = 7.015; IRR = 11.090), followed by Think Tanks (OR = 2.925; IRR = 2.696) and NGOs (OR = 1.143; IRR = 2.233). Government and Legislative Bodies fall consistently below the mean (OR = 0.351 and 0.281; IRR = 0.420 and 0.269), while Judicial Bodies show the lowest citation intensity (OR = 0.432; IRR = 0.132).

Topic effects are stable relative to the additive model. Health is the most citation-intensive topic (binary: OR = 3.777; count: IRR = 6.045), followed by Science & Technology (OR = 2.080; IRR = 2.661), Economy (OR = 1.777; IRR = 1.831), and Environment (OR = 1.766; IRR = 1.682). Politics and Justice remain the least citation-intensive domains (Politics: OR = 0.520; IRR = 0.412; Justice: OR = 0.557; IRR = 0.305).

Year is negative and significant in both models (OR = 0.969; count: IRR = 0.961), indicating a slight decline in citation probability and intensity over 2009–2024 after controlling for language, country, institutional type, and topic. A disaggregated check using year $\times$ source type interactions reveals that the grand-mean year effect becomes non-significant once source type is controlled (coef = -0.003 , $p = 0.717$): the negative trend is concentrated in Government documents (interaction coef = -0.059), while IGOs and Legislative Bodies show flat or increasing citation rates over time. This heterogeneity is consistent with a compositional coverage-expansion artifact rather than a genuine secular decline in evidence mobilization.

5. Discussion

Across three analytical stages, a consistent picture emerges: language and national institutional context jointly structure scholarly citation behaviour in European grey literature, but they do so through different mechanisms and at different levels of analysis. At the aggregate level, English-language documents are approximately 4.25 times more likely to cite scholarly work and accumulate citations at 6.6 times the rate of non-English documents. Di Bitetti and Ferreras

(2017) document a comparable directional advantage for English in peer-reviewed ecology journals — higher citation probability and count, net of journal, year, and paper length — though that work measures citations received by research outputs, whereas we measure citations given by policy actors. Our univariate estimates (OR = 2.060; IRR = 2.576) are consistent in direction and plausible in magnitude relative to their findings; the multivariate estimates (OR = 1.890; IRR = 1.828) represent a more conservative lower bound, net of institutional and national context. Both phenomena may share structural drivers — institutional prestige, international network access, database coverage biases (Ammon, 2007) — but they reflect different positions in the science-policy knowledge chain.

The language-by-country analysis reveals that the English advantage is not universal: it is powerful in Germany, the Netherlands, and Austria, but reverses in Spain and Portugal, where English-language documents fall well below the grand mean. This falsifies an access explanation — if language were the barrier, English documents would outperform everywhere. The binary-count dissociations deepen this point: English-Netherlands documents are among the most likely to cite but produce the lowest per-document counts (OR = 8.032; IRR = 0.04); English-Spain reverses (OR = 0.327; IRR = 7.822). Citation probability and citation intensity are partially independent outcomes, each shaped by different language-country configurations.

The multivariate model, incorporating language-country interactions alongside source type, topic, and year, resolves this complexity into a hierarchy of effects. Language retains strong independent associations with both citation probability (OR = 1.890) and intensity (IRR = 1.828) after controlling for all covariates — it is neither reducible to institutional context nor absorbed by it. Country modulates the English effect: Germany amplifies it (interaction OR = 1.806; IRR = 1.531), while Spain and Portugal suppress it (OR = 0.672; IRR = 0.530). But once language-country dynamics are accounted for, institutional source type dominates citation intensity: IGOs (IRR = 11.090) and Think Tanks (IRR = 2.696) produce structurally citation-dense documents regardless of language or country, while Government and Legislative Bodies remain near-absent from the citation-practising population. Topic follows the same pattern: Health (IRR = 6.045) and Science & Technology (IRR = 2.661) documents are citation-intensive across all national contexts. The slight negative year trend (OR = 0.969; IRR = 0.961) warrants caution; it may reflect Overton's expanding coverage as much as real temporal change.

A methodological note deserves emphasis. The near-perfect ZINB zero-prediction ($\Delta < 0.0001$ at both stages) confirms that the 93.77% zero rate reflects a structural division: most policy documents belong to a citation-free population regardless of their attributes. The convergence failure in the multivariate specification — due to non-identification of the inflation component under high-dimensional effect coding — is itself informative: the structural-zero process cannot be stably separated from sampling zeros once the full covariate space is introduced. Researchers in sparse corpora should design covariate structures with this identifiability boundary in mind.

6. Conclusion

Across 2,884,260 European policy documents spanning eight countries, eleven language varieties, and sixteen years, the language of publication is a statistically robust and functionally specific predictor of scholarly citation behaviour in European grey literature. The English citation advantage established in Section 4.1 persists across all model specifications and survives control for country, source type, topic, year, and language-country interactions — but its strength varies dramatically by national context: amplified in Germany and the Netherlands, reversed in Spain and Portugal. Citation probability and citation intensity are partially independent dimensions, each governed by a distinct configuration of factors.

The key implication is structural. Language is not the sole — or even the primary — lever for science-policy knowledge mobilization. The depth of scholarly engagement is governed by institutional mandate and thematic domain first, and by language second. Interventions targeting language barriers alone — translation programmes, multilingual submission policies — risk misdiagnosing the problem if any. Increasing knowledge mobilization in local-language contexts may require investment in the research-producing institutions and evidence-oriented policy cultures that generate citation-intensive documents, not only in the languages they use.

Future work should address three limitations. First, Overton's cross-nationally heterogeneous indexing coverage means that observed country differences may partly reflect what is captured rather than what is produced — a fundamental constraint on causal inference that no modelling strategy can fully resolve. Second, finer-grained topic classification, potentially from automated text analysis or topic modelling, would clarify whether the Health and Science & Technology advantages reflect genuine disciplinary citation norms or topic-assignment artefacts. Third, and most fundamentally, the observational design cannot establish whether language causally drives citation behaviour or proxies for the institutional characteristics of actors who self-select into English-language production. Quasi-experimental identification — difference-in-differences on institutional language policy shifts (Bleakley and Chin, 2004) or instrumental variables using national proficiency indices (Ku and Zussman, 2010) — remains the necessary next step toward understanding whether removing language barriers would meaningfully change knowledge mobilization, or whether the disparities are better addressed at the level of the institutions that produce citation-intensive policy knowledge in the first place.

Acknowledgement

F.L. acknowledges funding from SSHRC (756-2024-0557) and an Overton Micro-grant.

References

- Ammon, U. (2007) “Global Scientific Communication: Open Questions and Policy Suggestions,” *AILA Review*, 20, pp. 123–133.
- Bleakley, H. and Chin, A. (2004) “Language Skills and Earnings: Evidence from Childhood Immigrants,” *The Review of Economics and Statistics*, 86(2), pp. 481–496.
- Di Bitetti, M.S. and Ferreras, J.A. (2017) “Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications,” *Ambio*, 46(1), pp. 121–127.
- Isett, K. R. and Hicks, D. (2020) “Pathways From Research Into Public Decision Making: Intermediaries as the Third Community,” *Perspectives on Public Management and Governance*, 3, pp. 45–58.
- Jasanoff, S. (2004) *States of Knowledge: The Co-Production of Science and the Social Order*. Routledge.
- Ku, H. and Zussman, A. (2010) “Lingua franca: The role of English in international trade,” *Journal of Economic Behavior & Organization*, 75(2), pp. 250–260.
- Long, J.S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Sage.
- Overton (2025) “What is a policy citation?” <https://www.overton.io/blog/what-is-a-policy-citation>
- Pielke, R. A. Jr. (2007) *The Honest Broker: Making Sense of Science in Policy and Politics*. Cambridge: Cambridge University Press.
- Szomszor, M. and Adie, E. (2022) “Overton: A bibliometric database of policy document citations.” *Quantitative Science Studies*, 3(3), pp. 624–650.
- Torres-Salinas, D., Arroyo-Machado, W., Robinson-García, N. and Romero-Frías, E. (2025) “From Science to Policy: A Multilevel Framework for Measuring Institutional Policy Influence Through Overton.” [Preprint]. <https://digibug.ugr.es/handle/10481/104538>.

10 Lareau, Dion, Renaud-Desjardins, Barra, Ripoll-Alberola & Claveau

Yu, H., Murat, B., Li, J. and Li, L. (2023) “How can policy document mentions to scholarly papers be interpreted? An analysis of the underlying mentioning process,” *Scientometrics*, 128, pp. 6247–6266.

Mapping the languages of science: a comparative cross-media textual analysis of scientific expertise

Rosanna Cataldo¹, Ferdinando Iazzetta², Gabriella Punziano³, Barbara Saracino⁴

¹University of Naples Federico II – rosanna.cataldo2@unina.it

²University of Naples Federico II – ferdinando.iazzetta@unina.it

³University of Naples Federico II – gabriella.punziano@unina.it

⁴University of Bologna – barbara.saracino@unibo.it

Abstract

This study examines how scientific authority circulated across television, newspapers, and social media during the Italian vaccination campaign. Using a corpus of 26,933 texts, it combines computational text analysis and semantic modeling to explore cross-media discursive configurations. Findings show that newspapers emphasize evidence-based information, television institutional crisis management, and social media controversy and risk negotiation, despite substantial thematic overlap across platforms. The study highlights how scientific authority is discursively constructed within mediatized public communication.

Keywords: Computational Text Analysis, Scientific Expertise, Cross-Media Discourse, Science Communication.

1. Introduction

The COVID-19 pandemic intensified the public circulation of scientific knowledge, transforming scientific experts into central actors in media and political debate. During the Italian vaccination campaign, virologists, epidemiologists, physicians, and institutional experts occupied highly visible positions across television, newspapers, and digital platforms, contributing to the public interpretation of the health crisis and to the legitimization of political and health-related decisions. In our previous book, *Mediaticamente esperti. La costruzione pubblica dell'expertise scientifica nella campagna vaccinale anti-COVID-19 in Italia* (Cataldo, Iazzetta, Punziano & Saracino, 2026), we examined how scientific expertise was publicly constructed across television, print journalism, and social media. The present article extends that research by shifting attention from the visibility of experts to the linguistic and discursive configurations through which scientific authority is articulated across media environments. Comparative computational analyses of how scientific authority is constructed across communicative environments remain relatively limited. This study addresses this gap by integrating statistical text analysis, semantic modeling, and discourse-oriented approaches within a unified analytical framework. Drawing on a corpus of more than 26,000 texts collected from television broadcasts, national newspapers, and social media platforms, the article aims to map the discursive configurations through which scientific authority circulated during the Italian vaccination campaign. More specifically, the study addresses two main research questions:

- How do the lexical and semantic configurations of media discourse surrounding scientific expertise vary across television, print media, and social media?
- Which narrative and discursive patterns shape the public construction of scientific authority across different media environments?

2. Scientific Expertise Across Media: Discourse, Mediatization, and Computational Approaches

The public construction of scientific expertise has long represented a central issue within science communication and media studies. From Merton's (1973) classical sociology of science to Collins and Evans' (2002) work on expertise and legitimacy, scientific authority has been understood not merely as technical competence, but as a socially negotiated form of legitimacy emerging at the intersection between epistemic validation within scientific communities and public recognition in broader social arenas. The pandemic also intensified forms of "post-normal" public communication, in which scientific uncertainty, political decision-making, and media visibility became deeply intertwined (Funtowicz & Ravetz, 1993). Under these conditions, scientific expertise became increasingly exposed to processes of politicization, media amplification, and public contestation (Bucchi & Trench, 2021). As shown by Cataldo et al. (2022; 2024; 2026), scientific experts in Italy simultaneously occupied institutional fields—such as universities, research centers, and advisory committees—and media arenas including television studios, newspapers, and digital platforms. This dual positioning reveals a structural tension: while credibility depends on peer validation, public authority increasingly relies on media visibility. Mediatization studies (Hjarvard, 2008; Couldry & Hepp, 2013) have demonstrated how media reshape the conditions under which scientific knowledge becomes publicly legitimate. In highly mediatized societies, visibility itself becomes a form of symbolic capital (Couldry, 2014), and scientific authority is filtered through media formats, communicative constraints, and audience expectations. The transition from the "visible scientist" (Goodell, 1977) to the contemporary "mediagenic expert" reflects this transformation, whereby expertise must adapt to media-specific grammars of communication and public performance. Within this framework, Critical Discourse Analysis (CDA) provides an important theoretical foundation for examining how authority, legitimacy, and knowledge are linguistically constructed and reproduced (Fairclough, 2003; van Dijk, 2008). Linguistic strategies such as modality, hedging, metaphorical framing, and argumentative organization contribute to presenting scientific claims as authoritative, provisional, or contested. Lexical choices—for example, the distinction between expressions such as "data suggest" and "evidence proves"—play a crucial role in shaping epistemic authority. Similarly, metaphors such as "waves" or "war against the virus" contribute to framing the pandemic within emotionally and politically charged interpretative schemas. At the same time, Narrative Media Analysis (Roeh & Feldman, 1984; Benson, 2013) highlights how expertise is embedded within broader storytelling structures shaped by media conventions and communicative logics. Television frequently situates experts within dialogical and confrontational formats characterized by immediacy, personalization, and dramatization; print journalism tends to contextualize expert interventions within more stable argumentative structures; social media environments introduce further fragmentation, where expertise is negotiated through comments, threads, algorithmic amplification, and participatory interaction. While discourse-oriented and narrative approaches have traditionally relied on qualitative close reading, recent developments in computational linguistics and textual statistics make it possible to operationalize discourse analysis at scale. The quantitative analysis of textual corpora has a long tradition within textual statistics (Lebart & Salem, 1994), where techniques such as correspondence analysis, lexical co-occurrence mapping, and clustering have been used to identify structural regularities within large corpora. More recently, text mining and machine learning approaches have considerably expanded these possibilities (Peters, 2014). Embedding models such as Sentence-BERT enable the measurement of semantic similarity beyond simple lexical co-occurrence by capturing contextual meaning within high-dimensional vector spaces.

Likewise, topic modeling approaches such as BERTopic integrate transformer-based embeddings with clustering algorithms to generate semantically coherent thematic configurations. These computational approaches make it possible to map semantic structures, thematic constellations, and discursive regularities across heterogeneous corpora while preserving contextual relationships between texts. Such developments align with broader trends in computational social science and “text as data” approaches, which increasingly combine large-scale textual analysis with interpretative perspectives on public communication (Lazer *et al.*, 2009). During the pandemic, experts moved continuously between television broadcasts, newspaper editorials, and social media interactions, generating hybrid forms of visibility and discursive spillovers across media systems. A comparative cross-media perspective therefore becomes essential for understanding how scientific authority is constructed, negotiated, and adapted to different media logics and communicative contexts. By integrating heterogeneous corpora within a unified computational framework, this study bridges discourse-oriented approaches and computational text analytics, contributing to the analysis of science communication in contemporary mediatized societies. Rather than treating expertise as a self-contained form of communication, this perspective examines how scientific authority is discursively reconfigured across interconnected media arenas.

3. Methodology

3.1. Research Design and Corpus Construction

The study adopts a comparative cross-media research design focused on the Italian COVID-19 vaccination campaign, a period characterized by intense public debate and a high level of visibility of scientific experts. The corpus consists of 26,933 textual units collected between December 2020 and December 2021 and is structured into three distinct yet analytically integrated datasets:

- television transcripts (prime time talk shows, news broadcasts, and institutional briefings);
- national newspaper articles (editorials, interviews, and analytical reports);
- social media posts related to vaccination discourse and scientific expertise.

Data collection for television broadcasts and newspaper articles was conducted through the Volocom media monitoring database, which enabled the systematic retrieval of television transcripts from the seven selected national channels and articles from the three selected newspapers. To ensure consistency and replicability in corpus construction, data extraction was guided by a shared set of keyword-based queries. Boolean operators were employed to identify texts simultaneously referring to the pandemic, vaccination, and scientific expertise. More specifically, the extraction strategy combined terms¹ related to COVID-19 and coronavirus with references to vaccines and experts/scientists. These keyword combinations were designed to systematically identify media content in which references to the pandemic, vaccination, and scientific expertise co-occurred within the same communicative context. Social media data were collected following the same keyword strategy and restricted to publicly accessible posts and comments. The temporal segmentation reflects the principal institutional and communicative phases of the Italian vaccination campaign, from the initial preparation stage to the expansion of the Green Pass policy. The unequal distribution of documents across media environments reflects differences in media production intensity and publication frequency

¹ The main extraction queries included combinations such as: Coronavirus AND Scienziat AND Vaccin*; Covid AND Scienziat* AND Vaccin*; Vaccin* AND Espert*; Vaccin* AND Espert* AND Coronavirus; Vaccin* AND Scienziat**.

during the vaccination campaign. Descriptive statistics of the corpus are reported in Tables 1 and 2. Table 1 presents the distribution of documents across media environments and vaccination phases, while Table 2 summarizes the principal lexical characteristics of the corpus.

Table 1. Corpus distribution by media and phase

Media type	Pre-phase	Phase1	Phase2	Phase3	Phase4	total
Newspaper	601	469	661	589	633	2953
Social Network	831	949	989	735	1270	4774
TV	3287	4612	4949	2926	3432	19206
total	4719	6030	6599	4250	5335	26933

Table 2. Corpus lexical statistics

Metric	Value
Documents	26,933
Total tokens	28,752,959
Vocabulary size (types)	128,070
Type-Token Ratio	0.0045
Hapax legomena	15,608
Hapax ratio	12.2%
Mean document length	1,065 tokens
Median document length	1,354 tokens
Maximum document length	2,002 tokens

3.2. Preprocessing and First Analysis

All analyses were conducted in Python using a structured computational pipeline. Text preprocessing included lowercasing, punctuation removal, tokenization, lemmatization, stopword filtering, and the removal of URLs, hashtags, and metadata. Domain-specific terminology (e.g., immunization, variant, efficacy) was preserved to maintain the semantic specificity of the corpus. Following normalization, texts were transformed into document-term matrices for lexical analysis and vector representations for semantic modeling. Frequency analysis and TF-IDF weighting were first applied to identify salient and media-specific lexical patterns. A Lexical Correspondence Analysis (LCA) was then conducted to explore relationships between vocabulary structures and communicative contexts, with media environments and vaccination phases included as supplementary variables. Cluster analysis on factorial coordinates was subsequently used to identify recurring narrative configurations across the corpus. To complement lexical statistics with contextual semantic analysis, documents were encoded using the multilingual Sentence-BERT model (paraphrase-multilingual-MiniLM-L12-v2). Topic modeling was performed through BERTopic, while semantic clustering combined UMAP dimensionality reduction and KMeans clustering. Representative keywords for each cluster were extracted using class-based TF-IDF. By integrating lexical statistics, correspondence analysis, clustering techniques, and semantic modeling, the study develops a comparative framework for examining how scientific expertise was discursively articulated across media environments.

4. Results

TF-IDF reveals distinct lexical profiles across media environments. In the newspaper, distinctive terms include data, research, science, efficacy, and EMA (European Medicines Agency), pointing to references to empirical evidence and regulatory institutions. Political terms such as government, president, and regions also appear, situating expertise between scientific explanation and institutional governance. The television corpus displays a more explicitly political-institutional configuration. The highest TF-IDF scores are associated with terms such as government, president, regions, measures, ministry, and emergency. These lexical patterns indicate that television discourse tends to frame scientific expertise primarily within the context of political decision-making and crisis management. The social media corpus presents a more heterogeneous lexical profile characterized by references to vaccine-related

controversy and perceived risk. Terms such as anti-vaccine, risk, safety, and children, together with references to specific vaccines such as Pfizer and Moderna, emerge as distinctive. In this environment, expertise appears not only as a source of information but also as an object of debate and contestation. TF-IDF shows differentiated discourses across media: newspapers emphasize evidence, television institutional management, and social media a more polarized vocabulary. Table 3 lists the fifteen most distinctive lexical items.

Table 3 Most distinctive lexical items by media environment (TF-IDF)

Newspapers	TF-IDF	Social Media	TF-IDF	Television	TF-IDF
government	0.314	anti-vaccine	0.288	Government	0.566
Cases	0.302	pfizer	0.156	President	0.377
Data	0.286	Data	0.113	Regions	0.291
countries	0.227	children	0.111	Cases	0.214
Pfizer	0.223	government	0.094	Data	0.208
president	0.219	cases	0.088	Situation	0.198
Risk	0.217	Risk	0.084	Infections	0.181
campaign	0.197	efficacy	0.080	Politics	0.175
regions	0.182	research	0.078	Risk	0.174
infections	0.177	countries	0.076	Campaign	0.166
research	0.172	infections	0.064	Countries	0.153
science	0.167	emergency	0.058	Measures	0.141

The factorial map in Figure 1 provides a geometric representation of the semantic structure of the corpus. The first factorial axis (F1) reflects a primary opposition between scientific–biomedical discourse and political–institutional discourse. The positive side includes biomedical terms such as virus, coronavirus, scientists, patients, covid, and infection, reflecting a scientific and epidemiological framing of the pandemic. The negative pole of F1 is characterized by terms associated with political actors and institutional arenas, such as government, senate, center-right, and references to political leadership and institutional decision-making. The second factorial axis (F2) introduces a further differentiation within the health domain. The upper part of the map is dominated by terms related to epidemiological monitoring and the evolution of the pandemic, including epidemic, variant, monitoring, and public health. In contrast, the lower region includes terms associated with vaccination practices and health policy implementation, such as vaccine, dose, drug, administration, rollout, and authorization. This area reflects vaccination management and policy implementation. Taken together, the two axes reveal a structured semantic space organized around three major discursive domains: the biomedical interpretation of the pandemic, the institutional governance of the crisis, and the operational management of vaccination policies. When media environments are projected as supplementary variables onto the factorial space, a clear differentiation between communicative contexts emerges. National newspapers tend to occupy an intermediate position between the scientific and political poles, suggesting a mediating role between scientific interpretation and institutional decision-making. Social media appear closer to biomedical and vaccination-related vocabulary, reflecting the strong presence of debates about disease, vaccines, and related controversies in online environments. Television channels occupy a relatively central position within the semantic space, consistent with the hybrid and dialogic nature of broadcast discourse, where expert explanations frequently intersect with political discussion and institutional communication.

Figure 1 – Lexical Correspondence Analysis factorial map (F1–F2; total inertia explained = 21.3%)



The cluster analysis performed on the factorial scores derived from the Correspondence Analysis. Cluster 0 (6.84%) reflects the institutional leadership narrative, centered on political actors and the management of the crisis by government authorities (*Draghi, government, president, prime minister*). Cluster 1, which represents the largest portion of the corpus (37.69%), corresponds to an informational-scientific narrative of the pandemic, characterized by the vocabulary of health communication and epidemiological updates (*covid, virus, pandemic, data, experts, scientists*). Within this frame, the pandemic is represented as a health phenomenon to be understood and monitored through scientific expertise. Cluster 2 (10.19%) represents a vaccination mobilization narrative, focusing on the vaccination campaign and immunization strategies (*vaccine, dose, AstraZeneca, campaign, administration*). Cluster 3 (29.72%) can be interpreted as a public health emergency narrative, centered on the evolution of infections and the impact of the crisis on the hospital system (*contagion, curve, deaths, hospitals, intensive care, patients*). This register emphasizes the dramatic and clinical dimensions of the pandemic. Cluster 4 (11.00%) is associated with the territorial regulation narrative of the pandemic, referring to the implementation of containment measures and restrictions (*regions, zones, restrictions, curfew*). Finally, Cluster 5 (4.57%) reflects the political conflict narrative, characterized by references to political actors and partisan dynamics (*Conte, Renzi, opposition, political crisis*). The quantitative distribution of the clusters within the corpus is reported in Table 4.

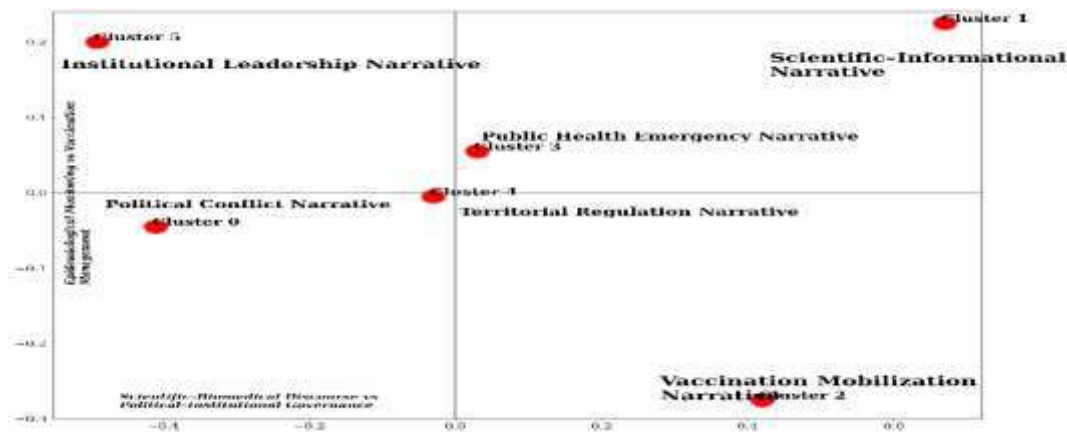
Table 4. Distribution and Interpretation of Narrative Clusters

Cluster	Documents	(%)	Thematic Interpretation
Cluster 0	1,833	6.84	Political leadership and institutional management of the crisis (<i>Draghi, government, president, prime minister, council, Salvini, majority, decree</i>)
Cluster 1	10,105	37.69	General narrative of the pandemic and scientific-health information (<i>covid, virus, pandemic, experts, scientists, physician, data, risk, contagion</i>)
Cluster 2	2,731	10.19	Vaccination campaign and immunization strategies (<i>vaccine, dose, AstraZeneca, administration, campaign, efficacy, production, elderly</i>)
Cluster 3	7,968	29.72	Epidemiological trends and monitoring of the health crisis (<i>contagion, curve, deaths, hospitals, patients, intensive care, data, wave</i>)
Cluster 4	2,949	11.00	Territorial governance and containment measures (<i>region, zone, restrictions, curfew, safety, schools, restaurants, controls</i>)

Cluster 5	1,224	4.57	Political conflict and government crisis dynamics (<i>Conte, Renzi, opposition, center-right, crisis, confidence vote, parliament, elections</i>)

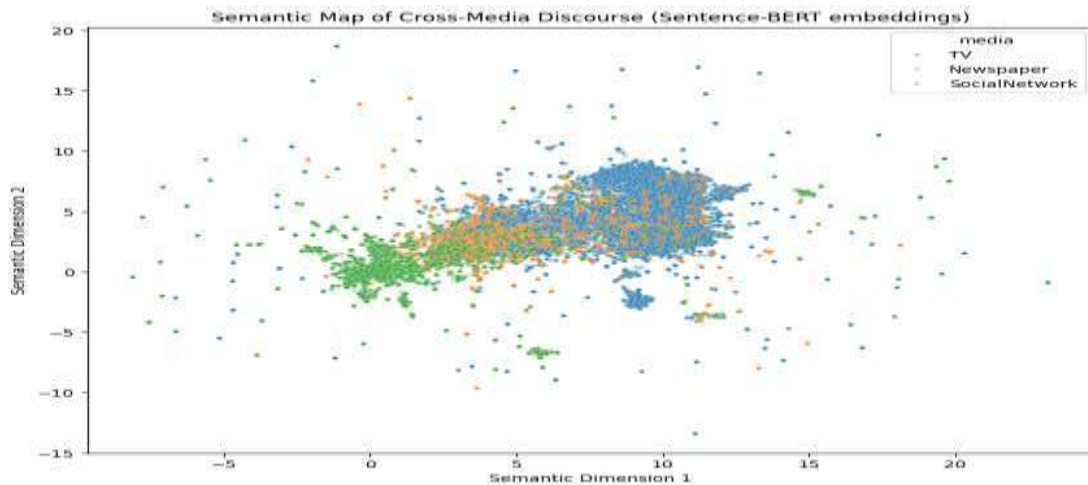
The projection of the clusters onto the factorial space (Figure 2) highlights the discursive structure of the corpus. While media environments privilege specific narrative modes, the overlap between clusters indicates that discursive configurations circulate across contexts rather than forming separate domains. Scientific expertise thus emerges as a dynamic and non-unified discourse, continuously reformulated across media environments.

Figure 2. Projection of narrative clusters in the factorial space (LCA)



The results suggest that the media construction of scientific authority is articulated along three main dimensions. The first concerns the epistemic density of discourse, ranging from highly technical language to more simplified registers. The second relates to narrative modes oscillating between scientific explanation and political controversy. The third involves strategies of legitimation based on institutional, moral, or experiential forms. Within this multidimensional space, different media environments tend to occupy relatively distinct positions. Television often functions as an arena of mediation between institutional expertise, political debate, and public reassurance. Print journalism contributes to consolidating scientific authority through more stable argumentative structures and contextualized interpretations. Social media, by contrast, introduce participatory dynamics that expose expert authority to processes of negotiation, contestation, and collective reinterpretation. To explore the semantic structure of the corpus at a deeper level and to verify the discursive configurations that emerged from the lexical analysis, a contextual sentence-embedding approach was adopted. The semantic map generated using Sentence-BERT represents the distribution of documents within the semantic embedding space. Each point corresponds to a document in the corpus, and colours indicate the media type (television, newspapers, and social media). Figure 3 reveals a substantial area of overlap among the three media environments, suggesting that a significant portion of the discourse on the pandemic is shared across television, newspapers, and social media. At the same time, some differences in the dispersion of points can be observed: documents originating from social media appear more widely dispersed across the semantic space, whereas television and newspaper texts tend to cluster more strongly within specific regions of the map. Overall, the semantic representation confirms the patterns identified in the previous lexical analyses: while different media environments share a common semantic foundation, they articulate and organize public debate on the pandemic through distinct communicative configurations.

Figure 3. Semantic Density Map of Cross-Media Discourse Based on Sentence-BERT Embeddings



A topic modeling analysis was subsequently conducted using the BERTopic algorithm to identify the main thematic configurations present in the corpus. The analysis identified eight main macro-topics, representing the principal semantic areas around which media discourse on the pandemic is organized. The distribution of documents across thematic clusters reveals a significant concentration within the first four topics, which account for most of the texts in the corpus (Table 5). In particular, the three largest clusters contain 6,563, 5,877, and 5,678 documents, respectively, indicating that a substantial portion of the media debate is dominated by content related to the political–institutional management of the pandemic. Alongside these major clusters, additional thematic configurations of smaller size but greater semantic specificity also emerge. Some topics are associated with discourse on the vaccination campaign, the scientific communication of experts, and discussions related to public health measures and emergency management. These clusters reflect the presence within the corpus of narratives related to the spread of the virus, population vaccination, and the public communication of science. Overall, the results show that media discourse on the pandemic is organized around two main discursive domains. The first concerns the political governance of the health crisis, characterized by references to institutions, governmental decisions, and political debates over emergency management. The second domain relates to the medical–scientific dimension of the pandemic, including discussions about vaccines, scientific research, the role of experts, and public health strategies. This thematic articulation is consistent with the findings of the previous lexical analyses. In particular, the LCA had already highlighted a similar opposition between a political–institutional discursive register and a medical–scientific register. The convergence between results obtained through lexical statistics and those derived from semantic modeling therefore reinforces the interpretation that media discourse on the pandemic is structured around these two principal discursive dimensions. Table 5 presents a summary of the main topics identified by the model, indicating for each cluster the number of associated documents and the most representative keywords.

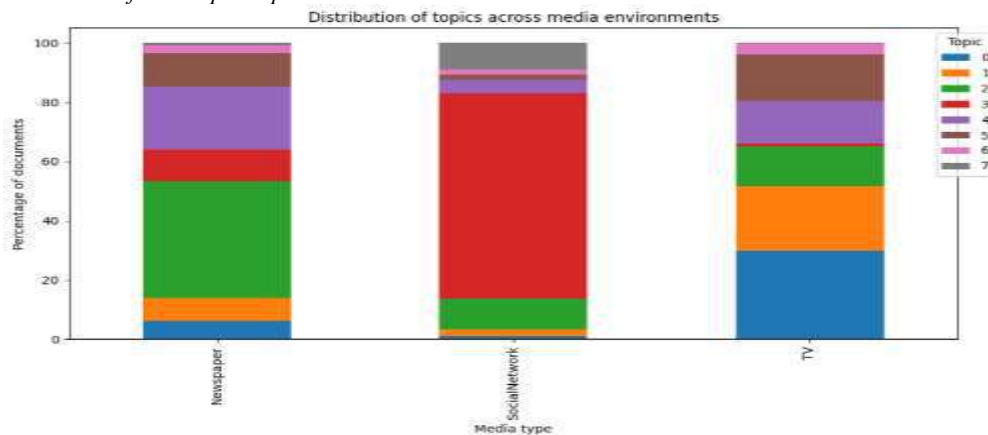
Table 5. Distribution of BERTopic Topics and Representative Keywords

Topic	Documents	Keywords	Discursive Configuration
0	6,563	government, state, institutions, politics, decisions	expertise in the political governance of the pandemic
1	5,877	government, parliament, decree, majority, opposition	political debate on the use of scientific expertise
2	5,678	vaccine, vaccines, campaign, doses, healthcare	organization of the vaccination campaign
3	3,930	covid, experts, scientists, virus, research	scientific communication and expertise

4	3,661	measures, restrictions, emergency, health security	institutional management of the health crisis
5	792	controversy, opposition, criticism, public debate	political controversy over scientific authority
6	225	vaccines, distribution, regions, vaccination centers	territorial implementation of vaccination
7	207	vaccine, social media, public opinion, risk	public negotiation of scientific expertise

Figure 4 shows the distribution of topics across the different media environments. Television channels are strongly associated with topics related to the political governance of the pandemic, with a significant presence of content referring to institutional decisions, public policies, and emergency management. National newspapers display a more balanced thematic distribution, combining political management of the crisis with topics related to the vaccination campaign and scientific communication. Social media, in contrast, show a stronger concentration in topics related to scientific expertise and public risk negotiation, suggesting a greater presence of participatory dynamics and opinion-based debate. Overall, this distribution confirms previous findings: although media discourse on the pandemic relies on a shared semantic foundation, it is articulated differently across communicative environments, each privileging specific narrative and thematic configurations.

Figure 4. Distribution of BERTopic Topics Across Media Environment



5. Discussion and Conclusion

The results show that scientific expertise during the Italian vaccination campaign took distinct linguistic and thematic forms across media. The TF-IDF analysis shows that newspapers tend to employ a lexicon more closely aligned with informational and scientific discourse, characterized by references to data, research, and regulatory institutions. Television discourse, by contrast, appears more oriented toward the political-institutional management of the crisis, with a strong presence of terms related to government actors, public decision-making, and emergency measures. Social media, finally, display a more heterogeneous lexical profile, in which references to risk, vaccine safety, and public controversies emerge prominently. Second, correspondence analysis and the clustering of lexical configurations reveal a semantic space structured around a limited number of recurring narrative configurations. Discourse on the pandemic appears organized around three main discursive domains: the biomedical interpretation of the disease, the political-institutional governance of the crisis, and the operational management of vaccination policies. Semantic embeddings confirm both differences and overlaps across media environments. The semantic map obtained through Sentence-BERT reveals a significant area of thematic overlap among television, newspapers, and social media. This indicates that different media environments do not necessarily produce

entirely separate thematic universes but rather tend to reinterpret similar topics through distinct communicative registers. Scientific authority therefore emerges less as a stable form of expertise than as a relational and mediatized configuration continuously shaped by different communicative logics. From a sociological perspective, these findings extend previous research on the public construction of expertise (Cataldo *et al.*, 2022; 2025) by shifting the analytical focus from the visibility of experts to the linguistic configurations through which expertise is articulated in the public sphere. The study contributes to the field of statistical text analysis by proposing a replicable cross-media analytical framework capable of comparing heterogeneous corpora within a unified computational pipeline. Several limitations should nevertheless be acknowledged. First, the analysis is limited to a single national context and to a specific temporal window corresponding to the Italian vaccination campaign. Second, although contextual embeddings enable the capture of semantic relationships between texts, they remain dependent on the training data of language models and may therefore reflect pre-existing linguistic biases. Finally, quantitative modeling cannot fully replace close qualitative interpretation of discourse; rather, it should be considered a complementary tool capable of identifying structural patterns that require further interpretative analysis. Future research could extend this approach to international comparative contexts or introduce dynamic topic modeling to analyze the evolution of discursive configurations over time. Overall, the study suggests that scientific authority during the COVID-19 vaccination campaign did not circulate as a stable or unified form of expertise but was continuously reconstructed through the communicative logics of different media environments.

References

- Benson, R. (2013). *Shaping immigration news: A French-American comparison*. Cambridge University Press.
- Bucchi, M., & Trench, B. (2021). *Routledge handbook of public communication of science and technology* (3rd ed.). Routledge.
- Cataldo, R., Iazzetta, F., Punziano, G., & Saracino, B. (2026). *Mediaticamente esperti: La costruzione pubblica dell'expertise scientifica nella campagna vaccinale anti-COVID 19 in Italia*. FrancoAngeli.
- Cataldo, R., Punziano, G., Saracino, B., & Iazzetta, F. (2022). Expert construction during the Italian vaccination campaign against COVID-19: positional, reputational, and communicational spheres in comparison. *Culture e Studi Del Sociale*, 7(1), 101-118.
- Cataldo, R., Punziano, G., Saracino, B., & Iazzetta, F. (2024). Experts and science communication in time of Covid vaccination campaign: the Italian case. *Sociologia Italiana*, (24).
- Collins, H., & Evans, R. (2002). The third wave of science studies. *Social Studies of Science*, 32(2), 235–296.
- Couldry, N. (2014). *The myth of media power*. Polity.
- Couldry, N., & Hepp, A. (2013). Conceptualizing mediatization. *Communication Theory*, 23(3), 191–202.
- Fairclough, N. (2003). *Analyzing discourse: Textual analysis for social research*. Routledge.
- Funtowicz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures*, 25(7), 739-755.
- Goodell, R. (1977). *The visible scientists*. Little, Brown.
- Hjarvard, S. (2008). The mediatization of society. *Nordicom Review*, 29(2), 105–134.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., & Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915), 721–723.
- Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Dunod.
- Peters, H. P. (2014). Gap between science and media revisited. *Public Understanding of Science*, 23(1), 1–6.
- Roeh, I., & Feldman, S. (1984). The rhetoric of news. *Journal of Communication*, 34(1), 20–39.
- van Dijk, T. A. (2008). *Discourse and power*. Palgrave Macmillan.

Discourse structure across technological maturity levels: a textometric approach

Mélusine Caillard^{1,2}, Gaël Lejeune¹

¹Sorbonne University — gael.lejeune@sorbonne-universite.fr

²IKOS Lab — melusine.caillard@etu.sorbonne-universite.fr

Abstract

Technological maturity is commonly assessed using the Technology Readiness Level (TRL) scale, but these levels are rarely explicit in textual data. This study proposes an original approach combining Natural Language Processing (NLP) and textometry to analyze a corpus of energy technology descriptions. It relies on POS distribution, morphosyntactic n-grams, and thematic lexical masking. The results show that maturity stages correspond to distinct morphosyntactic patterns, suggesting that technological maturity is partly encoded in discourse structure. This work provides a dedicated corpus, a methodological contribution and new perspectives for the automated analysis of technological maturity in large innovation corpora.

Keywords: Part of speech, morpho-syntactic patterns, technological maturity, innovation discourse

Résumé

Les niveaux de maturité technologique (TRL) sont rarement explicités dans les textes. Nous proposons une approche combinant traitement automatique des langues (TAL) et textométrie pour en inférer des indices à partir de descriptions de technologies énergétiques, via l'analyse des POS, des n-grammes morpho-syntaxiques et un masquage lexical thématique. Les résultats mettent en évidence des régularités morpho-syntaxiques propres aux différents stades de maturité, montrant que celle-ci est en partie inscrite dans la structure du discours. Cette contribution ouvre la voie à une estimation automatisée de la maturité dans de grands corpus d'innovation.

Mots clés: Discours de l'innovation, motifs morpho-syntaxique, POS n-grams

1. Introduction

The increasing availability of large textual data has opened new perspectives for the analysis of innovation processes. Technological descriptions, project reports, and technical documentation constitute a vast source of information reflecting the development, validation, and deployment of technologies. Despite this abundance of textual data, the linguistic structures underlying technological development remain underexplored. Most studies focus primarily on thematic content or metadata, while semantic and structural properties receive less attention. Different phases of technology development can be described by the Technology Readiness Level (TRL) scale. Originally introduced by NASA to assess the maturity of technological systems (Mankins, 1995), the TRL framework structures technological development into a sequence of stages ranging from early conceptual research (TRL 1–3), through experimental validation and prototyping (4–6), to system demonstration and full industrial deployment (7–9). Within this progression, a critical transition occurs between intermediate development stages and industrial application. This phase is frequently referred to as the “Valley of Death” where technologies

experimentally validated struggle to secure the financial, industrial, or institutional support required for large-scale demonstration and commercialization (Nwaka, 2021). As a result, many technologies stagnate between TRL 4 and TRL 6 and fail to reach operational deployment.

Assessing technological maturity is a critical task for industrial and technological organizations, as it informs strategic decisions related to technology investment, development priorities, and market entry (Rodríguez López et al., 2023). Maturity assessment frameworks such as Technology Readiness Levels (TRLs), S-curves, or the Gartner hype cycle are widely used to support these decisions, yet they often rely heavily on expert judgment and qualitative evaluation processes (Loginova et al., 2025). While expert-based approaches provide valuable insights, they remain time-consuming and difficult to scale when dealing with the rapidly growing volume of technological information available today. At the same time, the increasing availability of large textual datasets describing technological developments creates new opportunities for quantitative and automated maturity assessment based on big text data analysis (Martino, 2003).

Several studies have explored this perspective by applying computational methods to large textual corpora. Early work by Britt et al. (2008) proposed an automated TRL assignment approach based on machine classification of company reports using latent semantic indexing. Later contributions emphasized the potential of large-scale textual analysis to identify emerging technologies and track their maturity dynamics. For instance, Cunningham (2014) argued that big data analysis systems could help both public and private actors detect promising technologies at early stages of development. Despite these advances, the automation of technological maturity assessment remains an open methodological challenge, particularly in terms of identifying robust linguistic or semantic indicators of maturity in textual data (Loginova et al., 2025).

Existing approaches to maturity detection rely either on expert assessments or on machine learning approaches constrained by the limited availability of large annotated corpora (Voltan et al., 2024). Several studies have highlighted the methodological challenges involved in automating the assessment of technological maturity levels (Faidi and Olechowski, 2020; Lezama-Nicolás et al., 2018). However, explicit references to TRL levels are rarely present in textual descriptions of technologies (Voltan et al., 2024). As a result, the assessment of technological maturity often requires expert interpretation of implicit signals embedded in technical discourse.

From a linguistic perspective, the organization of discourse often reflects underlying conceptual structures. Research in discourse analysis and grammar has shown that recurrent patterns of language use contribute to the emergence and stabilization of grammatical constructions (Bowie and Popova, 2019; Thompson, 1990). In this view, linguistic patterns do not only convey thematic content but also disclose structural aspects of communicative practices. If technological development follows a structured progression, it is therefore reasonable to expect that different stages of maturity may correspond to distinct discursive configurations.

Textometric approaches provide a suitable framework for investigating such structural regularities. By analysing the statistical distribution of linguistic forms across textual corpora, textometry allows the identification of characteristic grammatical patterns beyond lexical content. Methods based on distributional analysis, specificity measures, and morpho-syntactic annotation make it possible to detect statistically significant variations in grammatical structures across textual categories (Evert, 2004). However, this potential remains underused in grammatical analysis: textometry is still mainly applied to lexical phenomena, despite its design enabling both quantitative corpus analysis and qualitative expert interpretation.

Based on this perspective, the present study investigates whether technological maturity can be observed through variations in the distribution and organization of lexico-grammatical forms.

More specifically, it examines whether different TRL levels correspond to distinct morpho-syntactic configurations in technological descriptions. Using a textometric analysis combining POS distribution analysis, sequential POS n-gram extraction, and lexical masking, the study aims to identify structural linguistic markers associated with technological maturity levels.

This study therefore investigates whether technological maturity follows a discursive trajectory that can be identified through variations in the distribution and sequential organization of morpho-syntactic patterns in technological descriptions.

2. Research hypotheses

Descriptions of technologies often contain specialized terminology related to a particular field, which may bias statistical analyses by emphasizing thematic content rather than structural properties of the discourse. This lexical bias may lead to the mistaken interpretation that maturity-related signals are reflections of domain-specific vocabulary. To address this issue, this study investigates whether technological maturity is encoded not only through lexical content but also through morpho-syntactic structures. If maturity is structurally represented in discourse, the associated linguistic patterns should remain observable even after neutralization of domain-specific terminology. Based on this assumption, three hypotheses are formulated:

H1 — Distributional variation. The distribution of morpho-syntactic categories (POS) varies significantly across Technology Readiness Levels (TRLs). Different stages of technological maturity are expected to correspond to distinct grammatical configurations reflecting conceptual description, validation processes, or industrial performance reporting.

H2 — Sequential patterns. Technological maturity levels are associated with specific morpho-syntactic sequences (POS n-grams). These sequential configurations are expected to correspond to recurrent discursive patterns such as capability framing, experimental validation, performance quantification, or operational scaling.

H3 — Structural robustness. The morpho-syntactic patterns associated with technological maturity remain largely stable after lexical neutralization of domain-specific terms. If this hypothesis holds, it would indicate that maturity-related signals are structurally encoded in discourse rather than solely dependent on thematic vocabulary.

These hypotheses are tested through a textometric analysis combining POS distribution analysis, sequential POS n-gram extraction, and comparison between the original and lexically masked corpora.

3. Corpus description

The corpus used in this study consists of technological descriptions extracted from the International Energy Agency (IEA) technology database¹. Each document corresponds to a technological entry describing the development, applications, and performance characteristics of a specific energy technology. The corpus consists of standardised technological descriptions that have been rated by IEA experts from TRL 2 to 9. This institutional and stylistic uniformity helps to identify maturity-related linguistic patterns by limiting external discursive variation. The resulting corpus contains 77,692 tokens and is organized at the document level, each document corresponding to a single technological description. The TRL value associated with each

¹ <https://www.iea.org/data-and-statistics/data-tools/etp-clean-energy-technology-guide>

entry is stored as a metadata attribute and used as the primary variable for the analysis. The dataset includes 628 technological descriptions with an uneven distribution across maturity levels (TRL 2: 5, TRL 3: 28, TRL 4: 63, TRL 5: 64, TRL 6: 81, TRL 7: 67, TRL 8: 69, TRL 9: 251). Across the entire corpus, explicit references to TRL levels occur only 42 times. Lafon specificity scores are computed relative to global frequencies, statistically identifying terms over- or under-represented in a sub-corpus while remaining robust to size variations (Lafon, 1980). In order to evaluate the influence of domain-specific vocabulary on the detection of maturity-related linguistic patterns, two structurally identical versions of the corpus were constructed: (1) **Original corpus**: the original technological descriptions preserving the full lexical content. (2) **Masked corpus**: a modified version in which domain-specific terms related to energy technologies were replaced by neutral placeholders corresponding to their part-of-speech category. The masking procedure aimed to neutralize thematic vocabulary while preserving the morpho-syntactic structure of the texts. Domain-specific terms were identified using iterative prompting and manual validation approach and replaced with their corresponding POS placeholder. In total, 96 lexical items were masked in the original dataset. This process ensures that the masked corpus retains the same syntactic organization, token count, and document structure as the original corpus while minimizing lexical domain effects.

4. Study design and methods

The objective is to identify whether stages of technological maturity correspond to distinct lexico-grammatical structures. The analysis combines distributional textometric methods and sequential pattern analysis in order to detect structural linguistic markers associated with TRLs. Experiments were conducted using the TXM platform (v0.8.4) (Heiden, 2010), an open-source environment for textometric analysis. The complete analytical protocol was independently replicated on original and masked corpora in order to evaluate the structural stability of the identified patterns. Both corpora were imported into TXM using an XML representation. Each technological description was encoded as a structured document including hierarchical units (article, body, text) and a TRL label stored as an attribute `article@label`. During import, automatic lemmatization and part-of-speech tagging were applied using the default TreeTagger pipeline integrated into TXM and the *englishpos* tagset. The resulting corpus representation includes POS annotations *englishpos* and associated lemmas *englishlemma*, which were subsequently used for the extraction of grammatical distributions and sequential morpho-syntactic patterns across maturity levels.

4.1. POS distribution analysis

The analysis first focuses on the distribution of morpho-syntactic categories across TRL levels (H1). For each POS category, specificity scores were computed using Lafon’s index in order to identify statistically over or under represented forms within each maturity level. Figure 1 presents a specificity heatmap for the main POS tags across TRL levels. Only values exceeding ± 2 are interpreted as statistically significant. The figure reports the overall frequency of each POS category in the corpus, providing contextual information for interpreting specificity scores. The most frequent categories include plural nouns (NNS), adverbs (RB), and proper nouns (NP), indicating that the specificity patterns include well-represented grammatical features rather than rare ones. This supports the robustness of the statistical signals identified. To ensure that statistically significant POS specificity correspond to coherent discursive functions rather than accidental frequency effects, a concordance analysis was conducted for each POS category (Tab. 1).

POS_tag	Freq in IEA classic	TRL2	TRL3	TRL4	TRL5	TRL6	TRL7	TRL8	TRL9	12,7
Noun plural (NNS)	5734	0,4	2,6	-0,3	-0,7	-2,6	-1,4	0,4	6,8	
Adverb (RB)	2393	-0,4	-0,4	0,9	0,2	2,7	0,5	-0,9	-3,3	
Proper noun (NP)	1870	-1,5	-1,9	-1,1	3	5,7	4,7	-0,7	-7,9	
Verb, gerund/participle (VVG)	1764	1	-0,3	-0,5	0,7	-2,2	0,7	2,6	-0,9	
Modal (MD)	1153	0,4	0,7	-0,5	2,1	0,5	-0,6	-0,4	-0,7	0
Cardinal number (CD)	975	-0,9	-4,2	0,9	-0,9	-0,7	-2,3	-1,1	12,7	
Verb, present (VVZ)	792	-0,7	-0,3	0,8	-0,3	1,2	-0,5	-0,6	-1,3	
Verb "be" (VBP)	605	0,4	2,3	-0,4	-0,4	-0,4	0,3	-0,6	-0,3	
Symbol (SYM)	421	-0,3	-3,1	1,9	0,9	-1	1,4	-0,3	-1,3	
Verb have (VHZ)	139	0,4	-0,9	0,9	-0,3	0,6	0,5	-0,4	-1,8	
Adjective, superlative (JJS)	78	-0,1	-0,3	-0,9	-0,7	-1,8	-0,5	-3,7	7,7	-7,9

Figure 1: Lafon specificity scores for major POS categories across TRL levels (original corpus).

4.2. POS n-gram analysis

To investigate whether technological maturity is associated with recurrent morpho-syntactic configurations (H2), sequential morpho-syntactic patterns were extracted in the form of POS bi-grams and trigrams. Only n-grams occurring at least 30 times in the global corpus were retained for statistical analysis, in order to avoid unstable specificity associated with low-frequency sequences. This allows identification of sequential configurations that are statistically characteristic of particular maturity levels.

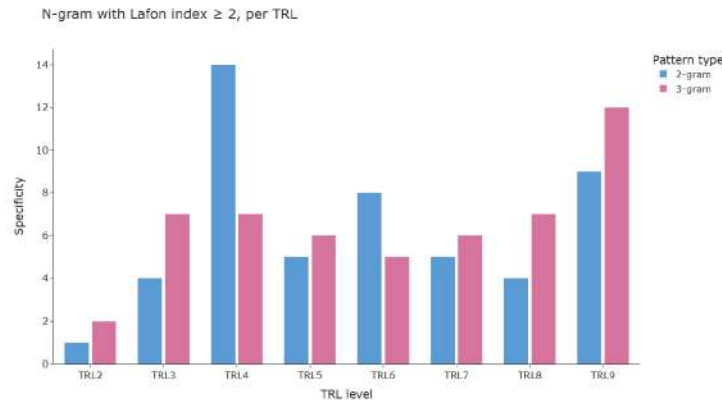


Figure 2: Number of statistically significant POS n-grams across TRL levels.

Figure 2 shows the number of statistically significant POS n-grams identified for each TRL level. The distribution shows that characteristic sequential patterns are not evenly spread across the different levels of maturity. TRL 9 has more statistically significant POS n-grams, suggesting prominent discourse configurations at higher maturity. In order to examine the global structural intensity of sequential patterns across maturity levels, the distribution of Lafon specificity scores for POS trigrams was analyzed. Rather than focusing only on individual significant sequences, this analysis considers the overall dispersion and central tendency of specificity scores within each TRL category. Figure 3 presents the distribution of specificity scores for POS n-grams across TRL levels using box-and-whisker plot visualization. Each box represents the interquartile range of n-gram specificity scores for a given maturity level, while whiskers indicate the dispersion of values and highlight the presence of highly specific sequences.

4.3. Concordance validation

Concordance analysis was conducted to validate the functional interpretation of statistically significant POS (and POS n-grams) categories. For each significant POS–TRL association, a sam-

ple of 20 occurrences was manually examined and classified according to discursive function (Tab. 1). Discursive functions identified during qualitative examination appeared to correlate with maturity. Early stages (TRL 2-3) are linked to *conceptual description*, introducing tech ideas and uses. Intermediate stages (TRL 4-6) are linked to *capability framing* and *comparative evaluation*, expressing possible functionalities and validation contexts. The final stage (TRL 9) is linked to *operational description* and *performance quantification*, focusing on industrial processes and measurable indicators. To evaluate the structural robustness of the identified linguistic markers, Lafon specificity scores obtained from the original and masked corpora were compared. Figure 4 presents a scatter plot of specificity scores for main POS category in both corpora. The strong alignment of the points along the diagonal indicates that most maturity-related linguistic signals remain stable after lexical masking.

Table 1: Representative examples illustrating discursive functions associated with maturity stages.

POS	TRL	Text	Discursive function
NNS	3	... This might have applications in regions such as Japan, ...	conceptual description
MD	5	... Ammonia can provide heat to low temperature heating ...	capability framing
RB	6	... production is significantly larger than in the case of gasification ...	comparative evaluation
VVG	9	... the rotor turns on a vertical axis; oscillating hydrofoils, attached to an arm ...	operational description
CD	9	... to deliver fuel to both at 350 and 700 bar. Hydrogen refuelling ...	performance quantification

5. Results

5.1. Grammatical markers of maturity

The POS specificity analysis highlights several grammatical markers associated with different stages of technological maturity. Early-stage maturity (TRL 3) is characterized by an over-representation of plural nouns (NNS: 2.6) and present tense verbs (VBP: 2.3), reflecting exploratory and conceptual discourse describing multiple potential technological applications. Intermediate levels (TRL 5–7) show stronger contextualization through proper nouns (NP: 3.0) and modal verbs (MD: 2.1). Modal constructions such as *can*, *may*, or *could* frequently occur in TRL 5, corresponding to capability framing during validation phases. Verb forms also evolve with maturity: present participles (VVG: 2.6 in TRL 8) become more frequent in later stages, reflecting operational descriptions of testing, operating, or scaling systems. Advanced maturity (TRL 9) is characterized by a strong over-representation of cardinals (CD: 12.7) and superlative adjectives (JJS: 7.7). Cardinals are associated with quantified performance indicators (e.g., voltage, capacity, efficiency), while superlatives express optimization (e.g., “highest efficiency”, “most cost-effective”). For instance, concordances for JJS frequently include expressions such as “highest energy density”, “most cost-effective”, “largest capacity”, or “best performance”, all referring to industrial-scale evaluation and comparative optimization (*cf.* Tab.1, Fig.2). This confirms that the statistical signal reflects a coherent discursive function aligned with advanced technological maturity. Interestingly, the intermediate stages display more heterogeneous and less stabilized linguistic patterns. This observation echoes the well-known “Valley of Death” phenomenon in technological development, where technologies

undergoing validation and demonstration phases often face uncertainty and slower progression toward industrial deployment. Overall, these distributional contrasts support the hypothesis that technological maturity is associated with distinct grammatical configurations.

5.2. Sequential morpho-syntactic patterns

Beyond individual grammatical categories, sequential morpho-syntactic configurations were analysed through POS n-grams. The analysis foregrounds that TRL 9 stands out as exhibiting markedly stronger sequential structuring than other levels. In particular, trigram configurations show higher dispersion and stronger specificity scores in advanced TRL levels. In addition to identifying POS n-grams, the distribution of the top 20 most specific sequences for each TRL had been statistically analyzed. Whisker boxplot visualizations (Fig. 3) show an increasing dispersion and higher positive specificity values at advanced maturity stages, particularly for TRL 9. Only the 20 most specific n-grams per TRL were retained in order to focus on the most characteristic sequential patterns while keeping the analysis interpretable. Higher maturity levels therefore exhibit a greater variability and intensity of specific sequential patterns, indicating a stronger stabilization of discursive structures. This effect is particularly pronounced for trigram configurations, which display higher specificity values and wider dispersion at advanced TRLs.

In contrast, early maturity stages show lower dispersion and more moderate specificity values, suggesting less stabilized and more heterogeneous discursive structures. These observations support the hypothesis that technological maturity is reflected not only in the distribution of individual grammatical categories but also in the progressive stabilization of sequential morpho-syntactic patterns (Fig. 2). The presence of highly specific trigram patterns in advanced maturity stages suggests that industrialized technologies tend to rely on more standardized discourse structures related to performance reporting, operational processes, and system optimization.

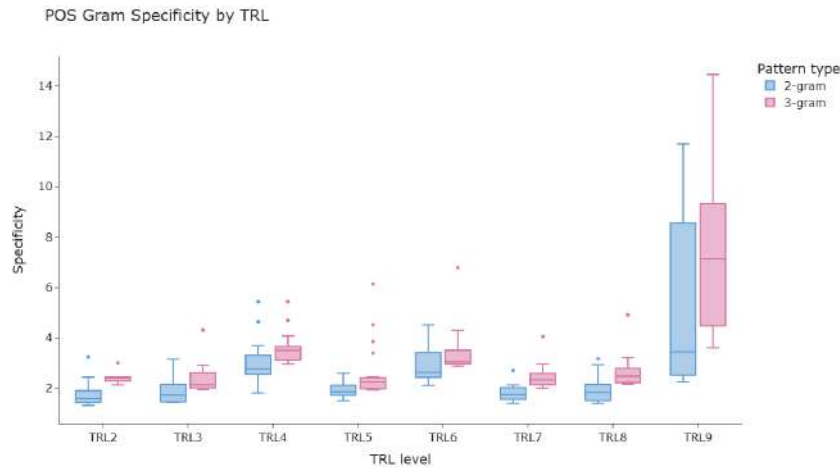


Figure 3: Distribution of Lafon specificity scores for POS n-grams across TRL levels.

5.3. Structural establishment of maturity

The replication of the POS specificity analysis on the masked corpus shows a strong structural stability of most maturity-related linguistic markers. In particular, categories associated with quantification (CD), plural generalization (NNS), superlative constructions (JJS), and modal usage (MD) remain consistently distributed across TRL levels. Only proper nouns (NP) show a noticeable attenuation, suggesting that contextualization markers may partially depend on

domain-specific lexical content. Overall, these patterns correspond to discursive functions such as conceptual description, validation reporting, performance quantification, and optimization.

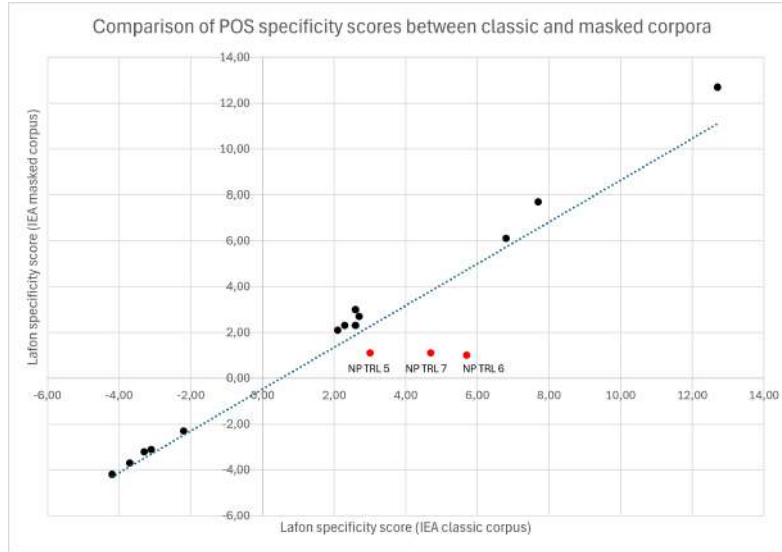


Figure 4: Comparison of POS specificity scores between original and masked corpora

To evaluate the structural robustness of the identified linguistic markers, the Lafon specificity scores obtained for the original and masked corpora were compared for each POS–TRL pair exhibiting significant specificity. As shown in Figure 4, most points lie close to the diagonal, indicating a strong correspondence between specificity scores in the two corpora. This result suggests that the majority of maturity-related linguistic signals remain stable after lexical masking. A small number of deviations can still be observed, corresponding to the proper nouns (NP) category for TRL 5, 6 and 7 (in red on Fig. 4). In the original corpus, proper nouns display relatively high specificity scores because they refer to domain-specific entities such as acronyms, organisations, programmes or industrial initiatives. After thematic masking, these lexical items are replaced by neutral placeholders, which substantially reduces their statistical specificity. As a consequence, proper nouns (NP) appear less specific in the masked corpus and deviate from the regression line. Overall, the results support the hypothesis that technological maturity is primarily encoded through morpho-syntactic structures rather than lexical content. To further validate the functional interpretation of statistically significant POS categories, concordance were manually examined for each POS–TRL association. Table 1 presents representative examples illustrating the discursive roles associated with different maturity stages. Early stages of technological development (TRL 3) are characterized by plural conceptual references, such as *applications*, reflecting exploratory descriptions of potential technological uses. Intermediate stages exhibit modal constructions that frame technological capabilities and possible applications during validation phases. At TRL 6, comparative adverbs frequently occur in evaluative contexts describing performance differences between technological alternatives. In contrast, advanced maturity levels (TRL 9) are associated with operational descriptions and quantified performance indicators. Present participles typically describe ongoing industrial processes, while numerical expressions provide precise technical specifications related to system performance. Superlative adjectives are frequently associated with optimization statements in mature technological descriptions. These concordance observations confirm that the statistical POS specificity correspond to coherent discursive functions aligned with maturity stages.

6. Discussion & Conclusion

The results of this study show that technological maturity is reflected not only in the lexical content of technological descriptions but also in their structural linguistic organization. The distribution of grammatical categories and the presence of recurrent morpho-syntactic sequences vary systematically across TRLs. These linguistic patterns should therefore be interpreted as statistical associations with maturity levels rather than direct causal markers of technological maturity itself. Some observed configurations also reflect institutional or genre conventions associated with mature technological reporting. Yet the consistency of the patterns across TRL levels suggests they capture something beyond mere genre effects. These variations suggest that technological development is supported by a progressive transformation of discursive practices.

Early-stage technologies tend to be described through exploratory and conceptual discourse, characterized by plural generalizations and definitional structures. Intermediate stages exhibit stronger contextualization and capability framing, reflecting phases of experimental validation and technological demonstration. In contrast, advanced maturity levels are associated with quantified evaluation and optimization language, typical of industrial deployment and performance reporting. This progression mirrors the evolution of the TRL scale, in which innovation becomes progressively formalized as technologies move toward industrialization, while the intermediate stages reflect the uncertainties and challenges of the so-called “Valley of Death”.

The comparison between the original and masked corpora further strengthens this interpretation. The stability of most linguistic signals after lexical neutralization confirms that maturity-related information is not only conveyed through domain-specific terminology, but is also encoded in the structural organization of discourse. This finding highlights the importance of considering morpho-syntactic structures when analysing technological texts. From a methodological perspective, this study illustrates the potential of textometric approaches for analysing innovation processes through large textual corpora. By focusing on distributional and sequential linguistic patterns, it becomes possible to identify implicit indicators of technological maturity without relying exclusively on thematic vocabulary or expert annotation. Such indicators may contribute to the development of analytical tools supporting expert analysis of technological development across large document collections. Beyond description, POS distribution profiles and recurrent sequential patterns could serve as interpretable input features for supervised models trained to predict TRL categories, complementing or partially replacing costly expert annotation.

Several limitations should be acknowledged. First, the analysis is based on a corpus restricted to a technological domain. Although this corpus provides a coherent experimental framework, the generalization of the observed linguistic patterns remains to be tested across other technological fields. Second, the analysis relies on automatic POS tagging, which introduces errors affecting the extraction of morpho-syntactic patterns. Future research should apply this methodology to multi-domain corpora to test the stability of maturity markers across technological contexts. Integrating sequential morpho-syntactic patterns, like mixed POS and lemma sequences, allows quantitative stylistics (Sacy and Legallois, 2025) to isolate discursive nuances. These motifs provide interpretable features for hybrid NLP architectures, augmenting semantic embeddings to refine automated TRL estimation and innovation monitoring. A natural next step would be to evaluate their predictive power in a classification setting, for instance with cross-validated experiments on labelled corpora. Overall, we showed that technological maturity can be approached not only as a technical or managerial concept but also as a phenomenon reflected in the structure of technological discourse. Textometric analysis offers promising perspectives for exploring the dynamics of technological development through large-scale textual data.

References

- Bowie J. and Popova G. (2019). Grammar and discourse. In Aarts B., Bowie J., and Popova G., editors, *The Oxford Handbook of English Grammar*. Oxford University Press.
- Britt B. L., Berry M. W., Browne M., Merrell M. A., and Kolpack J. (2008). Document classification techniques for automated technology readiness level analysis. *Journal of the American Society for Information Science and Technology*, 59(4):675–680.
- Cunningham S. W. (2014). Big data and technology readiness levels. *IEEE Engineering Management Review*, 42(1):8–9.
- Evert S. (2004). The statistical analysis of morphosyntactic distributions. In Lino M. T., Xavier M. F., Ferreira F., Costa R., and Silva R., editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. European Language Resources Association.
- Faidi S. and Olechowski A. (2020). Identifying gaps in automating the assessment of technology readiness levels. In *Proceedings of the Design Society: DESIGN Conference*, volume 1, pp. 551–558.
- Heiden S. (2010). The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24)*, pp. 389–398.
- Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1:127–165.
- Lezama-Nicolás R., Rodríguez-Salvador M., Río-Belver R., and Bildosola I. (2018). A bibliometric method for assessing technological maturity: The case of additive manufacturing. *Scientometrics*, 117(3):1425–1452.
- Loginova I. V., Grozovskiy F., and Piekalnits A. (2025). Prospects for big text data application in technology maturity assessment. *Automatic Documentation and Mathematical Linguistics*.
- Mankins J. C. (1995). Technology readiness level, a white paper. Technical report, NASA. White Paper.
- Martino J. P. (2003). A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change*, 70(8):719–733.
- Nwaka S. (2021). Technology readiness levels, the valley of death and scaling up innovations. In *Social and Technological Innovation in Africa*. Palgrave Macmillan, Singapore.
- Rodríguez López N., Alvarez Cabal J. V., Cuiñas M. C., and Ortega Fernández F. (2023). Applicability of technology maturity level evaluation methodologies within small- and medium-sized organizations: Prospects and proposals. *Systems*, 11(8):387.
- Sacy A. d. and Legallois D. (2025). Les librairies motiver et pymotifs pour identifier les patrons lexico-grammaticaux significatifs (ou motifs séquentiels). *Corpus*, (27).
- Thompson S. A. (1990). Discourse and grammar. *Text - Interdisciplinary Journal for the Study of Discourse*, 10(1-2):113–116.
- Voltan J., Girardi R., Santos T., Abreu D., Farias J., Araujo D., Cruz V., and Goldschmidt R. (2024). Methodology to build labeled corpora and classification models to assess technological readiness: A case study with defense technologies described by texts in pt-BR. In *Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBBDD 2024)*, pp. 848–854.

Structuring sectorial knowledge: automated detection of lexical relations

Claudia Brunini¹, Eugenio Concetti², Giampiero D'Alessandro³

¹Istat – brunini@istat.it

²Istat – eugenio.concetti@istat.it

³Istat – g.dalessandro@istat.it

Abstract 1

The study addresses the challenge of identifying and managing semantic relations in the specialised language of official statistics, where terminological control is essential to ensure clarity, consistency, and traceability. To enhance the efficiency and scalability of the process, the study tests an automatic technique drawn from computational linguistics and natural language processing—Latent Semantic Analysis (LSA)—which is particularly well suited to identifying distributional semantic proximity and associative relations. The experiment is conducted on a diachronic corpus of Istat Annual Reports covering the period 2001–2023. The methodology combines corpus segmentation, advanced text preprocessing, and the construction of a weighted term–document matrix, followed by dimensionality reduction through Singular Value Decomposition. The results show that LSA can effectively support the identification of semantic relations such as synonymy, antonymy, and associative links within a specialised lexicon. The latent semantic space further reveals coherent domain-specific clusters, highlighting the underlying semantic and discursive organisation of the corpus. While the model proves robust in capturing semantic proximity and terminological structures, it shows limitations in detecting polysemy and in capturing long-term diachronic variation. Overall, the study confirms the value of distributional semantic methods as tools supporting terminology management in institutional contexts, whilst also pointing to the need for complementary approaches, such as context-sensitive models, to address more complex semantic phenomena.

Keywords: Semantic relations; Statistical text analysis; Natural language processing.

Abstract 2

Lo studio affronta la sfida dell'identificazione e della gestione delle relazioni semantiche nel linguaggio settoriale della statistica ufficiale, in cui il controllo terminologico è essenziale per garantire chiarezza, coerenza e tracciabilità. Per migliorare l'efficienza e la scalabilità del processo, lo studio valuta una tecnica automatizzata derivata dalla linguistica computazionale e dal trattamento automatico del linguaggio naturale, la Latent Semantic Analysis (LSA), particolarmente adatta all'individuazione della prossimità semantica distribuzionale e di relazioni associative. La sperimentazione viene condotta su un corpus diacronico di Rapporti Annuali dell'Istat, relativi al periodo 2001–2023. La metodologia combina la segmentazione del corpus, un preprocessing testuale avanzato e la costruzione di una matrice termini-documenti pesata, seguita da una riduzione della dimensionalità tramite decomposizione in valori singolari. I risultati mostrano che la LSA può supportare efficacemente l'identificazione di relazioni semantiche quali sinonimia, antonimia e associazioni lessicali all'interno di un lessico specialistico. Lo spazio semantico latente rivela inoltre cluster coerenti specifici del dominio, mettendo in luce l'organizzazione semantica e discorsiva sottostante del corpus. Sebbene il modello si dimostri robusto nel catturare la prossimità semantica e le strutture terminologiche, presenta limitazioni nell'individuazione della polisemia e nella rilevazione di variazioni diacroniche di lungo periodo. Nel complesso, lo studio conferma il valore dei metodi di semantica distribuzionale come strumenti a supporto della gestione terminologica nei contesti istituzionali, evidenziando anche la necessità di approcci complementari, quali i modelli sensibili al contesto, per affrontare fenomeni semantici più complessi.

Parole chiave: Relazioni semantiche; Analisi statistica testuale; Trattamento automatico del linguaggio naturale.

1. Introduction

In managing, exchanging, and disseminating statistical information, the effective treatment of textual descriptors is crucial. Data must be linked to unambiguous and unique designators to ensure traceability and prevent confusion. This requires precise mapping of semantic relations among designators, including synonymy and polysemy, which, if not properly monitored, may lead to errors. Mapping these relations is essential yet demanding, due to the expertise required and the costs of continuous monitoring. Since specialised language evolves rapidly, a permanent system of semantic control is necessary, and automatic methods for extracting semantic relations can provide valuable support. This study proposes a structured, partially automated approach for managing semantic resources in institutional contexts. While automatic methods for extracting semantic relations are growing, their application to institutional texts, especially diachronic corpora, remains limited. By applying Latent Semantic Analysis (LSA) to a longitudinal corpus of Istat Annual Reports (ARs), this work demonstrates how computational techniques can support expert-driven terminology management.

2. Data

The analysed corpus consists of the first chapter of ten ARs from Istat – Italian National Institute of Statistics, covering the period 2001-2023. The documents, published online in PDF format, were converted into .txt and subjected to segmentation and normalisation in the R environment. Since this study aims to analyse and extract semantic relations among terms, the corpus was designed to produce sufficiently homogeneous textual units and a lexical representation that preserves conceptually relevant multiword expressions. Data processing proceeded in three main phases: document segmentation, text preparation, and construction of the document-term matrix (DTM), used for subsequent analyses.

2.1. Document segmentation

The decision to construct the corpus from the first chapter of the Istat ARs is grounded in the fact that the introductory chapter provides a particularly representative synthesis of the statistical domain, with specific reference to its economic dimension, displaying both a high terminological density and strong thematic coherence. With a view to applying LSA, the corpus was subjected to a segmentation procedure aimed at defining the local context units on which the semantic model is based. LSA does not use syntactic information or the sequential order of words; rather, it relies on the distribution of terms across discrete textual contexts. The quality of segmentation therefore has a direct impact on the structure of the term-by-document matrix and, consequently, on the emergence of latent semantic relations. Following the methodological principles outlined by Thomas K. Landauer (Landauer et al. 1998), the corpus was divided into textual units characterised by local semantic coherence, each of which was treated as an autonomous contextual entity. Operationally, texts were imported into R through a procedure that identifies lines formatted within square brackets as segment markers. Each marker was interpreted as the heading of a new textual unit and used to generate a unique identifier by combining the name of the source file with the segment label, thus preserving both documentary provenance and sequential position within the text. Segmentation was based primarily on the editorial structure of the reports (paragraphs, subparagraphs, and thematic boxes), subsequently refined according to both dimensional and semantic criteria: each unit was defined so as not to exceed 11,000 characters and 2,000 words, while at the same time maintaining internal thematic coherence.

Table 1. Descriptive statistics of corpus segments

Segment with the lowest n. of chars. [2021 Scheda4]	Segment with the highest n. of chars. [2011 1.2.5]	Average number of characters per segment	Segment with the lowest n. of words [2021 Scheda4]	Segment with the highest n. of words [2011 1.2.5]	Average number of words per segment
831	10,996	5,469	140	1,727	854

In some cases, particularly extensive subsections were further subdivided in order to isolate distinct semantic nuclei; for example, in the 2001, 2003, 2006, and 2008 editions, subsection 1.2.5, devoted to the labour market, was systematically divided into three segments corresponding to three specific thematic areas: the state of employment, the sectors driving employment growth, and the state of unemployment. The resulting corpus therefore does not coincide with a mere collection of complete documents, but rather with a structured set of textual segments treated as autonomous analytical units, better suited to the detection of co-occurrence patterns and semantic proximity.

Table 2. Distribution of corpus segments by reference year of Annual Reports

Reference year	2001	2003	2006	2008	2011	2013	2016	2018	2021	2023	Total
N. segments	26	26	26	20	25	13	14	13	24	23	210

Through this procedure, the corpus was structured into 210 textual segments, each of which constitutes a column of the term-by-document matrix used for the analysis. During import, several technical aspects relevant to textual data quality were also addressed, including the removal of any initial spurious characters and the normalisation of file encoding. The resulting units were finally converted into a corpus object within the quanteda package (Benoit et al., 2018), with the segment taken as the basic documentary unit.

2.1. Text pre-processing

The text preparation process aimed to reduce lexical noise and make the corpus more suitable for the analysis of semantic relations. In an initial phase, the segments were tokenised by removing punctuation and numbers; subsequently, all tokens were converted to lowercase so as to neutralise purely graphic differences arising from capitalisation. A particularly relevant step involved the identification of multiword expressions, for which an external inventory of multiword lemmas from the Istat glossary was employed. In order to broaden the coverage of forms occurring in the corpus, a plural variant was also generated for each multiword expression, whenever possible, by means of a set of morphological rules for Italian, integrated with dictionary-based validation through the Hunspell R package (Ooms, 2026). This procedure allowed the inclusion of both the base form and some of its inflected variants, thereby reducing the risk of lexical dispersion. The multiword expressions selected in this way were then compacted into single tokens by concatenation with underscores. Lexical sequences such as “tasso di crescita” were thus treated as single units (tasso_di_crescita), preserving their semantic cohesion and preventing them from being split during subsequent stages of analysis. This step is methodologically important because it allows stable conceptual units to be represented more faithfully. In a second cleaning phase, Italian stop words were removed, supplemented by a domain-specific stop word list including highly frequent but weakly informative terms with respect to the aims of the analysis. A further filter was then applied, retaining only alphabetic tokens, optionally containing underscores, and with a minimum length of three characters. This choice made it possible to eliminate low-information items without compromising the preservation of the previously identified multiword expressions. At the end of preprocessing, the corpus was represented by means of a TDM, in which each segment

corresponds to a document and each token to a lexical feature. Subsequently, through a preliminary analysis of the TDM, additional stop words and new multiword expressions not included in the initial list were identified. This made it possible to carry out a second stage of textual refinement, aimed at further improving the quality of the lexical representation by reducing residual noise and incorporating compound lexical units emerging directly from the data. This iterative step contributed to making the corpus more coherent and better suited to semantic analysis, thereby increasing the reliability of the subsequent procedures used to analyse relations among terms.

3. Methods

According to Landauer (Landauer et al., 1998), LSA is a method for extracting and representing the contextual-usage meaning of words through statistical analysis of large text corpora. It relies on patterns of word co-occurrence to infer meaning, mirroring how humans reconstruct word meaning from linguistic regularities. The technique aims to identify the latent contextual structures that humans implicitly exploit in understanding language (Landauer and Dumais, 1997). These structures are derived statistically: words and multiword expressions are represented as points in a high-dimensional semantic space (typically 50–1,500 dimensions), where the dimensions capture latent patterns of co-occurrence that approximate the conceptual structures underlying human knowledge. To derive the latent semantic space, the Singular Value Decomposition (SVD) is applied to the input matrix, a term–document matrix (TDM). Before applying SVD, the TDM matrix is transformed through an appropriate weighting scheme, since raw word frequencies alone do not adequately capture the informational value of terms across the corpus. The weighting procedure typically involves three steps, each designed to address a specific issue: controlling the disproportionate influence of very frequent domain-specific terms, allowing terms that occur frequently in only a few documents to emerge as informative and compensating for the heterogeneous length of documents. Each of these steps is addressed through its own weighting factor, ensuring that the TDM matrix is adjusted in a way that reflects the true informational contribution of each term. In the first step a local weighting scheme is applied to moderate the influence of terms with very high raw frequencies:

$$w_{local} = \begin{cases} 1 + \log(tf_{t,d}) & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In the second step the global weighting scheme (entropy) allows informative terms to emerge by down-weighting words that occur with similar frequency across all documents, thereby highlighting those that carry discriminative value.

Entropy is calculated as follows:

$$H_t = - \sum_{d=1}^N p_{t,d} \log(p_{t,d}) \quad (2)$$

$p_{t,d}$ is the relative frequency of term t in in document d :

$$p_{t,d} = \frac{tf_{t,d}}{\sum_d^k tf_{t,d}} \quad k \text{ is the number of documents}$$

E_t represents the normalized value:

$$E_t = 1 - \frac{H_t}{\log(k)} \quad (3)$$

The global weight is:

$$w_{global}(t) = E_t \quad (4)$$

The weighted term-document matrix is obtained by multiplying local and global weights:

$$\mathbf{w}_{t,d} = [1 + \log(tf_{t,d})]E_t \quad (5)$$

Finally, the normalisation step ensures that all documents receive the same overall weight, regardless of their length. For each vector document d :

$$\vec{h}_d^{norm} = \frac{\vec{h}_d}{\|\vec{h}_d\|} \quad \vec{h}_d = (\mathbf{w}_{1,d}, \mathbf{w}_{2,d} \dots \mathbf{w}_{n,d}) \quad (6)$$

where:

$$\|\vec{h}_d\| = \sqrt{\sum_t w_{t,d}^2}$$

Together, these weighting steps produce a refined weighted term–document matrix $[TDM_w]$ that captures the most informative distributional patterns, providing an optimal basis for the subsequent SVD and for detecting semantic relations such as synonymy.

SVD decomposes the matrix into three components:

$$TDM_w = U \Sigma V^T \quad (7)$$

$[U]$ represents the terms as vectors in the latent space.

$[V]$ represents the contexts (documents, paragraphs or other text units) in the same latent space.

$[\Sigma]$ is a diagonal matrix whose singular values indicate the importance of the latent dimensions.

Together, these three matrices reconstruct the original matrix. SVD makes it possible to discard the least important dimensions, those associated with the smallest singular values. This dimensionality reduction yields a more compact representation, reduces noise, enhances generalisation and produces semantic relations that more closely resemble those observed in human cognition. It is at this stage that LSA can infer semantic similarity even in the absence of direct co-occurrence. For example, it can recognise that *Italia* and *Paese* refer to closely related concepts, or that *tasso_di_occupazione* and *forze_lavoro* are terms semantically connected, even if they rarely appear near one another in the corpus.

4. Results

This section presents the analysis conducted to identify semantic relations between terms, along with a description of the contextual semantic patterns from which these relations emerge.

4.1. Text pre-processing

To identify semantic relations within the corpus, the analysis follows a multi-step procedure. First, the application of the LSA to the TDM_w generates, for each term, a vector in a semantic space that reflects its distributional properties. Semantic similarity between terms is then assessed through cosine values: vectors with higher cosine scores are interpreted as semantically closer, potentially indicating synonymy, antonymy, singular–plural pairs, compound–component relations, or other forms of lexical association. Once the pairs with the highest absolute cosine values ($\geq |0.7|$) are identified, correlation coefficients are examined both on the original matrix $[X]$ and on the reconstructed matrix $[\hat{X}]$. This comparison makes it possible to assess whether the semantic proximity observed in the reduced vector space—reflected by the cosine values—corresponds to a meaningful improvement in the underlying distributional structure, rather than being an artefact of noise or accidental co-occurrences. As shown by Landauer, Foltz and Laham (1998, p. 13), the reduced-dimensional reconstruction produced by LSA can substantially alter word–word similarity estimates, revealing latent semantic relations that are not observable in the raw co-occurrence data. Finally, the analysis

is restricted to pairs of terms that are morphosyntactically identical, ensuring that the detected proximity is not driven by formal or grammatical differences. This multi layered procedure combining cosine-based selection, correlation-based validation, and morphosyntactic filtering, helps isolate genuine semantic relations and strengthens the reliability of the semantic equivalence detection process. A selection of meaningful examples of synonymic, antonymic, and associative relations (ISO, 2022) is presented in the tables below.

Table 3. Cosine and correlations for pairs of terms

SYNONYMS	sgravi/ agevolazioni	disoccupati/ persone_in_cerca_di_ occupazione	decelerazione/ rallentamento	deficit/ disavanzo
Cosine	0.748	0.923	0.892	0.769
Correlation $[X]$	0.490	0.699	0.553	0.428
Correlation $[\hat{X}]$	0.732	0.919	0.842	0.742
ANTONYMS	decelerazione/ accelerazione	importazioni/ esportazioni	negativo/ positive	emergenti/ avanzate
Cosine	0.698	0.902	0.803	0.919
Correlation $[X]$	0.289	0.720	0.366	0.782
Correlation $[\hat{X}]$	0.518	0.869	0.668	0.910
ASSOCIATIVE RELATIONS	sgravi/ politiche_fiscali	inflazione/ prezzi	importazioni/ interscambio	pressione_fiscale/ contributi_sociali
Cosine	0.790	0.845	0.825	0.975
Correlation $[X]$	0.413	0.645	0.646	0.856
Correlation $[\hat{X}]$	0.773	0.768	0.810	0.973

As expected, in the examined cases involving domain-specific technical terms, the model successfully identifies relevant semantic relations – synonymy, antonymy, and lexical association – characterised by high cosine similarity values. These values correspond to correlation coefficients in the original matrix $[X]$ that are positive, though not always high, and that systematically increase in the reconstructed matrix $[\hat{X}]$. This increase indicates that the latent representation produced by the model reinforces the underlying distributional structure, making semantic relations between terms more salient. In particular, the higher correlations observed in the reconstructed matrix suggest that dimensionality reduction mitigates noise due to accidental co-occurrences and highlights shared semantic patterns.

Overall, the results confirm the effectiveness of the two-step procedure, cosine-based selection followed by correlation-based validation, in discriminating genuine semantic relations within a specialised lexicon, thereby improving the reliability of synonym detection and, more broadly, of semantic association identification among terms. The method appears less effective in the case of generic lexemes or items that are strongly dependent on the discursive structure of the text. In such contexts, the results do not yield synonymy or antonymy in the strict lexicographic sense, but rather configurations of contextual proximity whose interpretation depends on the actual usage of the terms within the corpus. In line with the objective of preserving usage-based distinctions between formally related lexical items, no lemmatisation was applied to the corpus. This methodological choice allows semantic and functional nuances to remain observable, preventing the loss of information that would result from morphological normalisation. A significant example is provided by the pair *Paese/Italia*, which can be interpreted as a form of contextual near-synonymy. Their proximity appears to stem from the fact that, in the analysed texts, *Italia* is frequently treated as the reference country. This closeness does not imply full semantic equivalence: *Italia* typically occurs in comparative or supranational contexts, whereas

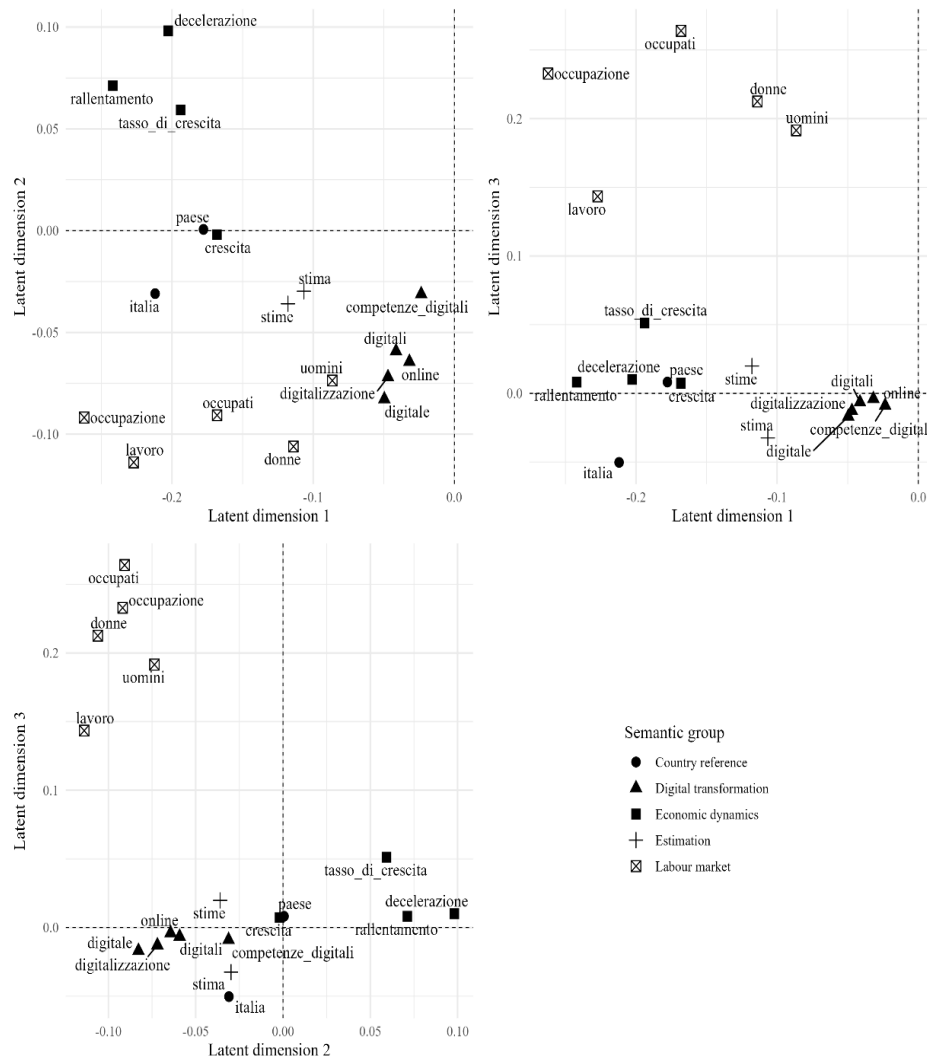
Paese is more often used in general statements. The result reflects a plausible context-dependent relation rather than strict synonymy. A similar pattern emerges with *stima/stime*. Despite their morphological relatedness, the two forms display distinct functional distributions: *stima* tends to occur in methodological or theoretical contexts, while *stime* is more closely associated with operational or statistical discourse. Their proximity therefore captures differentiated usage rather than interchangeability. These cases underscore the relevance of the decision not to lemmatise the corpus. Morphological normalisation would have neutralised the semantic contribution of the specific forms: in the first instance, collapsing singular and plural would obscure their functional differentiation; in the second, reducing *Paese* to its lemma, and potentially merging it with *paesi*, would weaken the contextual relation with *Italia*. Overall, the findings suggest that retaining surface forms preserves fine-grained meaning distinctions tied to actual usage, which would otherwise be lost through lemmatisation, thereby supporting the suitability of a non-lemmatised corpus for analyses aimed at capturing subtle semantic variation.

4.2. Interpretation of the latent space

The interpretation of the latent space obtained through LSA confirms and further develops the patterns already observed in the correlation matrix. The first dimensions should not be understood as directly observable semantic categories, but rather as interpretive axes reconstructed ex post based on the terms displaying the strongest loadings. From this perspective, their main interest lies in their ability to make visible the principal semantic-discursive structures that organise the corpus. The first dimension appears to function as a general macroeconomic conjunctural factor, dominated by terms such as *prezzi*, *esportazioni*, *trimestre*, *pil*, *domanda*, *importazioni* and *occupazione*. Rather than opposing two distinct poles, it seems to gather the basic lexical repertoire through which the corpus describes overall economic trends. It is significant that terms such as *Italia*, *Paese*, *crescita*, *decelerazione* and *rallentamento* are also located along this same axis, confirming their membership in a common discursive space. The subsequent dimensions are more clearly structured. The second dimension distinguishes the lexicon of public finance and taxation from that of foreign trade and prices; the third opposes macroeconomic aggregates to the lexicon of the labour market and the socio-territorial dimension; and the fourth separates the domain of inflation and domestic prices from that of export performance. Taken together, these axes show that the model organises terms not only according to lexical similarity, but also across different thematic and analytical planes. The fifth dimension prominently captures a digital transformation lexicon (e.g., *digitale*, *digitali*, *digitalizzazione*, *competenze_digitali*, *online*), indicating that the latent space can capture not only macroeconomic clusters but also specific, semantically cohesive subdomains. Overall, the analysis of the first dimensions confirms that LSA is particularly effective in representing domain-specific semantic clusters, whereas it appears less informative in the case of generic or metadiscursive terms. The latent space, therefore, does not merely reveal relations of synonymy or antonymy, but makes it possible to reconstruct the principal semantic-discursive structures that organise the specialised corpus. The three two-dimensional projections of the selected terms onto the first three dimensions of the latent space show that LSA organises the lexicon not so much on the basis of strict synonymy as according to areas of semantic-discursive proximity (Laham, 1997). Across all representations, relatively stable clusters emerge corresponding to three main domains: economic dynamics (*crescita*, *rallentamento*, *decelerazione*, *tasso_di_crescita*), digital transformation (*digitale*, *digitali*, *digitalizzazione*, *competenze_digitali*, *online*), and the labour market (*lavoro*, *occupazione*, *occupati*, *donne*, *uomini*). Their separation in space confirms that the model effectively

distinguishes specialised semantic fields, whereas more cross-cutting terms, such as Italia and Paese, tend to occupy intermediate positions, consistent with a function that is less rigidly anchored to a single domain.

Graph 1. Projection of selected semantic groups onto Latent dimensions 1, 2, 3



The plots also make it possible to refine certain findings already suggested by the correlation-matrix analysis. The relative proximity between Italia and Paese suggests a form of contextual near-synonymy, though not full semantic equivalence; similarly, the arrangement of crescita, rallentamento and decelerazione shows that the model captures the unity of the semantic field of economic dynamics more than sharply defined binary oppositions. Finally, the only partial proximity between stima and stime confirms that morphological similarity does not automatically imply identity of use. Taken together, the projections therefore reinforce the interpretation according to which LSA is especially effective in identifying domain-related proximity and semantic families – synonyms, antonyms, and contextual near-synonymies – while proving less informative for generic or weakly specialised terms.

4.3. Analysis of the diachronic dimension

The diachronic dimension of the analysis makes it possible to assess whether, over the period considered and within the domain-specific terminology represented in Chapter 1 of the ARs,

any significant linguistic changes have occurred, such as shifts in the preferred use of specific terms, the introduction of new terminology, or the disappearance of existing lexical items. To investigate possible shifts in usage preferences over time, the word-word correlation coefficients computed on the reconstructed matrix were examined. The lowest values in the matrix range from -0.40 to -0.48, indicating the absence of substantial changes in usage patterns. Accordingly, the core relationships among terms appear to have remained relatively stable across the editions considered. These findings suggest that a time span of little more than twenty years, within such a highly specialised linguistic domain, is not sufficient to reveal significant shifts in terminological preferences. Since the first edition of the ARs dates back only to 1992, this publication does not appear to be especially well suited to this type of diachronic investigation. By contrast, a study aimed at detecting long-term lexical change might prove more productive if applied to documentary series with greater historical depth, such as Istat's Italian Statistical Yearbook, whose editorial tradition extends over nearly one hundred and fifty years. Regarding the other two types of variation considered, the analysis of the DTM highlights terms that, within the context of the publication, have fallen out of use, as well as terms newly introduced in specific years. One example is provided by *indagine/indagini* and *rilevazione/rilevazioni*, which refer to one of Istat's core activities. In the first five editions of Chapter 1 examined here (covering the years 2001–2011), these terms occur 65 times, whereas in the subsequent five editions (2013–2023) they occur only 11 times. This marked decline may be interpreted as reflecting a redefinition of the publication's aims: while earlier editions were primarily oriented towards presenting the most significant results emerging from statistical surveys, later editions became more strongly focused on macroeconomic analysis, with particular attention to the European context and to developed countries. An opposite trajectory can be observed for terms related to digital innovation, such as *digitalizzazione*, *digitale/digitali*, *competenze digitali*, and *online*, which begin to appear only from the Annual Report referring to 2021 onwards, that is, only in the last two editions included in the analysis. This development is linked to the transformation triggered in the country following the COVID-19 pandemic, which reshaped social and economic policies and promoted the large-scale adoption of interoperable forms of remote interaction. Overall, over the last twenty years usage patterns have remained broadly stable, while the lexical repertoire of the domain has undergone continuous yet localised adjustments.

4.4. Detection of polysemy

The LSA approach adopted in this study does not enable the direct detection of polysemy. In LSA, each term is represented by a single vector in the term-space matrix produced through SVD factorisation, and this representation does not vary according to the specific context in which the term occurs. In other words, all instances of a term are aggregated into a single point in semantic space, representing an average of its possible shades of meaning. For this reason, LSA can capture polysemy only partially, and only indirectly, insofar as different occurrences of a term may generate different statistical distributions across lexical clusters; however, it cannot explicitly distinguish the different senses of a word within a corpus. Nevertheless, examination of the DTM revealed a significant case. The term *transizione* occurs in 2001 and 2003 with reference to the transition from the lira to the euro, whereas in 2021 and 2023 it is used in a digital context. While in general language this word functions as a broad top-concept that simply adapts to different contexts, within this sector-specific language it crystallises into distinct, narrow meanings. This semantic shift points to a case of polysemy within the specialised domain under analysis. The observation emerged from the diachronic perspective, since the term disappears for a period (from 2006 to 2018) and subsequently re-emerges with a

different, uniquely defined meaning. The systematic and contextual identification of polysemy requires different semantic analysis tools, such as models that generate context-sensitive embeddings. These models do not assign a single vector representation to each word; rather, they produce different vectors for each occurrence in context, thereby allowing an implicit differentiation of word senses according to the surrounding contextual information. This principle underlies pre-trained contextualised embedding models such as ELMo and BERT, which derive a dynamic representation of a term’s meaning by modulating it according to the sentence in which it appears. Recent literature has shown that static models such as LSA and LDA suffer from what has been termed ‘meaning conflation deficiency’, that is, the tendency to collapse all possible interpretations of a word into a single vector, thereby limiting their ability to distinguish complex cases of polysemy.

5. Conclusion

The analysis confirms that LSA, applied through both the word-word and reconstructed correlation matrices, effectively identifies semantic relations – including synonyms, antonyms, and associative links – within a specialised lexicon. The approach is less effective for generic or weakly specialised items. The examination of latent dimensions further highlights the emergence of semantic families, grouping terms that co-occur across multiple factors. However, the method shows limitations for diachronic analysis: the twenty-year span of the corpus proved insufficient to capture significant shifts in term usage, and negative correlations did not provide meaningful insights. These limitations also suggest two possible methodological extensions. First, diachronic variation could be explored by performing separate LSAs on temporally distinct subcorpora, for example by contrasting the earliest period, 2001–2006, with the most recent one, 2018–2023. This would reduce the risk that temporal effects are hidden by the dominant structural features of a global LSA model and would allow word associations and distance patterns to be compared across periods. Second, cluster analysis could be applied to the term vectors or distance matrices obtained from LSA, to identify broader semantic families and assess their stability or transformation over time. Similarly, the systematic identification of polysemy requires context-sensitive models, as static approaches like LSA are unable to differentiate word senses in different contexts.

References

- Benoit K., Watanabe K., Wang H., Nulty P., Obeng A., Müller S., and Matsuo A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30): 774. <https://doi.org/10.21105/joss.00774>
- Landauer T.K. and Dumais S.T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2): 211–240. <https://doi.org/10.1037/0033-295X.104.2.211>
- Landauer T.K., Foltz P.W. and Laham D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25: 259–284. <https://doi.org/10.1080/01638539809545028>
- Laham D. (1997). Latent semantic analysis approaches to categorization. In Shafto M.G. and Langley P., editors, *Proc. of the 19th Annual Conference of the Cognitive Science Society*: pp. 979–984.
- Ooms J. (2026). hunspell: High-performance stemmer, tokenizer, and spell checker (R package, Version 3.0.7). <https://docs.ropensci.org/hunspell>
- ISO (2022). ISO 704:2022 – Terminology work – Principles and methods. Geneva: International Organization for Standardization.

A hybrid rule-based and data-driven pipeline for author name normalization in IRIS institutional repositories

Pasquale Pavone^{1,3}, Simone Righi^{2,3} Sara Colombini³

¹ Department of Literary, Linguistic and Philosophical Studies, Pegaso University –
pasquale.pavone@unipegaso.it

² Department of Economics Marco Biagi, University of Modena and Reggio Emilia –
simone.righi@unimore.it

³ Centro di Analisi delle Politiche Pubbliche, University of Modena and Reggio Emilia –
sara.colombini@unimore.it

Abstract

Author name normalization is a crucial preprocessing step for bibliometric analysis, research evaluation, and the study of scientific collaboration networks, especially when data originate from heterogeneous and weakly standardized institutional repositories. This paper presents a Python-based, conservative pipeline for the normalization of author strings in bibliographic records extracted from IRIS/CRIS repositories, with a case study focused on the Department of Economics at the University of Modena and Reggio Emilia. The proposed approach addresses recurrent sources of syntactic heterogeneity in author fields, including inconsistent separators, capitalization variants, initials, orthographic variants, onomastic particles, compound surnames, and given-name/surname inversions. The pipeline combines explainable rule-based preprocessing and parsing with a corpus-adaptive expansion of initials based on internal frequency evidence and a final majority-based canonicalization of inverted full-name pairs applied after frequency stabilization. Unlike full author-name disambiguation systems, the method does not aim to resolve author identity through external or relational metadata such as ORCID identifiers, affiliations, co-author networks, venues, titles, or topics. Rather, it provides an auditable preprocessing layer that consolidates surface variants only when the available corpus evidence supports a reliable decision, while explicitly retaining unresolved cases when evidence is insufficient or ambiguity remains. Applied to 10,271 publication records and 34,746 author–publication pairs, the pipeline reduces syntactic heterogeneity and produces traceable outputs for downstream bibliometric analyses. The revised evaluation complements internal outcome counts with manual validation and sensitivity analysis of the majority threshold, thereby clarifying the precision–coverage trade-off of the proposed approach. The contribution lies in providing a transparent, reproducible, and adaptable normalization workflow for institutional repository data, designed to improve metadata quality before subsequent bibliometric or science-mapping analyses.

Keywords: Bibliographic metadata, Text preprocessing, Institutional repositories, Author name normalization, IRIS DB.

1. Introduction

In recent years, Current Research Information Systems (CRIS) have become a central infrastructure for managing the research lifecycle in universities. They collect and integrate metadata on publications, projects, patents, and researcher profiles, with direct implications for both reporting practices and policies on dissemination and open science. In Italy, a particularly relevant role is played by IRIS (Institutional Research Information System), a CRIS platform developed by CINECA and widely adopted nationwide, designed to support the integrated management of research information and its alignment with institutional processes (Bollini et al., 2016). The widespread adoption of CRIS platforms makes databases such as IRIS

potentially useful not only for administrative and evaluative purposes, but also for exploratory analyses and bibliometric studies. The availability of structured metadata, including output types, authors, publication venues, years, affiliations, and related fields, makes it possible to reconstruct scientific profiles, collaboration dynamics, and thematic trajectories at different levels, such as individual researchers, groups, departments, and institutions. Furthermore, analyzing IRIS gives a view on what local and national evaluators typically observe. However, as is often the case in complex information systems that involve automated imports, manual entries, migrations from legacy systems, and subsequent corrections, metadata quality is not homogeneous and can substantially affect the reliability of downstream analyses.

Beyond national-scale bibliometric databases, IRIS enables local, institution-specific analyses, for instance at the level of a department or a single university, where analysts often need transparent and reproducible preprocessing steps rather than opaque disambiguation procedures. In this context, author-string heterogeneity becomes a critical bottleneck, as it can quickly undermine the reliability of collaboration networks and descriptive indicators.

In particular, the literature has shown that author data quality is a critical issue. Inconsistencies and variants in name strings can distort co-authorship networks and descriptive metrics, producing duplicated nodes, artificial fragmentation of authors, and biased centrality and community structures, with relevant interpretive consequences (De Stefano et al., 2023). This concern is consistent with a well-established line of work in record linkage theory. When identifiers are not unique and records are not perfectly structured, cleaning and normalization become essential to reduce systematic errors (Fellegi & Sunter, 1969).

Among the most problematic fields in institutional repositories, the representation of authors is often the most critical source of downstream distortion. Even when the system provides procedures for recognition or linkage to internal profiles, in practice a non-negligible share of authors can remain not fully structured and highly heterogeneous in form, therefore not uniquely linked, especially due to the presence of co-authors external to the department or the institution, whose information is frequently entered manually according to non-uniform conventions. This leads to recurring phenomena such as orthographic variants, inconsistent use of initials, given-name and surname inversions, multiple separators, onomastic particles, and cases of homonymy. In the absence of systematic unique identification, variability in the author field can directly affect the quality of datasets exported from IRIS and, consequently, the exploratory or bibliometric analyses conducted on those exports.

In parallel, the scientometrics and informetrics community has proposed many approaches to Author Name Disambiguation (AND), ranging from rule-based solutions and string matching, to probabilistic record linkage, to graph-based and machine learning approaches, including recent review and systematization contributions (Cappelli et al., 2025; De Bonis et al., 2023; Ferreira et al., 2012; Shamly & Subaveerapandiyam, 2026). In addition, the use of persistent identifiers can reduce ambiguity, but adoption is not complete and metadata coverage remains uneven across many institutional databases (Meadows & Haak, 2018).

Although AND toolkits and benchmarks developed on large bibliographic datasets and designed for advanced ML workflows are available, these tools often assume relatively rich and standardized metadata and do not automatically resolve the syntactic idiosyncrasies that are typical of exports from institutional CRIS platforms (Kim & Kim, 2024; Subramanian et al., 2021). In this context, there remains a need for tools that can make data analysis-ready without relying on external integrations or non-reproducible procedures.

In this work, we propose a Python pipeline for conservative author-name normalization in bibliographic records extracted from IRIS, initially applied to the Department of Economics of the University of Modena and Reggio Emilia. The aim is not to solve full author-name disambiguation, but to consolidate syntactically heterogeneous author strings when this can be done on the basis of explicit rules and corpus-internal evidence. The pipeline is therefore conceived as an auditable preprocessing layer that improves the quality and consistency of author metadata before downstream bibliometric or science-mapping analyses.

The approach is deliberately hybrid and explainable. It combines (i) rule-based, linguistically aware parsing, handling commas, capitalization, particles, and initials, (ii) data-driven expansion of initials based on internal corpus evidence with a majority threshold, and (iii) final canonicalization of given-name and surname inversions via majority swap, applied only after counts have been stabilized. Rather than replacing bibliometric analysis frameworks, the pipeline is designed to prepare heterogeneous IRIS author fields for more reliable downstream analyses by reducing syntactic variability, preserving traceability, and avoiding unsupported identity assignments (Aria & Cuccurullo, 2017).

2. Essential background and positioning of the contribution

Author Name Disambiguation (AND) has been extensively studied in bibliometrics and digital libraries, as name ambiguity affects attribution, indicators, and co-authorship network reconstruction. Existing solutions span rule-based normalization and string matching, probabilistic record linkage, and more recent graph-based and machine learning approaches, including deep learning pipelines evaluated on large-scale scholarly datasets (Cappelli et al., 2025; De Bonis et al., 2023; Fellegi & Sunter, 1969; Ferreira et al., 2012; Shamly & Subaveerapandiyam, 2026). Persistent identifiers such as ORCID mitigate ambiguity, but adoption and metadata completeness remain uneven across institutional repositories, so name-based normalization is still required in many real-world CRIS exports (Meadows & Haak, 2018). In parallel, recent benchmarks and open-source toolkits have accelerated AND research and evaluation, often targeting web-scale settings and assuming relatively rich, standardized metadata (Kim & Kim, 2024; Subramanian et al., 2021). Institutional CRIS data, however, frequently exhibit upstream heterogeneity in author strings, including inconsistent separators, initials, and systematic given-name and surname inversions, which can propagate as artifacts into downstream science mapping workflows (Aria & Cuccurullo, 2017; De Stefano et al., 2023). Therefore, the present contribution focuses on a reproducible preprocessing layer for IRIS exports, combining explainable rule-based parsing with a constrained, auditable data-driven component, designed to stabilize author representations before downstream analyses.

3. Data and application context

The case study relies on bibliographic records extracted from IRIS (Institutional Research Information System), a CRIS platform widely adopted by Italian universities to collect and manage research outputs (Bollini et al., 2016). The empirical application focuses on the Department of Economics “Marco Biagi” at the University of Modena and Reggio Emilia. The elementary unit of analysis is a publication record, associated with a “paper_id” (or an equivalent identifier) and an author field stored as a textual string.

In the dataset analysed here, the input consists of 10,271 publication records, corresponding to 7,368 distinct publications in IRIS, spanning the period 1974-2025. Before normalization, the

initial extraction yields 34,746 author–publication pairs (author tokens), which represent the main source of heterogeneity for downstream analyses.

The author field exhibits several recurring issues, often in combination, including multiple and inconsistent separators, alternating surname and given-name order, inconsistent use of initials and punctuation, the presence of onomastic particles and compound surnames, and cases of genuine or near homonymy that require conservative decisions. These patterns are not marginal noise. Even a limited number of systematic inconsistencies can inflate the number of distinct author entities, generate duplicated nodes in co-authorship networks, and produce artificial fragmentation of authors, ultimately biasing descriptive statistics and collaboration structures (Aria & Cuccurullo, 2017; De Stefano et al., 2023).

4. Methodology: a hybrid rule-based and data-driven pipeline

The proposed procedure is designed as an analysis-ready preprocessing layer for author metadata extracted from IRIS. It integrates three tightly coupled components: rule-based and linguistically aware normalization of author strings, data-driven expansion of initials based on internal corpus evidence, and final canonicalization of given-name and surname inversions after frequency counts have been stabilized. The overall strategy is deliberately conservative. Whenever the available evidence does not support a reliable decision, the pipeline avoids arbitrary assignments and explicitly records cases that remain non-expanded.

Throughout the paper, we use the term normalization to refer to the conservative consolidation of syntactically different author-string variants. This differs from full Author Name Disambiguation, which aims to infer real-world author identities and typically requires richer metadata, such as persistent identifiers, affiliations, co-author relations, venues, titles, or topical information. Our pipeline deliberately operates at the author-string level and leaves unresolved any case for which corpus-internal evidence is insufficient.

To make the rule ordering and decision logic explicit, the pipeline follows a fixed sequence of conservative operations. First, superficial variants are removed by normalizing spacing, punctuation, and delimiters, without inferring name structure. Second, author lists are segmented by giving priority to semicolons and, only when they are absent, by reconstructing repeated surname/given-name patterns in multiple-comma strings. Third, individual tokens are standardized into either “Surname, Given” or “Surname, Initials”, using commas, capitalization cues, particles, compound surnames, and explicit initial patterns. Fourth, initials are expanded only when corpus-internal evidence provides either a unique candidate or a majority candidate above the selected threshold. Finally, inverted full-name pairs are canonicalized only after frequency counts have been stabilized. At each stage, irregular or weakly supported cases are not forced into a canonical form but are retained as non-expanded or unresolved outputs for manual inspection.

4.1 Input representation, preliminary cleaning, and author list segmentation

The input is a tab-delimited file in which each row links a publication identifier (“paper_id”) to a textual author field containing one or more authors. Before any parsing decision, the pipeline applies a lightweight but crucial cleaning stage aimed at removing superficial variability that would otherwise prevent deterministic matching. This includes normalizing spacing around punctuation, harmonizing punctuation variants, and ensuring consistent formatting of delimiters. The purpose is not to infer name structure at this stage, but to ensure that equivalent

strings become comparable, for instance, by bringing variants such as “Pavone ,Pasquale” and “Pavone, Pasquale” to the same surface form.

A key challenge in IRIS exports is that separators are not always used consistently, and commas may serve either as intra-author separators (surname, given name) or as inter-author delimiters in malformed strings. For this reason, segmentation is performed with a priority strategy. When a semicolon is present, it is treated as the primary delimiter, since it typically separates authors reliably in CRIS exports. Thus, a string like “Pavone, Pasquale; Righi, Simone” is directly segmented into two author tokens.

When semicolons are absent and multiple commas appear, the pipeline avoids naïve splitting and instead attempts pattern-based reconstruction. In practice, it detects repeated “surname, given-name/initials” pairs and reconstructs author tokens accordingly. This allows a string such as “Pavone, Pasquale, Righi, Simone” to be interpreted as two authors rather than four fragments. A further source of heterogeneity is the presence of concatenated author sequences with weak separators, which are sometimes recognizable only through initial patterns. The segmentation component includes specific handling for these cases, reconstructing minimal informative representations (e.g., “Pavone, P.” and “Righi, S.”), while leaving the full expansion of initials to a later evidence-based step.

4.2 Rule-based normalization of individual author tokens

After segmentation, each author token is normalized into a canonical form. The pipeline aims to produce “Surname, Given” when a full given name is available, or “Surname, Initials” when only initials are present. Parsing is rule-based but linguistically aware, in the sense that it exploits strong formal cues (the comma, capitalization patterns typical of institutional exports, and explicit initial patterns) and includes dedicated handling for particles and multiword surnames.

When a comma is present and the token matches a well-formed “Surname, Given” structure, the interpretation is straightforward and the output is standardized through consistent titlecasing. For instance, “PAVONE, PASQUALE” and “Pavone, Pasquale” are unified as “Pavone, Pasquale”. For abbreviated forms, the pipeline enforces a consistent punctuation convention for initials, so that variants such as “Righi, S”, “Righi, S.”, and “Righi, S .” converge to “Righi, S.”. Similarly, multi-initial blocks are normalized to a stable representation, bringing “ML”, “M.L.” and “M. L” to “M. L.”. This step is essential, since many downstream operations depend on comparing initials-based signatures reliably.

Tokens without commas are handled through additional rules based on patterns that are common in CRIS exports, for example “SURNAME Given” formats driven by capitalization cues. When ambiguity is unavoidable, the pipeline remains conservative, preferring a minimally standardized output over a risky inversion that may generate false positives.

4.3 Data-driven initial expansion with a majority threshold and a unique-mapping single-initial step

The data-driven component addresses abbreviated author forms such as “Pavone, P.” or “Righi, S.” by leveraging internal corpus evidence. The pipeline builds an index that maps each pair (surname, initials signature) to the set of observed full given names and their frequencies in the corpus. The initials signature is defined as the ordered sequence of given-name initials extracted after punctuation normalization, so that variants such as “ML”, “M.L.”, and “M. L.” are treated

as the same signature. For instance, if “Righi, Simone” is observed multiple times and “Righi, S.” appears elsewhere, the index captures that the signature “S” associated with surname “Righi” co-occurs with the full given name “Simone” with a given frequency.

Initial expansion follows transparent decision rules. When a (surname, signature) pair has a single candidate full given name in the index, the abbreviated form is expanded directly. When multiple candidates exist, expansion is applied only when the top candidate accounts for at least 60% of all compatible occurrences in the corpus (conservative configuration); otherwise, the case is retained as unresolved and recorded for audit.

The majority threshold is a configurable parameter. In addition to the 0.60 setting used in this work, the pipeline can be executed with the majority filter disabled, producing a separate “no-threshold” output that maximizes coverage at the cost of potentially higher ambiguity, while preserving full traceability of the decisions made.

Finally, the pipeline includes a targeted unambiguous single-initial step. It constructs a mapping (surname, first initial) $\rightarrow$ full given name only when the corpus supports a single candidate for that combination. This step improves coverage while remaining conservative, since it expands only those cases supported by uniquely identifying evidence.

4.4 Canonicalization of given-name and surname inversions and audit outputs

A recurring issue in institutional exports is the presence of the same two tokens in inverted order, as in “Pavone, Pasquale” versus “Pasquale, Pavone”. In the finalized workflow, inversion handling is performed through majority-based canonicalization, applied at the appropriate stage to avoid unstable decisions. Specifically, after initial expansion and the unambiguous single-initial step, the pipeline applies a majority swap on inverted full-name pairs, restricted to simple cases with one token on each side of the comma. The swap is applied only to full-name pairs (excluding initials-only forms), and only after initial expansion to ensure stable frequency estimates. The most frequent orientation is selected as the canonical form, and occurrences of the inverted variant are rewritten accordingly. Applying this step only after counts have been stabilized is crucial, because abbreviated forms such as “Pavone, P.” must be consolidated first, otherwise the observed frequencies of the full-name variants may be artificially distorted. As a result, inverted and abbreviated surface forms can be consolidated into a single canonical author representation, reducing duplicate entities and preventing artificial fragmentation in co-authorship networks (De Stefano et al., 2023).

To ensure traceability, the pipeline produces four main artifacts, namely “authors_parsed.csv”, “authors_expanded.csv”, “authors_unresolved.csv”, and a per-publication summary report. Each artifact carries a source label and a trace note that document the decision path, enabling full auditability and supporting both methodological transparency and practical reuse.

5. Results

This contribution is primarily methodological, with the aim of presenting a reproducible and auditable pipeline rather than a large-scale benchmark evaluation. Nevertheless, even on a departmental CRIS export, the procedure produces measurable outcomes that are directly relevant for downstream exploratory and bibliometric analyses, because they quantify how much heterogeneity in author strings is removed and how much ambiguity remains after applying conservative decision rules.

5.1 Dataset-level outcomes

In the dataset analysed here, the input consists of 10,271 publication records, corresponding to 7,368 distinct publications in IRIS. After splitting the author field, the initial extraction yields 34,746 author–publication pairs. After deduplication at the author–publication level, the pipeline processes 24,096 unique author tokens, corresponding to a reduction of 10,650 duplicates (30.65%). Here, deduplication refers to collapsing identical author–publication pairs that arise from repeated or redundant author strings in the raw export, so that each unique author token is counted once per publication. For completeness, we also produce a no-threshold output, intended for exploratory scenarios where coverage is prioritized; the results reported here refer to the conservative configuration (threshold = 0.60).

Table 1. Author name normalization outcomes: counts and percentages of author–publication pairs.

Outcome labels	Author tokens (n)	Share of author tokens (%)
as_is_fullname	11,003	45.66%
expand:majority	229	0.95%
expand:single	3,776	15.67%
non_expanded (no_evidence)	9,036	37.50%
non_expanded (ambiguous candidates)	19	0.08%
missing_given_name	33	0.14%

As shown in Table 1, out of 24,096 author tokens, 11,003 (45.66%) were already available as full names, 3,776 (15.67%) were expanded through a single-candidate rule, and 229 (0.95%) were expanded through the majority rule. The remaining cases were conservatively retained as non-expanded, almost entirely because no corpus evidence was available for expansion (9,036 tokens, 37.50%), while genuinely ambiguous cases with competing candidates were rare (19 tokens, 0.08%). Finally, only 33 tokens (0.14%) lacked a usable given name (“missing_given_name”), representing the main source of potentially noisy nodes for downstream analyses. Non-expanded labels do not indicate parsing failure; they capture either conservative non-expansion due to missing evidence or, in a small minority of cases, genuine ambiguity among competing candidates.

Before initial expansion and final canonicalization, the dataset contained 10,169 distinct author strings (“author_norm”), which decreased to 9,430 after the full workflow (−7.27%), with 26 inverted full-name occurrences rewritten via the final majority-swap step.

As regards initials-based tokens, in the conservative configuration 3,776 abbreviated author tokens were expanded via the single-candidate rule and 229 via the majority rule, while 9,055 remained non-expanded due to insufficient evidence. These outcomes correspond to 28.91%, 1.75%, and 69.33% of all initials-based tokens, respectively, providing a compact summary of the precision–coverage trade-off enforced by the chosen configuration.

5.2 Manual validation and threshold sensitivity

To complement the internal outcome counts, we carried out a manual validation on a stratified sample of author tokens covering the main decision labels produced by the pipeline: rule-based full-name normalizations, single-candidate expansions, majority-based expansions, unresolved cases, unique single-initial expansions, malformed or incomplete strings, and inverted full-name pairs rewritten by the majority-swap step. Each sampled case was manually inspected to assess whether the pipeline decision was consistent with the available author-string evidence

and with the internal rules of the workflow. The validation was not intended to certify real-world author identity, which would require external metadata, but to provide a targeted quality check of the main normalization decisions.

The “as_is_fullname” group was not treated as a set of already clean records. Rather, it includes cases in which cleaning, separator handling, casing normalization, and rule-based parsing produced a complete Surname, Given representation without requiring initial expansion. The validation also distinguishes between “expand_single”, which refers to expansions based on a unique compatible candidate for the complete initials signature, and “expand_unique_single_initial”, which refers to a targeted rule applied to unresolved single-initial forms when the same surname and first initial identify a unique full-name candidate in the corpus. For “expand_majority”, validation assessed consistency with corpus-internal majority evidence rather than external identity verification.

Table 2. Manual validation of pipeline decision groups.

Decision group	Sample d cases	Validated / consistent	Error s	Main issue
as_is_fullname	40	40	0	—
expand_single	59	59	0	—
expand_majority	55	55 consistent with internal evidence	0	identity not externally verifiable
expand_unique_single _initial	23	23	0	—
unresolved	60	60 correctly unresolved	0	insufficient evidence
majority_swap	23	23	0	—
no_given	20	20 correctly flagged	0	malformed/incomplete strings
Total	280	280	0	

Manual inspection did not reveal false merges, false splits, or incorrect inversions in the sampled cases. The results confirm that the conservative strategy avoids unsupported assignments and makes residual uncertainty explicit. In particular, unresolved cases were correctly left unexpanded when author-string evidence was insufficient, while “no_given” cases corresponded to malformed or incomplete strings correctly flagged by the pipeline. Majority-based expansions should therefore be interpreted as decisions consistent with corpus-internal evidence, not as externally certified author identity resolutions.

A comparison with full author-name disambiguation toolkits was not performed because the available IRIS export lacks the metadata required by such systems, including ORCID identifiers, stable internal person identifiers, affiliations, e-mail addresses, co-author histories, venues, titles, and topical information. This limitation is partly due to the historical depth of the archive, which includes older records with missing or incomplete structured author metadata.

We also assessed the sensitivity of the majority-based expansion step to different threshold values. As shown in Table 3, lowering the threshold to 0.50 slightly increases the number of majority-based expansions, but admits decisions supported by weaker evidence. Conversely, increasing the threshold to 0.90 substantially reduces coverage, while retaining only cases supported by very strong internal evidence. The 0.60 threshold was retained as a conservative operational setting because it excludes the weakest majority decisions accepted at 0.50, while preserving substantially higher coverage than the very strict 0.90 configuration.

Table 3. Sensitivity analysis of the majority threshold for initial expansion.

Majority threshold	Majority expansions	Ambiguous unresolved	Mean majority share
0.50	245	3	0.89
0.60	229	19	0.92
0.75	218	19	0.93
0.90	139	19	0.98

5.3 Residual ambiguity

In the analysed dataset, 9,055 author tokens (37.58% of all author–publication pairs) remain non-expanded under the conservative configuration. This reflects a deliberate design choice: initials are expanded only when corpus evidence supports a reliable and auditable decision. From an operational perspective, it is crucial to distinguish two non-expanded subtypes with very different implications. The vast majority of non-expanded cases are classified as “non_expanded (no_evidence)” (9,036 cases). These instances are non-expandable because no full-name candidates are observed in the corpus for the same (surname, initials signature), a situation that typically occurs for external co-authors recorded only through initials. Importantly, these forms remain fully usable as consistent author labels for downstream analyses, including co-authorship network reconstruction.

In contrast, the only cases that reflect genuine residual ambiguity are those classified as “non_expanded (ambiguous candidates)” (19 cases, 0.08% of all author–publication pairs), where multiple competing full-name candidates exist and none reaches the majority threshold. These cases cannot be safely resolved using author strings alone and therefore mark the boundary beyond which full author identity resolution would require additional metadata, such as persistent identifiers, affiliations, co-author relations, venues, titles, or topical information. Finally, the main source of potentially noisy entries is concentrated in the very small “missing_given_name” subset (33 cases, 0.14%), which captures incomplete or malformed tokens lacking a usable given name or initials and may generate uninterpretable nodes unless filtered or manually inspected.

Overall, these results provide a minimal but robust empirical characterization of the pipeline’s effect on IRIS exports. They quantify consolidation while making residual non-expansions and genuinely ambiguous cases explicit and measurable, thereby establishing a clear baseline for scaling the approach to larger institutional datasets and for future extensions that incorporate additional metadata signals.

6. Conclusion and future work

We presented a hybrid pipeline, combining rule-based and data-driven components, for author-name normalization in heterogeneous bibliographic corpora extracted from IRIS. The main methodological contribution is a CRIS-oriented, auditable preprocessing layer that bridges rule-based normalization and corpus-adaptive consolidation, integrating linguistically aware and explainable parsing, evidence-based initial expansion under a configurable majority threshold, and a final majority swap for systematic given-name and surname inversions applied only after frequency stabilization. The workflow is designed as a robust preprocessing layer for downstream bibliometric and exploratory analyses, with particular attention to reproducibility, transparency, and conservative decision-making. The results show that, even in a departmental

setting, the pipeline can reduce duplication and inconsistency in author metadata, providing an initial empirical basis for extending the approach to larger institutional datasets and for releasing a reusable Python library. Concretely, on the departmental IRIS export, the workflow reduced the number of distinct normalized author strings from 10,169 to 9,430 (−7.27%) and removed 10,650 duplicate author–publication pairs during deduplication (−30.65%). Moreover, non-expanded cases are overwhelmingly due to missing corpus evidence for expansion, while genuinely ambiguous candidates are rare (19 cases, 0.08%), confirming that the conservative strategy primarily avoids unsupported expansions rather than failing to parse the data.

Future work will extend the application beyond the current departmental dataset to the full University of Modena and Reggio Emilia, enabling a broader assessment across disciplines and potentially more heterogeneous author populations. In parallel, the codebase will be packaged as an open-source Python library, with modular components for parsing, normalization, initial expansion, and canonicalization, and with configurable parameters to support different precision-versus-coverage trade-offs. Finally, where available, future versions may integrate persistent identifiers and controlled lookups to external services, while preserving reproducibility and keeping the rule-based and auditable design principles as a core requirement (Meadows & Haak, 2018). Such extensions would move beyond string-level normalization toward richer identity-resolution workflows, while keeping the present pipeline as a transparent preprocessing layer.

References

- Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975.
- Bollini, A., Mennielli, M., Mornati, S., & Palmer, D. T. (2016). Iris: Supporting & managing the research life-cycle. *Universal Journal of Educational Research*.
- Cappelli, F., Colavizza, G., & Peroni, S. (2025). Deep learning approaches to author name disambiguation: A survey. *International Journal on Digital Libraries*, 26(4), 1–17.
- De Bonis, M., Falchi, F., & Manghi, P. (2023). Graph-based methods for Author Name Disambiguation: A survey. *PeerJ Computer Science*, 9, e1536.
- De Stefano, D., Fuccella, V., Vitale, M. P., & Zaccarin, S. (2023). Quality issues in co-authorship data of a national scientific community. *Network Science*, 11(1), 98–112.
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210.
- Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record*, 41(2), 15–26.
- Kim, J., & Kim, J. (2024). ANDez: An open-source tool for author name disambiguation using machine learning. *SoftwareX*, 26, 101719.
- Meadows, A., & Haak, L. (2018). How persistent identifiers can save scientists time. *FEMS Microbiology Letters*, 365(15), fny143.
- Shamly, H. A. H. E., & Subaveerapandiyar, A. (2026). Author Name Disambiguation in Scholarly Research: A Bibliometric Perspective. *Open Information Science*, 10(1), 20250035.
- Subramanian, S., King, D., Downey, D., & Feldman, S. (2021). S2and: A benchmark and evaluation system for author name disambiguation. *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 170–179.

From Knowledge Gaps to Scientific Discovery: Structural Holes Theory as a Framework for Literature-Based Discovery

Andrej Kastrin¹, Bojan Cestnik², Alja Nike Kastrin³, Nada Lavrač⁴

¹University of Ljubljana, Ljubljana, Slovenia – andrej.kastrin@mf.uni-lj.si

²Temida d.o.o & Jožef Stefan Institute, Ljubljana, Slovenia – bojan.cestnik@temida.si

³Jožef Stefan Institute, Ljubljana, Slovenia – nike.kastrin@gmail.com

⁴Jožef Stefan Institute & University of Ljubljana, Ljubljana, Slovenia – nada.lavrac@ijs.si

Abstract

Literature-based discovery (LBD) seeks to uncover implicit knowledge by identifying novel relationships between a starting concept A (i.e., a keyword or keyphrase) and a target concept C through intermediate bridging concepts (B). Current LBD approaches face two major limitations: limited mechanistic insight into how connections emerge across domains and a reliance on expert judgment for intermediate selection. To address this, we apply structural hole theory (SHT) to systematically prioritize intermediate B concepts. Rooted in Granovetter's theory of weak ties and formalized by Burt, SHT posits that nodes positioned between disconnected communities gain informational advantages. We demonstrate how SHT-based indicators of cross-domain brokerage identify B concepts capable of linking conceptually distant clusters, grounding LBD in a topological model of scientific discovery. Our approach supports both the closed and open discovery modes. To quantify brokerage, we extract lexical networks from large-scale scientific corpora and compute three complementary indicators: weighted betweenness centrality (bridging potential), Burt's network constraint (brokerage efficiency), and cluster span (cross-domain reach). These indicators are integrated into a composite statistical brokerage score to rank candidate intermediates. Validation using benchmark datasets (Raynaud's disease–fish oil and migraine magnesium) shows that SHT-based indicators successfully recover canonical A – B – C pathways. These findings demonstrate that SHT is statistically grounded and scalable framework for automating hypothesis generation in textual data analysis.

Keywords: natural language processing, text mining, complex networks, computational scientific discovery, literature-based discovery.

1. Introduction

A pivotal concept in (social) network analysis is the structural hole (SH) proposed by the leading sociologist Burt (2004). Intuitively, a SH is a gap between disconnected communities of nodes that limits interactions and communications between those communities. Individuals who bridge these gaps can take advantage of the resulting constraints to gain benefits, and such individuals are referred to as SH spanners (Lou and Tang, 2013).

To illustrate this idea, Figure 1 depicts a simple example in which community (group) A and community (group) B form two separate clusters of nodes with no direct links between them, creating an empty region between the communities that represents the SH. This gap limits information flow and interaction across the two communities. A broker (SH-spanning) node connects to members of both groups, thereby bridging the gap and enabling communication, coordination, and potential influence across the otherwise disconnected communities.

The mechanism by which a broker links disconnected communities extends beyond social interaction networks and can likewise be applied to knowledge structures that exhibit similar

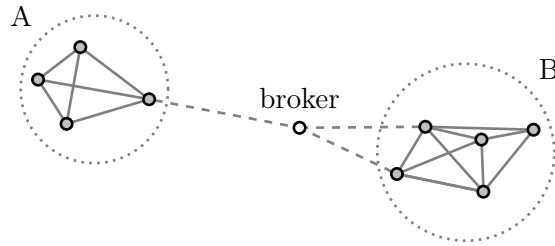


Figure 1: Conceptual depiction of two unconnected node communities (A and B) and the SH created by the absence of links between them.

fragmentation. Theoretical and empirical work increasingly conceptualizes SHs not only as social configurations but also as knowledge structures: they represent the “spaces” where communication breaks down, yet also where inter-group knowledge transmission and learning can occur (Rost, 2011). Research on innovation shows that actors who bridge SHs gain access to diverse, nonredundant knowledge and are more likely to generate innovative or disruptive outcomes (Guo et al., 2021). Extending Burt’s framework from social to intellectual networks, Chen et al. (2009) argue that SHs in co-citation and other associative scientific networks delineate disconnected domains of knowledge. Bridging such domains is theorized to be a high-risk, high-reward mechanism for transformative scientific discovery because it introduces concepts and methods from disparate fields that can reframe existing problems and catalyze new research trajectories. In this view, SHs can be reconceptualized as intellectual gaps between literatures, and bridging them becomes a central mechanism of scientific discovery and knowledge diffusion.

In this work, we examine how structural hole theory (SHT) can be used for literature-based discovery (LBD) by leveraging its core principle of bridging disconnected information or knowledge clusters to uncover novel connections or insights from scientific publications. LBD is a computational methodology that identifies implicit associations between knowledge concepts in disjoint literatures, often based on the ABC model, where if concept *A* is linked to *B* and *B* is linked to *C*, a novel indirect association between *A* and *C* can be hypothesized.

Applying SHT to LBD involves treating knowledge clusters or topics in literature as social groups in a network, with SHs representing gaps or disconnections between these clusters. A researcher or algorithm acting as a “broker” or “spanner” across these holes can facilitate the discovery of previously unconnected but potentially insightful links between separate literature domains, enabling innovation or new hypotheses. The approach presented here employs the theory’s idea that bridging SHs provides access to diverse, non-redundant information and alternative thinking, which is essential for generating innovative research questions or discovering novel knowledge in literature.

The article is organized as follows. Section 2. provides the necessary background and reviews related work. Section 3. outlines the methods employed in our study. Section 4. presents the results, in which we demonstrate our approach through two case studies. Section 5. concludes the article with a summary of the main findings and reflections on future directions.

2. Background and Related Work

2.1. Literature-Based Discovery

LBD is a subfield of artificial intelligence (AI) at the intersection of information retrieval, natural language processing (NLP) and computational scientific discovery, which is dedicated to automating the scientific discovery process. The early Swanson’s approach to LBD can be formalized using a generic ABC model (Figure 2) that considers two independent literature sets, A and C (Swanson, 1986b). In this model, a represents a source concept, c is a target concept, and b serves as a bridge or intermediate concept that connects the two. The key idea is that if a is associated with b in one body of literature and b is associated with c in another—yet a and c have not been directly linked in any publication—there may be a novel, undiscovered relationship between a and c worth exploring.

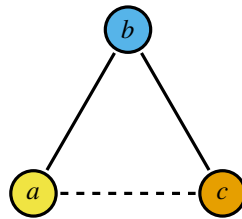


Figure 2: Swanson’s ABC model of discovery. When a is related to b , and b is related to c , it suggests the possibility of an undiscovered indirect relationship between a and c .

A seminal example of this model is Swanson’s (1986a) groundbreaking discovery linking dietary fish oil (c) to Raynaud’s disease (a). He found that Raynaud’s disease was associated with reduced blood viscosity in one set of articles, while another set linked high blood viscosity to fish oil. Although no studies at the time had made a direct connection between Raynaud’s disease and fish oil, Swanson’s hypothesis suggested a new therapeutic use for fish oil, which was later confirmed by clinical research (DiGiacomo et al., 1989).

In general, LBD encompasses two tasks: hypothesis validation and hypothesis generation, which correspond to closed and open discovery modes, respectively. In closed discovery (Figure 3a), the process starts with two known elements, a starting concept (a) and a target concept (c), and seeks to validate or elaborate the relationship between them by identifying intermediate mechanisms (b) that connect a and c . Conversely, in open discovery (Figure 3b), the process begins with a starting concept (a) and aims to uncover previously unrecognized intermediate concepts (b) and target concepts (c) that could suggest novel research hypotheses.

In a standard formulation, the LBD workflow comprises two principal components (Thilakaratne et al., 2019): a filtering stage and a ranking stage. The purpose of the filtering stage is to constrain the discovery space by retaining only those concepts that are plausibly relevant to the target domain. This reduction step is essential for eliminating overly broad, trivial, or otherwise uninformative terms that could introduce noise into the downstream analysis.

Following filtration, the ranking stage evaluates the remaining concepts and assigns priority according to their potential contribution to novel knowledge discovery. This component is intended to attenuate the influence of weak or spurious associations while amplifying the prominence of concept pairs that exhibit stronger or more meaningful relationships. Approaches for ranking concepts during LBD process can be grouped into four categories according to their statistical properties (Aizawa, 2003): (i) popularity-based measures, which quantify how frequently a term appears or the probability of its occurrence; (ii) specificity measures, which

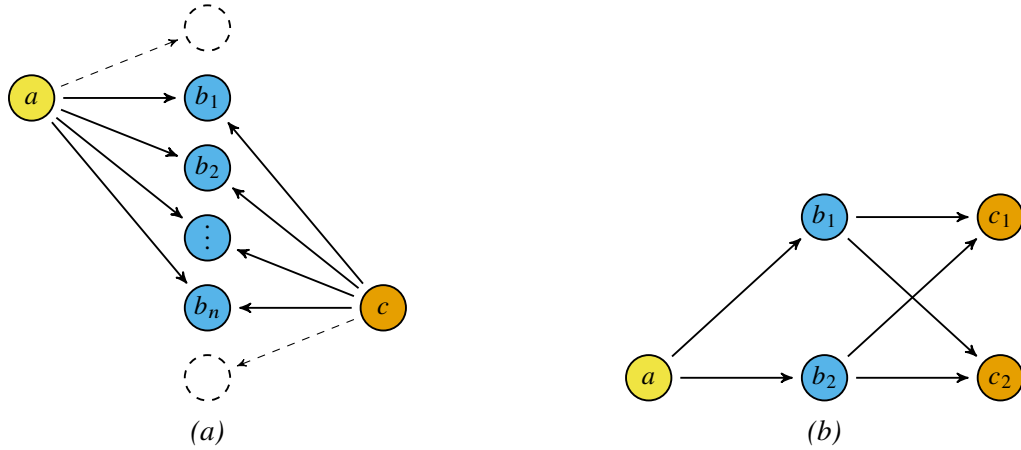


Figure 3: Closed (a) and open (b) LBD mode.

capture the informational content or entropy associated with a term (e.g., mutual information); (iii) discriminative measures, which assess how well a term supports a chosen classification or decision function (e.g., information gain); and (iv) representational measures, which describe how effectively a term characterizes the documents in which it occurs, such as tf-idf.

2.2. Structural Hole Theory

The origins of SHT can be traced to Granovetter’s (1973) work on the strength of weak ties, which demonstrated the critical role of weaker social connections in facilitating information flow across disparate networks. When two individuals share strong relationships with the same set of contacts, their social worlds largely overlap. In contrast, weak ties frequently span otherwise separate groups and therefore act as channels through which unfamiliar ideas or opportunities can spread. Burt (1992) extended this perspective by arguing that the benefits often attributed to weak ties arise not simply from tie strength but from the structural position these ties create. In his formulation, individuals who bridge gaps between disconnected groups—those occupying SHs—gain informational advantages by accessing diverse, non-redundant sources.

To formalize this idea, Burt (2004) introduced a set of metrics designed to assess whether an actor’s network position reflects meaningful brokerage. These measures are computed on the ego network, which consists of a focal individual (the ego) and all of the ego’s immediate neighbors (the alters), together with the connections among those alters. Indicators such as effective size or constraint quantify the degree to which an ego’s contacts are themselves interconnected. Actors with larger effective size and lower constraint are more likely to function as brokers spanning SHs, since their alters do not simply replicate the same information or social environment.

3. Methods

We conduct two experiments to illustrate how SHT can guide the LBD process. The first case study demonstrates a closed discovery validation on the Raynaud’s Syndrome–Dietary Fish Oil (RS–DFO) dataset, where both the starting (a) and target (c) concepts are known. The second case study applies the open discovery approach to the Migraine–Magnesium (Mig–Mg) dataset, where novel connections are identified beginning from a single concept.

In Section 3.1., we describe the data preprocessing steps, including semantic filtering using Unified Medical Language System (UMLS) semantic types to reduce the search space and the

transformation of filtered bibliographic records into co-occurrence networks. The proposed approach for ranking candidate bridging terms and systematically guiding the LBD process is presented in Section 3.2., where we introduce three SH metrics along with the rank aggregation procedure.

3.1. Data Acquisition and Network Construction

MEDLINE is the premier bibliographic collection in the life science disciplines. As of November 2025, it includes more than 32 million citations that begin in the late 19th century. Since the mid-1940s, MEDLINE records have been indexed by qualified annotators from the U.S. National Library of Medicine using the Medical Subject Headings (MeSH) thesaurus (Lipscomb, 2000). MeSH is a hierarchically organized thesaurus containing biomedical concepts at distinct levels of complexity. The vocabulary includes different categories of terms, among which only the main headings (descriptors) were used in this study. Descriptors are the core components of the MeSH thesaurus and signify the main subject of a particular bibliographic record. For example, for a citation which reports the results of analysis of the NLP approach for the studying protein–ligand interactions, MeSH descriptors might be Ligands, Natural Language Processing, Protein Binding, and Proteins among others. In each record, some MeSH descriptors are labeled as major descriptors, which designate the principal subject of the record. The 2024 MeSH vocabulary, that was exploited in this work, consists of 30,764 descriptors.

In the closed-discovery setting, the LBD process begins with both the starting concept (Raynaud’s syndrome) and the target concept (fish oil). The bibliographic records for the RS–DFO dataset were retrieved using the following PubMed queries:

1. raynaud [TI] AND 1900/01/01:1985/11/30 [PDAT]
2. fish oil [TI] AND 1900/01/01:1985/11/30 [PDAT]

Swanson formulated the RS–DFO hypothesis at the end of November 1985 (Swanson, 1986a); therefore, we restricted the upper bound of the time window using the PubMed “Publication Date” (PDAT) tag to reflect only the literature available at that time.

The second dataset corresponds to the Mig–Mg domain pair (Swanson, 1988), here used to illustrate open discovery mode. In this setting, only the starting concept (migraine) is required to initiate the LBD process, and the target concept as well as bridging terms are identified later. We constructed this dataset using the query `migraine [TI] AND 1900/01/01:1987/12/31 [PDAT]`. As in the RS–DFO dataset, the PDAT tag was used to restrict the retrieved literature to the period available up to the end of 1987.

To narrow the search space and accelerate the computation of SH measures, we retained only MeSH terms assigned to the preselected semantic types (Weeber et al., 2001). As a semantic filter for terms in set *A* we used the following semantic types (filter #1): (i) Amino Acid, Peptide, or Protein, (ii) Finding, (iii) Pathologic Function, (iv) Biologic Function, (v) Phenomenon or Process, (vi) Cell Function, (vii) Molecular Function, (viii) Organism Function, (ix) Organ or Tissue Function, and (x) Physiologic Function. Similarly, in set *C*, we included MeSH terms categorized under the semantic types (filter #2) (i) Vitamin and (ii) Element, Ion, or Isotope.

Finally, to set up a co-occurrence network, we considered each MeSH term as a node. An edge between two MeSH descriptors was defined if they appear together in the same MEDLINE citation. Note that we did not consider the direction of the relations (i.e., the relation between descriptors *u* and *v* is the same as the relation between descriptors *v* and *u*), that is, the edges are undirected.

3.2. Structural Hole Metrics

Indicators for detecting SHs can be selected in various ways to support effective analysis across different types of networks. In this work, we follow Burt's (1992) recommendation and quantify SHs using betweenness centrality, effective network size, and constraint.

Betweenness centrality of a node is defined as the portion of the number of shortest paths (between all pairs of nodes) that pass through the given node divided by the number of shortest path between any pair of nodes (regardless of passing through the given node). More precisely, for a specific node i the betweenness centrality is

$$C_B(i) = \sum_{i < j}^n \frac{g_{ijk}}{g_{ij}}, \quad i \neq j, i \neq k, j \neq k$$

where g_{ij} is the geodesic distance (shortest paths) linking i and j and g_{ijk} is the geodesic distance linking i and j that contains k . Nodes positioned on the shortest paths between many other node pairs exhibit higher betweenness centrality, indicating a greater likelihood of serving as brokers spanning SHs.

Effective size of a node measures its nonredundant connections and is computed as

$$ES(i) = \sum_j \left(1 - \sum_q p_{iq} m_{jq} \right), \quad q \neq j$$

where m_{jq} represents i 's interactions with q divided by j 's strongest relationship and p_{iq} is the proportion of i 's energy invested in relationship with q (Burt, 1992).

Burt's (1992) constraint captures a node's ability to exploit SHs by measuring the degree to which its contacts are interconnected, with lower constraint indicating greater brokerage opportunity. Constraint for node i can be computed as

$$C_i = \sum_{j \in \Gamma(i)} \left(p_{ij} + \sum_{q \in \Gamma(i)} p_{iq} p_{qj} \right)^2, \quad q \neq i, j$$

where $\Gamma(i)$ refers to the set of nodes which are direct neighbors of node i , and p_{ij} denotes the proportion of i 's investment allocated to its relationship with node j ; p_{iq} and p_{qj} are determined analogically.

Each metric produces an independent ranking of nodes. To combine these perspectives, we convert the metric values into relative ranks and then aggregate the individual ranks into a single composite brokerage rank, which reflects the overall prominence of a node in spanning SHs (Kolde et al., 2012). Specifically, we employed the robust rank aggregation (RRA) method, which uses order statistics to evaluate how each node is positioned across the individual lists compared to a baseline null model of randomly shuffled rankings. The algorithm calculates a significance score for each node, reflecting how consistently higher it is ranked than expected by chance, thereby ensuring that the final composite ranking is robust to noise and outliers.

4. Results

This section is structured around two case studies: the closed discovery experiment on the RS–DFO dataset (Section 4.1.) and the open discovery experiment on the Mig–Mg dataset (Section 4.2.).

4.1. Closed Discovery Mode

The closed discovery experiment was conducted on the RS–DFO dataset using two initial sets of documents. We considered 1,268 documents related to Raynaud’s syndrome (*A*) and 150 documents representing the fish oil (*C*). In this setting, the objective was to evaluate the hypothesis that blood viscosity acted as a bridging concept between sets *A* and *B*. Documents in set *A* expanded to 1,616 unique MeSH terms, while set *C* consisted of 482 MeSH terms.

After the semantic filtering step (set *A* was processed with semantic filter #1 and set *C* with semantic filter #2), set *A* contained 274 unique MeSH terms and set *C* contained 90 MeSH terms. Two co-occurrence networks of MeSH terms were subsequently constructed. Network *A*, derived from terms in set *A*, consisted of 235 nodes and 715 edges. Network *C* was smaller, consisting of 78 nodes and 146 edges. Note that the original number of MeSH terms in each set and the final number of nodes in the corresponding network differ; this is because all isolated nodes (i.e., terms that did not co-occur with any other term in the network) were removed.

Table 1 summarizes the aggregated ranks of the 27 MeSH terms that are common to both literature sets, *A* and *C*. The terms are ordered by their ranks in the set *A*.

Table 1: Ranks of MeSH terms for literature sets A and C, along with the aggregated value across both rank lists (AC) for the RS–DFO domain (closed discovery mode).

MeSH term	R_A	R_C	R_{AC}
Blood Pressure	1.0	5.0	2.5
Blood Viscosity	2.0	10.0	6.0
Hemoglobins	3.0	9.0	5.0
Arteriosclerosis	4.0	3.0	1.0
Platelet Aggregation	5.0	1.0	2.5
Thrombosis	6.0	15.5	11.0
Blood Coagulation	7.0	2.0	4.0
Hemostasis	8.0	12.0	8.0
Erythrocyte Deformability	9.0	15.5	10.0
Stimulation, Chemical	10.0	20.0	16.0
Lipid Metabolism	11.0	4.0	7.0
Oxidoreductases	12.0	22.0	18.0
Cerebrovascular Disorders	13.0	6.0	9.0
Histocompatibility Antigens Class II	14.0	17.0	12.0
Aging	15.0	23.0	19.5
Prolactin	16.0	25.0	24.0
Lysine	17.0	19.0	14.0
Cytochrome P-450 Enzyme System	18.0	26.0	25.5
Cell Adhesion	19.0	7.0	13.0
Lipoproteins	20.0	8.0	15.0
Death, Sudden	21.0	24.0	23.0
Cell Survival	22.0	11.0	17.0
Lung Compliance	23.0	21.0	19.5
Intestinal Absorption	24.0	14.0	22.0
Alkaline Phosphatase	25.0	27.0	27.0
Amino Acids	26.0	18.0	25.5
Adaptation, Physiological	27.0	13.0	21.0

Our ranking procedure clearly identifies both platelet aggregation and blood viscosity as highly ranked bridging concepts. Platelet aggregation achieved an aggregated rank of 2.5, placing it among the top-ranked terms, while blood viscosity obtained an aggregated rank of 6.0. These findings align closely with Swanson’s original discovery pathway (Swanson, 1986a), where he identified these specific physiological mechanisms as key intermediates connecting fish oil consumption to the amelioration of Raynaud’s syndrome symptoms.

4.2. Open Discovery Mode

The open discovery experiment was performed on the Mig–Mg dataset starting with a single concept *A* (migraine). The aim here was to find meaningful connections between a starting concept and previously unknown related target concepts (*C*) through bridging intermediate concepts (*B*).

We started the analysis with 1,156 papers on migraine headache (*A*) containing 1,812 unique MeSH terms. Following the procedure from the first experiment (Section 4.1.), we removed terms not relevant to the discovery task. Applying the semantic filter #1 yielded 283 potential bridging terms and reduced the document set to 626 records. The resulting MeSH co-occurrence network comprised 100 nodes and 183 edges. SH measures were computed for all nodes, and the ranks for betweenness centrality, constraint, and effective size were aggregated into a combined brokerage rank. The top five MeSH terms were (i) Cerebrovascular Circulation, (ii) Vasoconstriction, (iii) Blood Pressure, (iv) Cerebrovascular Disorders, and (v) Muscle Contraction. The decision to present only the top five terms was arbitrary and intended for demonstrative purposes.

In the second step (*B* to *C*), we retrieved bibliographic records for each of the five selected MeSH terms, applied the semantic filter #2, and generated the corresponding networks. SH metrics were calculated separately for each network and then aggregated. Table 2 summarizes the ranks for the 14 MeSH terms common to all five networks, together with the row-wise aggregated rank shown in the last column.

Table 2: Ranks of MeSH terms for literature sets C_i for the Mig–Mg domain (open discovery mode).

MeSH term	R_{C_1}	R_{C_2}	R_{C_3}	R_{C_4}	R_{C_5}	R_C
Potassium	1.0	2.0	5.0	7.0	2.0	2.0
Oxygen	2.0	3.0	1.5	1.0	3.0	1.0
Rubidium	3.0	14.0	11.0	13.0	7.0	11.0
Nicotinic Acids	4.5	11.0	7.0	2.0	13.0	8.0
Hydrogen	4.5	8.0	14.0	12.0	9.0	12.0
Xenon Radioisotopes	6.0	6.0	12.0	5.0	14.0	7.0
Sodium	7.0	5.0	1.5	8.0	4.0	4.0
Magnesium	8.0	10.0	10.0	9.0	5.0	6.0
Tritium	9.0	7.0	4.0	6.0	6.0	5.0
Barium	10.0	13.0	8.0	3.0	10.0	10.0
Calcium	11.0	1.0	3.0	4.0	1.0	3.0
Nitrogen	12.0	4.0	9.0	11.0	12.0	9.0
Ions	13.0	9.0	13.0	14.0	8.0	14.0
Manganese	14.0	12.0	6.0	10.0	11.0	13.0

The suggested approach successfully identified magnesium as a target concept, achieving an aggregated rank of 6 among the 14 common terms across all five bridging networks. Mag-

nesium ranked consistently within the top ten across all five intermediate concept networks, demonstrating robust structural positioning across multiple pathways.

5. Conclusion

LBD has long struggled with the substantial manual effort required to identify meaningful bridging concepts across disconnected domains. Although computational approaches have made progress toward automating parts of the discovery workflow, most existing methods lack a coherent theoretical foundation for explaining why certain intermediate concepts successfully link disparate literatures. This study addresses that gap by showing how SHT offers both explanatory value and practical guidance for the LBD process.

Our results from two benchmark datasets demonstrate that SH metrics consistently highlight bridging terms aligned with known discovery pathways. In the RS–DFO case, blood viscosity and platelet aggregation emerged among the top-ranked intermediates, mirroring Swanson’s original findings. Likewise, in the Mig–Mg analysis, magnesium was recovered as one of the highest-ranked target concepts when starting from physiologically relevant bridging terms. These outcomes indicate that principles governing information flow in social networks translate effectively to knowledge structures represented in scientific literature.

The theoretical grounding offered by SHT constitutes an important advancement. Instead of relying solely on statistical co-occurrence patterns, our approach draws on an established social-science framework to explain how cross-domain connections facilitate discovery. Concepts positioned at SHs derive their value not merely from frequency but from their capacity to broker between otherwise isolated knowledge communities. This perspective clarifies the mechanisms underlying successful discovery and provides more transparent criteria for prioritizing candidate hypotheses.

Several limitations merit consideration. First, our reliance on MeSH descriptors restricts the analysis to biomedical literature indexed in MEDLINE. Extending this framework to full-text corpora or to other scientific disciplines would require adapting the concept extraction process. Second, the semantic filters applied in this study were tailored to the two benchmark datasets; developing more generalizable filters would improve applicability across broader discovery scenarios. Third, although our rank aggregation procedure combines multiple SH indicators, the optimal weighting of these metrics remains an open question and may depend on domain-specific characteristics.

Future work should include rigorous statistical evaluation of the proposed metrics in comparison with established LBD ranking methods. Such analyses would help clarify when and where SH-based indicators offer genuine advantages. Additionally, exploring whether SH metrics can be combined with traditional statistical measures to yield improved performance may uncover complementary strengths worth leveraging.

Integrating SHT into the LBD workflow opens promising avenues for computational scientific discovery. By grounding hypothesis generation in a well-established theory from network science, this approach moves beyond purely data-driven techniques toward theory-informed discovery. As scientific literature continues to grow at an accelerating pace, such principled frameworks will become increasingly important for navigating the expanding landscape of human knowledge and uncovering the hidden connections that drive innovation.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency through the Knowledge Technologies (Grant No. P2-0103), and Methodology for Data Analysis in Medical Sciences (Grant No. P3-0154) core research projects.

References

- Aizawa A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1):45–65.
- Burt R. (1992). *Structural holes: The social structure of competition*. Harvard University Press, Cambridge, MA.
- Burt R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399.
- Chen C., Chen Y., Horowitz M., Hou H., Liu Z., and Pellegrino D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3):191–209.
- DiGiacomo R. A., Kremer J. M., and Shah D. M. (1989). Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *The American Journal of Medicine*, 86(2):158–164.
- Granovetter M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.
- Guo M., Yang N., Wang J., Zhang Y., and Wang Y. (2021). How do structural holes promote network expansion? *Technological Forecasting and Social Change*, 173:121129.
- Kolde R., Laur S., Adler P., and Vilo J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580.
- Lipscomb C. E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–266.
- Lou T. and Tang J. (2013). Mining structural hole spanners through information diffusion in social networks. In *Proceedings of the 22nd international conference on World Wide Web, WWW '13*, pp. 825–836, New York, NY, USA. Association for Computing Machinery.
- Rost K. (2011). The strength of strong ties in the creation of innovation. *Research Policy*, 40(4):588–604.
- Swanson D. R. (1986a). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18.
- Swanson D. R. (1986b). Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118.
- Swanson D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557.
- Thilakaratne M., Falkner K., and Atapattu T. (2019). A systematic review on literature-based discovery workflow. *PeerJ Computer Science*, p. e235.
- Weeber M., Klein H., de Jong-van den Berg L. T., and Vos R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557.

S06

**Specialized Session
Text Analysis in Economics**

09-lug
15:00 - 16:30
Aula 11

From Fed Speeches to Signals: Embedding Central Bank Communication in Macroeconomic Forecasting

Angelo Famà¹

¹Department of Economics, University of Insubria

Abstract

This paper examines whether semantic novelty in Federal Reserve speeches contains incremental information for US macroeconomic density forecasting. Speeches are processed with TextRank extractive summarisation, encoded using FinBERT and Word2Vec, aggregated at the monthly frequency, and summarised by the cosine distance between consecutive monthly embeddings. The resulting communication-novelty indicator is included as an endogenous variable in Bayesian VARs estimated with hierarchical shrinkage and a COVID-19 volatility correction. The empirical results show that point-forecast gains are concentrated in expectation-sensitive variables such as industrial production, equity prices, hours worked, and the federal funds rate. Density-forecast gains are stronger, especially at medium and long horizons. A regime decomposition using the continuous ranked probability score shows that the indicator is not a universal short-run predictor: its contribution is largest during the COVID emergency and remains positive during the subsequent inflation-tightening cycle, while pre-COVID improvements appear at longer horizons. Benchmark comparisons with PCA-based embedding factors and the news-based monetary policy uncertainty index show that neither absorbs the predictive content of the cosine-distance measure. The results suggest that Federal Reserve speech novelty is best interpreted as a parsimonious indicator of communication-regime instability.

Keywords: Bayesian VAR; FinBERT; Word2Vec; central bank communication; density forecasting; state dependence.

1. Introduction

Central banks communicate extensively through speeches, statements, and press conferences. These communications often carry forward-looking information about inflation, growth, and the likely policy path not yet visible in standard macroeconomic aggregates (Hansen et al., 2018; Kalamara et al., 2022). A growing literature documents that textual information improves macroeconomic forecasting (Kalamara et al., 2022), and financial-domain language models such as FinBERT (Araci, 2019) provide contextual representations of economic text. Most applications use textual signals as external sentiment or uncertainty regressors; less is known about whether semantic representations improve forecasts when embedded directly into a multivariate time-series system. The closest contribution is De et al. (2025), who embed ECB introductory statements into a BVAR for euro-area inflation forecasting; we extend this line to Federal Reserve speeches and the United States.

We process more than two decades of Fed speeches through TextRank extractive summarisation, encode the extracted text with FinBERT and Word2Vec, aggregate monthly via the element-wise median, and compute the cosine distance between consecutive monthly embeddings. The resulting scalar is real-time with respect to the textual information set and has a direct geometric interpretation as the angular displacement between successive communication regimes. We include it as an additional endogenous variable in Small and Medium Bayesian VARs estimated under the hierarchical framework of Giannone et al. (2015) with the COVID-19 heteroskedasticity correction of Lenza and Primiceri (2022), sampling the seven hyperparameters jointly via adaptive Random-Walk Metropolis-Hastings.

Three findings emerge. First, point-forecast gains are heterogeneous and concentrated in expectation-sensitive variables: industrial production, equity prices, hours, and the federal funds rate. Second, density-forecast gains are stronger and broader at medium and long horizons, consistent with the unsigned nature of the cosine distance, which captures the magnitude but not the hawkish–dovish direction of a communication shift. Third, the evidence is state-dependent: pre-COVID gains exist at $h = 12$ under CRPS, the COVID emergency window is the period of largest per-origin density improvement, and the post-2022 tightening cycle delivers intermediate gains. Benchmark comparisons with PCA factors and with the news-based monetary policy uncertainty index of Husted et al. (2020) show that neither absorbs the predictive content of the cosine-distance measure.

2. Methodology

We process each Federal Reserve speech in four steps. First, the speech is condensed via TextRank (Mihalcea and Tarau, 2004), which selects the most salient sentences subject to FinBERT’s 512-token input constraint by building a sentence-similarity graph from TF-IDF vectors and applying PageRank (damping $\alpha = 0.85$); selected sentences are re-ordered by original position. The procedure is fully deterministic. Second, the extracted text is encoded using two alternative architectures: FinBERT (Araci, 2019), a transformer pre-trained on financial text yielding 768-dimensional contextual embeddings, and Word2Vec (Mikolov et al., 2013), a static model trained on Google News yielding 300-dimensional vectors averaged over content words.

Third, individual speech embeddings within each calendar month are aggregated via the element-wise median, $\tilde{\mathbf{e}}_t = \text{median}(\{\mathbf{e}_{i,t}\}_{i=1}^{N_t})$, which is robust to outlier speeches (the arithmetic mean yields a Pearson correlation of 0.972 on the FinBERT series). Fourth, the monthly communication-novelty indicator is the cosine distance between consecutive monthly aggregate embeddings,

$$d_t = 1 - \frac{\tilde{\mathbf{e}}_t \cdot \tilde{\mathbf{e}}_{t-1}}{\|\tilde{\mathbf{e}}_t\| \|\tilde{\mathbf{e}}_{t-1}\|}, \quad d_t \in [0, 2]. \quad (1)$$

The cosine distance is scale-invariant, real-time by construction, and stationary by both ADF and Phillips–Perron tests (rejection at the 1% level for both series). It measures the magnitude of the month-to-month semantic shift in central bank communication rather than a level of policy uncertainty in the sense of Husted et al. (2020).

2.1. Text-augmented Bayesian VAR and forecast evaluation

Let y_t denote the $n \times 1$ vector of macroeconomic variables observed at the monthly frequency, specified in (log) levels. The augmented system stacks y_t and d_t ,

$$\tilde{y}_t = (y_t', d_t)' \in \mathbb{R}^{n+1}, \quad \tilde{y}_t = c + B_1 \tilde{y}_{t-1} + \dots + B_p \tilde{y}_{t-p} + s_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma). \quad (2)$$

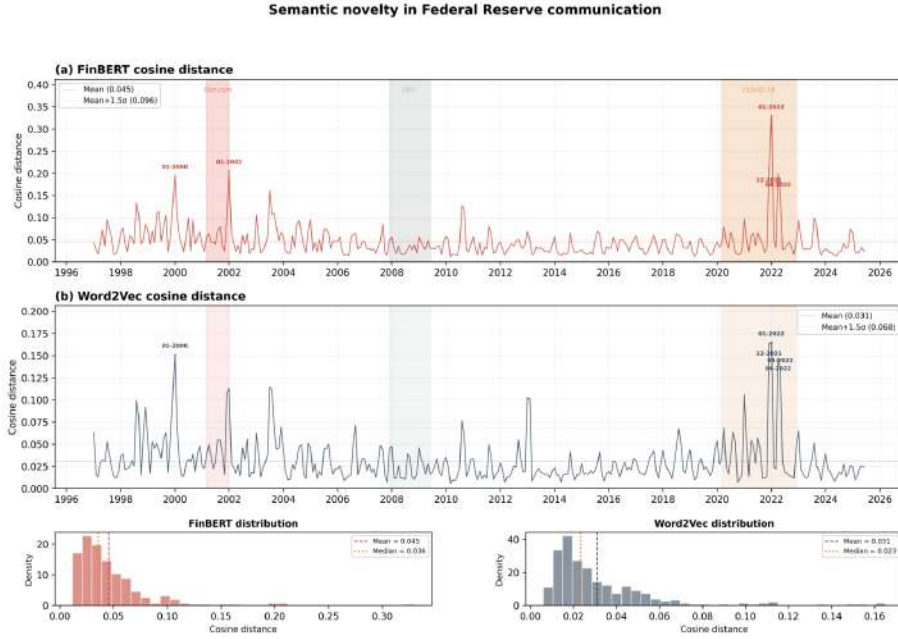


Figure 1: Cosine distance series and speeches mapping: FinBERT and Word2Vec. Panel (a) displays the monthly cosine distance computed from FinBERT embeddings; panel (b) shows the analogous series from Word2Vec embeddings. Dashed horizontal lines indicate the sample mean. Shaded bands denote the dot-com recession, the Global Financial Crisis, and the COVID-19 pandemic through the subsequent tightening cycle. Panels (c) and (d) report the empirical density of each series together with its mean and median.

We adopt the Minnesota prior (Giannone et al., 2015) in the conjugate Normal-Inverse-Wishart family. The macroeconomic block receives $\mathbb{E}[(B_1)_{ii}] = 1$ on the first own lag, while the cosine-distance variable receives $\mathbb{E}[(B_1)_{n+1,n+1}] = 0$ to avoid inducing artificial persistence in a stationary signal; the sum-of-coefficients and single-unit-root dummy observations are extended consistently to exempt the textual factor from the unit-root constraint. The scaling factor s_t accommodates COVID-19 heteroskedasticity following Lenza and Primiceri (2022): $s_t = 1$ for $t < T_c$, $s_t \in \{S_0, S_1, S_2\}$ for $t \in \{T_c, T_c + 1, T_c + 2\}$, and $s_t = 1 + (S_2 - 1)\rho^{t-T_c-2}$ for $t \geq T_c + 3$, where T_c is the COVID onset month (March 2020) and the four parameters (S_0, S_1, S_2, ρ) are jointly estimated. The hyperparameter vector $\gamma = (\lambda, \mu, \theta, S_0, S_1, S_2, \rho)'$ is sampled from the joint posterior $p(\gamma | Y) \propto p(Y | \gamma) p(\gamma)$ under proper hyperpriors (Gamma for scale parameters, Beta for ρ) via adaptive Random-Walk Metropolis-Hastings initialised at the posterior mode; the marginal likelihood admits a closed-form. Conditional on each draw, the VAR coefficients retain the conjugate Normal-Inverse-Wishart form and predictive paths are obtained by iterating the VAR forward, with draw-specific COVID scaling.

Forecasts are generated under an expanding-window scheme at horizons $h \in \{1, 6, 12\}$ months. Point-forecast accuracy is summarised by the MSFE ratio of the augmented over the baseline model, with significance assessed via the one-sided Diebold and Mariano (1995) test using Newey-West HAC standard errors with bandwidth $h - 1$. Density-forecast accuracy is summarised by the average log predictive score difference,

$$\text{ALPSD}_{i,h} = \frac{1}{P} \sum_{t=1}^P [\log \hat{p}^{\text{aug}}(y_{i,t+h} | Y_{1:t}) - \log \hat{p}^{\text{base}}(y_{i,t+h} | Y_{1:t})], \quad (3)$$

where the predictive log density is evaluated non-parametrically by Gaussian kernel density estimation on the posterior predictive draws with Silverman’s bandwidth and significance is

assessed via the one-sided HAC log-score test with the same HAC correction. To complement the log score, we also report the continuous ranked probability score, $\text{CRPS}(F, y) = \int (F(z) - \mathbb{1}\{y \leq z\})^2 dz$, with F the empirical predictive CDF. CRPS is a proper scoring rule that does not require density smoothing. We report cumulative CRPS improvements decomposed across three sub-periods (pre-COVID, COVID emergency March 2020–December 2021, tightening/post-COVID from January 2022) as the central robustness check on the density-forecast evidence.

2.2. Data

The dataset is monthly, spans January 1997 to June 2025, and contains $T = 339$ observations after the complete-case filter. Macroeconomic series are sourced from FRED and Yahoo Finance; Federal Reserve speech transcripts are retrieved from the Bank for International Settlements central bank speeches database. The semantic indicators are real-time with respect to the textual information set, but the macroeconomic series are latest-vintage, so the forecasting exercise is pseudo real-time with respect to the macro block. The *Small* BVAR contains the four core monetary-policy variables: log CPI, log industrial production, the unemployment rate, and the federal funds rate. The *Medium* BVAR augments this core with the log S&P 500, average weekly hours, log hourly wages, and a macro uncertainty index. The semantic indicator (FinBERT or Word2Vec cosine distance) is added one at a time.

3. Empirical Results

We focus on lag order $p = 6$, which delivers a clean balance between macroeconomic persistence and the high-frequency nature of the communication signal.¹ The effective out-of-sample target windows are 2012:M10–2025:M6 for $h = 1$ ($N = 153$), 2013:M3–2025:M6 for $h = 6$ ($N = 148$), and 2013:M9–2025:M6 for $h = 12$ ($N = 142$).

3.1. Small BVAR

The Small BVAR removes the financial, labour-market, wage, and uncertainty variables and tests whether the semantic-novelty signal retains predictive content with a deliberately parsimonious information set. Tables 1 and 2 report the results.

Table 1: *Point forecast evaluation: Small BVAR, Lag = 6, MSFE ratios. See Table 3 for definitions.*

Variable	$h = 1$		$h = 6$		$h = 12$	
	FinBERT	W2V	FinBERT	W2V	FinBERT	W2V
Fed Funds	0.945	0.975	0.860**	0.910*	0.869**	0.897**
Unemployment	1.001	1.001	0.962	0.964	0.947	0.948
Log CPI	0.912***	0.980	0.992	1.073	1.177	1.222
Log IP	0.951*	0.956*	0.915	0.951	1.148	1.189

¹ Lag-12 results, Granger-causality diagnostics, pre/post-COVID decompositions, terminal heatmaps, and speech-level identification of major cosine spikes are reported in the full version.

Table 2: *Density forecast evaluation: Small BVAR, Lag = 6, ALPSD. See Table 4 for definitions.*

Variable	$h = 1$		$h = 6$		$h = 12$	
	FinBERT	W2V	FinBERT	W2V	FinBERT	W2V
Fed Funds	+0.128*	+0.109*	+0.929	+0.694	+0.871	+0.613
Unemployment	+0.223	-0.343	-0.012	-0.092	-0.425	-0.381
Log CPI	+0.130**	+0.098**	+0.120	+0.103	+0.500	+0.725
Log IP	-0.118	+0.034	+0.176	+0.219	+0.473	+0.448

The Small BVAR confirms that the cosine-distance factor retains predictive content under a parsimonious information set, but the gains are more concentrated: FinBERT improves MSFE in 9 of 12 cells, with the most robust point-forecast gains on the federal funds rate. Density-forecast evidence remains positive in 9 of 12 cells for both embeddings, with the largest gains again on the federal funds rate and on log CPI. CPI and industrial production deteriorate at the long horizon in point-forecast terms but improve in density-forecast terms, consistent with the unsigned-novelty channel operating through predictive uncertainty rather than conditional means. Its core contribution to policy-rate dynamics and density calibration is therefore preserved even in this smaller system.

3.2. Medium BVAR

Tables 3 and 4 report the MSFE ratios and ALPSD differentials for the Medium BVAR. The MSFE evidence shows broad directional reductions: 19 of 24 variable-horizon cells are below unity for both embeddings. The largest gains arise at medium and long horizons in industrial production, the S&P 500, hours, the federal funds rate, and at $h = 12$ also in CPI and wages. Diebold-Mariano rejections are concentrated in fewer cells, reflecting the limited power of cell-level HAC tests in a finite, COVID-affected window. The density evidence is stronger than the point-forecast evidence: positive ALPSD in 17–18 of 24 cells, with the largest gains for wages, the federal funds rate, and CPI at $h = 12$.

Table 3: *Point forecast evaluation: Medium BVAR, Lag = 6, MSFE ratios. Augmented over baseline; values below unity (bold) indicate lower squared forecast errors. Stars: one-sided Diebold–Mariano with Newey–West HAC (bandwidth $h - 1$). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

Variable	$h = 1$		$h = 6$		$h = 12$	
	FinBERT	W2V	FinBERT	W2V	FinBERT	W2V
Log CPI	1.073	1.109	0.926*	0.967	0.768*	0.780*
Log IP	0.892**	0.890*	0.634	0.633	0.598	0.599
Unemployment	1.007	1.008	0.994	0.992	0.906	0.899
Fed Funds	0.915*	0.946	0.866**	0.932*	0.904*	0.953
Log S&P 500	0.904*	0.899*	0.651	0.636	0.686	0.701
Hours	0.855**	0.847**	0.880*	0.877*	0.868*	0.862*
Uncertainty	0.933*	0.944	0.939	0.940	1.055	1.052
Log Wages	1.099	1.095	1.131	1.134	0.670	0.678

Table 4: *Density forecast evaluation: Medium BVAR, Lag = 6, ALPSD. Augmented minus baseline; positive values (bold) indicate higher average predictive density. Stars: one-sided HAC log-score test with Newey–West HAC (bandwidth $h - 1$). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

Variable	$h = 1$		$h = 6$		$h = 12$	
	FinBERT	W2V	FinBERT	W2V	FinBERT	W2V
Log CPI	+0.173*	+0.021	+0.130	−0.042	+0.682*	+0.584
Log IP	−0.225	+0.041	+0.223	+0.303	+0.241	+0.239
Unemployment	−0.676	−0.589	−0.542	+0.037	−0.382	+0.034
Fed Funds	+0.106**	+0.073*	+0.579*	+0.335*	+0.366	+0.293*
Log S&P 500	−0.222	−0.154	+0.005	+0.018	+0.004	−0.016
Hours	+0.027	+0.069	+0.282	+0.540	+0.402	+0.757
Uncertainty	−0.156	−0.021	+0.069	+0.008	+0.191	+0.299
Log Wages	−0.074	−0.035	+1.335*	+1.203*	+2.998*	+2.971*

3.3. State dependence: regime-decomposed CRPS

Because the log score is sensitive to the kernel bandwidth, we complement it with a regime-decomposed CRPS analysis, which avoids density smoothing and is robust to heavy tails. We split the out-of-sample period into three sub-periods: pre-COVID (target dates before March 2020), COVID emergency (March 2020–December 2021), and tightening/post-COVID (January 2022 onward). Table 5 reports cumulative CRPS improvements of the augmented Medium BVAR over the baseline, summed across the eight macroeconomic variables.

Table 5: *Cumulative CRPS contributions by regime: Medium BVAR, Lag = 6. Augmented minus baseline; positive values indicate that the augmented model has lower CRPS than the baseline. Stars: one-sided HAC tests on the mean aggregate CRPS differential within each regime. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

h	Model	Pre-COVID	COVID emergency	Tightening/Post-COVID	Total
1	Cosine FinBERT	−0.05	−0.20	+0.03	−0.22
1	Cosine W2V	+0.03	+0.08	−0.05	+0.06
6	Cosine FinBERT	+1.95***	+2.29*	+1.27	+5.52
6	Cosine W2V	+1.69**	+3.97**	+0.03	+5.70
12	Cosine FinBERT	+2.28*	+7.42*	+4.38**	+14.07
12	Cosine W2V	+2.46*	+9.30**	+3.48**	+15.24

Three patterns are visible. First, at $h = 1$ the indicator delivers no consistent improvement; CRPS contributions are small in absolute value and sometimes negative during the COVID emergency, consistent with an unsigned novelty measure operating through medium-horizon predictive uncertainty rather than one-step conditional means. Second, at $h = 6$ and $h = 12$ the indicator delivers positive cumulative improvements in *every* regime, including pre-COVID; at $h = 12$ the pre-COVID contribution of FinBERT is +2.28 and of Word2Vec is +2.46, both statistically significant. The evidence therefore does not support a pure-COVID interpretation: a real, if smaller, signal exists before the pandemic. Third, the COVID emergency window remains the period of largest per-origin gains. Normalising by the number of months in each regime ($N \approx 85$ pre-COVID, $N = 22$ COVID emergency, $N \approx 42$ tightening), at $h = 12$ the per-origin FinBERT CRPS improvement is approximately +0.027 pre-COVID, +0.34 in the COVID emergency, and +0.10 in the tightening cycle, exceeding the pre-COVID intensity by roughly an order of magnitude and supporting a graduated state-dependent reading rather than a binary COVID-versus-rest split. The pattern is robust to embedding architec-

ture: FinBERT and Word2Vec deliver qualitatively similar regime contributions, indicating a broad communication-novelty signal rather than a purely deep-contextual one.

4. Benchmark comparisons: PCA factors and HRS-MPU

A natural benchmark for our scalar specification is the principal-component strategy of De et al. (2025), who add latent embedding factors to a BVAR. We re-estimate the Medium BVAR with one and four PCA factors from FinBERT (F_1^{FB} , $F_{1:4}^{\text{FB}}$) and from Word2Vec (F_1^{W2V} , $F_{1:4}^{\text{W2V}}$), constructed in real time to avoid look-ahead bias in the loadings, and re-evaluate the cosine specifications on the same restricted common window.

Figure 2 reports the cumulative log predictive score paths by variable and horizon. PCA factors are useful and often competitive, especially at long horizons and for industrial production and the S&P 500, and four-factor specifications occasionally dominate the cosine distance in point-forecast terms by absorbing additional conditional-mean variation. The cosine specifications, however, deliver more stable cumulative density gains across variables and horizons at lag $p = 6$, with fewer reversals during the COVID emergency and the tightening cycle. Each PCA factor is a linear combination of hundreds of embedding dimensions, is opaque to direct economic interpretation, and depends on recursively estimated loadings that are hard to monitor in real time. The cosine distance, by contrast, is a single scalar with a direct geometric meaning: it measures the angular displacement between consecutive monthly communication regimes, can be linked to identifiable speech episodes (such as the withdrawal of the “transitory” inflation narrative in December 2021), and is updated each month. The choice between the two strategies is therefore between flexibility and interpretability.

4.1. HRS-MPU benchmark

To check that the cosine distance is not merely capturing broad monetary policy uncertainty, we benchmark it against the news-based index of Husted et al. (2020) (HRS-MPU). The two indicators are conceptually distinct: HRS-MPU measures the *level* of policy uncertainty in newspaper coverage, while the cosine distance measures the *change* in the semantic content of Fed communication itself. Their cross-correlation is 0.21, confirming the distinction; the comparison uses the common HRS sample, which ends earlier than our speech sample.

Figure 3 reports the cumulative log predictive score paths. HRS-MPU is competitive, especially for CPI, industrial production, hours, and the S&P 500, where it tracks or sometimes exceeds the cosine specifications, but it is generally weaker on the federal funds rate and on aggregate density performance. The trajectories also differ qualitatively: HRS-MPU behaves like a persistent stock that accumulates evidence gradually, while the cosine distance behaves like a high-frequency flow that accumulates it in sharp bursts at major communication shifts. The two indicators therefore extract complementary signals, and a joint specification including both would be a natural extension.

Taken together, PCA factors and HRS-MPU contain useful information, but neither absorbs the predictive content of the cosine distance, which provides a complementary, parsimonious signal whose forecasting value is most visible in density forecasting at medium-to-long horizons.

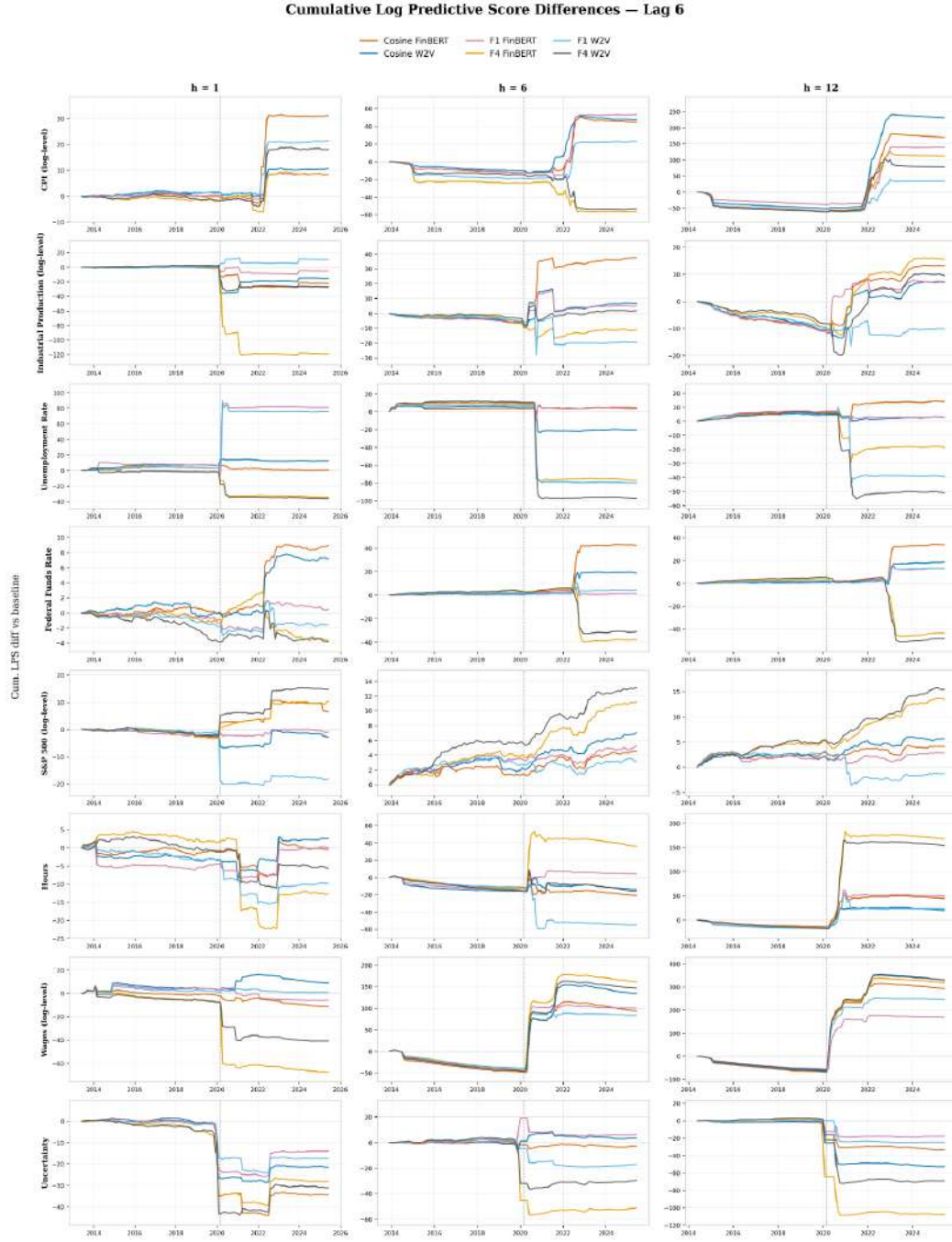


Figure 2: Cumulative log predictive score differences: PCA benchmark, Medium BVAR, Lag = 6. Cumulative log predictive score differences of the cosine and PCA textual specifications relative to the macroeconomic baseline on the common PCA evaluation window. Rows correspond to macroeconomic variables and columns to forecast horizons. Positive values indicate cumulative predictive evidence in favour of the augmented specification. The dashed vertical line marks March 2020.

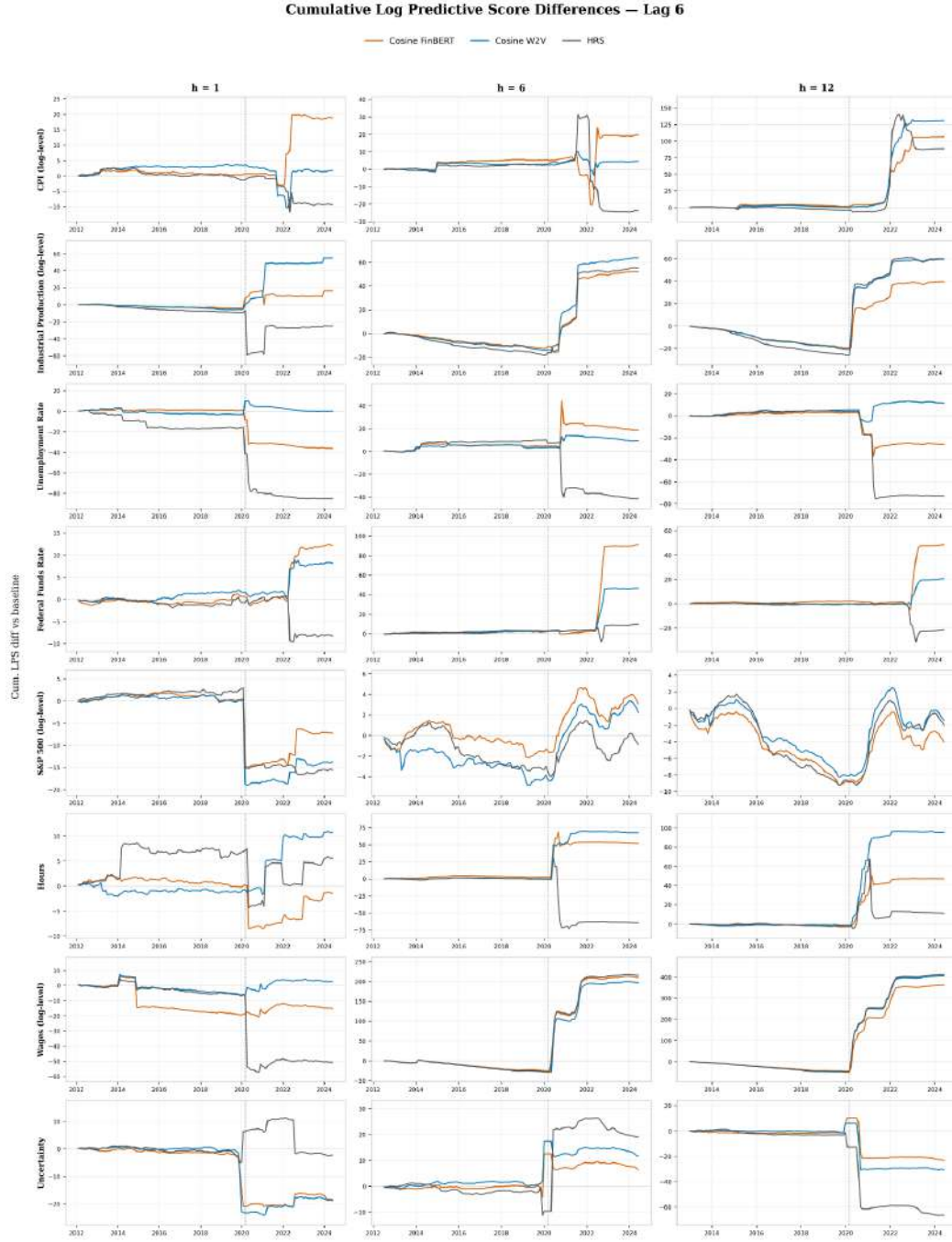


Figure 3: **Cumulative log predictive score differences: HRS-MPU benchmark, Medium BVAR, Lag = 6.** Cumulative log predictive score differences of the cosine FinBERT, cosine Word2Vec, and HRS-MPU specifications relative to the macroeconomic baseline on the common HRS evaluation window. Rows correspond to macroeconomic variables and columns to forecast horizons. Positive values indicate cumulative predictive evidence in favour of the augmented specification. The dashed vertical line marks March 2020.

5. Conclusion

This paper embeds Federal Reserve speech semantics into a hierarchical Bayesian VAR through a single communication-novelty indicator: the cosine distance between consecutive monthly speech embeddings, a real-time, parsimonious measure of the angular displacement between successive communication regimes.

The evidence supports a state-dependent forecasting interpretation. In the Medium BVAR, gains are strongest for density forecasts at medium and long horizons, especially for variables linked to expectations and policy transmission. The regime-decomposed CRPS results show that the signal is not purely a COVID artefact: pre-COVID improvements are smaller but positive at longer horizons, while the COVID emergency and the post-2022 tightening cycle generate larger gains. The Small BVAR confirms that the signal survives in a parsimonious information set, mainly through the federal funds rate and density calibration.

The cosine distance is unsigned, so it measures the magnitude rather than the direction of a communication shift; a signed hawkish–dovish companion measure is a natural extension. Even with this limitation, the results suggest that Fed speech novelty is a useful complementary signal for probabilistic macroeconomic forecasting when the policy narrative changes sharply.

References

- Araci D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv:1908.10063*.
- De A. et al. (2025). *Word2Prices: Central bank communication and inflation forecasting*. Working paper.
- Diebold F. X. and Mariano R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3): 253–263.
- Giannone D., Lenza M. and Primiceri G. E. (2015). Prior selection for vector autoregressions. *The Review of Economics and Statistics*, 97(2): 436–451.
- Hansen S., McMahon M. and Prat A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics*, 133(2): 801–870.
- Husted L., Rogers J. and Sun B. (2020). Monetary policy uncertainty. *Journal of Monetary Economics*, 115: 20–36.
- Kalamara E., Turrell A., Redl C., Kapetanios G. and Kapadia S. (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, 37(5): 896–919.
- Lenza M. and Primiceri G. E. (2022). How to estimate a vector autoregression after March 2020. *Journal of Applied Econometrics*, 37(4): 688–699.
- Mihalcea R. and Tarau P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP 2004*, pp. 404–411.
- Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations Workshop*.

C03

Public, political and media discourse

09-lug
15:00 - 16:30
Aula 12

Détection des isotopies du discours présidentiel d'Emmanuel Macron : de la cooccurrence à la self-attention

Damon Mayaffre¹, Laurent Vanni²

¹CNRS – Université Côte d’Azur – Damon.Mayaffre@univ-cotedazur.fr

²CNRS – Université Côte d’Azur – Laurent.Vanni@univ-cotedazur.fr

Abstract

A central challenge in Textual Data Analysis (TDA) is bridging the gap between statistical regularities observed in corpora and the semantic interpretation of discourse. This study addresses this issue by proposing a method to detect and analyze isotopies—the recurring semantic features that structure discourse. It does so by confronting two representations of lexical distribution: one derived from classical co-occurrence analysis, and the other from the self-attention mechanisms of Transformer-based language models.

The approach is applied to a synchronic analysis of the presidential discourse of Emmanuel Macron since his accession to power (148 speeches; 501,817 tokens). After part-of-speech filtering to retain only nouns—considered the primary carriers of semantic content—the study compares the co-occurrence profiles of words with their attention-based profiles generated by a Transformer model.

This confrontation reveals four major isotopic domains that structure the corpus: the regalian domain, the economic, the social, and the societal. Beyond these empirical findings, the comparison raises a fundamental methodological question: do the self-attention mechanisms of Transformers offer a genuine extension to the classical distributional hypothesis in TDA, or do they rather provide a computational reformulation of long-established principles in discourse analysis, particularly the concept of isotopy itself?

Keywords: co-occurrence, self-attention, transformer, isotopy, correspondence analysis, textometry, logometry, political discourse, Emmanuel Macron.

Résumé

La détection d’isotopies dans les discours constitue une préoccupation méthodologique centrale de l’ADT, dans la mesure où elle vise à transformer des corrélations statistiques lexicales calculées en corpus en corrélats sémantiques interprétables en discours.

Cette contribution propose une analyse synchronique des grandes isotopies structurant le discours présidentiel d’Emmanuel Macron depuis son accession au pouvoir (148 discours, 501 817 occurrences).

Le corpus est filtré morpho-syntaxiquement sur les seuls substantifs, réputés plus immédiatement porteurs de sens. L’étude croise alors deux approches qui apparaissent complémentaires. D’une part, l’analyse des profils cooccurrence des mots, issue de l’ADT classique, d’autre part l’analyse des profils attentionnels des mots (self-attention), fondée sur des modèles neuronaux récents de type *Transformer*,

Ce regard croisé entre ADT et IA permet d’identifier quatre isotopies majeures du discours présidentiel : le régalien, l’économique, le social et le sociétal.

La discussion interroge la plus-value effective des modèles *Transformer* dans l’outillage méthodologique de l’ADT et questionne leur capacité à dépasser, ou à reformuler, les principes distributionnels classiques et la notion clef en analyse du discours : l’isotopie.

Mots clés : cooccurrence, self-attention, transformer, isotopie, AFC, textométrie, logométrie, discours politique, Emmanuel Macron.

1. Introduction

Le passage des réalités statistiques du corpus aux structures de sens des discours constitue l'un des problèmes majeurs de l'ADT. En effet, si les traitements quantitatifs de données textuelles permettent aisément d'objectiver des phénomènes linguistiques à grande échelle et de caractériser un texte ou un auteur, leur interprétation en termes sémantiques ne va jamais de soi. Dans ce cadre, l'identification automatique d'isotopies – entendues comme les régularités sémantiques structurant les textes et leur lecture – cristallise, pour la discipline, la tension fondatrice entre calcul et interprétation.

Dans le champ du discours politique, le passage du chiffre au sens est particulièrement sensible. De fait, les discours présidentiels étudiés sont fortement ritualisés et sont produits en masse : ils offrent pour cela un terrain très favorable à l'analyse quantitative. Cependant, ils posent la question concrète pour l'analyse du discours, les sciences politiques, l'histoire ou la sociologie de l'interprétation des régularités textuelles dégagées par la statistique. Précisément, la détection d'isotopies vise à dépasser la simple analyse de contenu à base de fréquences lexicales pour accéder à des organisations textuelles ou distributionnelles plus profondes susceptibles d'éclairer le sens et les fondements politiques, historiques, idéologiques des discours, ici, ceux d'Emmanuel Macron.

Parallèlement, l'essor récent des modèles neuronaux de langage, et en particulier des architectures de type *Transformer*, a renouvelé les outils disponibles pour l'analyse des textes. Fondés sur des mécanismes de self-attention (Vaswani *et al.*, 2017), ces modèles proposent une autre manière de capturer ou de formaliser les relations entre mots dans les textes, c'est-à-dire d'appréhender de manière systématique les co(n)textes locaux des mots, producteurs de sens, jusqu'ici traités en ADT sous la forme de collocations (via les concordanciers par exemple) ou par des calculs savants de cooccurrences. Toutefois, l'intégration des *Transformers* dans le champ de l'ADT soulève une question : les mécanismes attentionnels entre les mots-neurones des textes constituent-ils une rupture méthodologique effective, ou constituent-ils une simple reformulation computationnelle de principes distributionnels déjà bien établis, notamment dans le traitement de n-grams, des segments répétés, des motifs ou plus généralement de cooccurrences ?

C'est à cette question qu'essaye de répondre le présent papier, en confrontant deux approches – cooccurrence et attentionnelle – appliquées à un même corpus politique : l'ensemble des discours présidentiels d'Emmanuel Macron depuis 2017. L'hypothèse défendue est que, malgré leurs différences conceptuelles et algorithmiques, les deux approches tendent à faire émerger des isotopies macroniennes largement convergentes. Cette convergence invite d'un point de vue analytique à confirmer que les discours de Macron s'organisent autour de quelques pôles isotopiques homogènes facilement identifiables par une approche distributionnelle : le régalién, l'économique, le social et le sociétal. D'un point de vue méthodologique, elle interroge la plus-value supposée des modèles *Transformer* pour l'ADT qui ne semblent finalement qu'apprendre le profil cooccurentiel des mots dans le corpus.

2. Cadre théorique et méthodologique

2.1. Isotopie, isotropie et structuration du sens

La notion d'isotopie, directement issue de la sémantique structurale (Greimas, 1966 ; puis Rastier, 1972 et 1987), désigne la récurrence de traits assurant la cohérence ou l'homogénéité sémantiques d'un texte ou d'un corpus. En analyse du discours politique, elle permet de

dépasser une approche du lexique dictionnaire ou ontologisante, ou une simple analyse de contenu, pour mettre au jour les régularités et le « réglage du sens » (Siblot, 1989 : 107) qui traversent et organisent les textes, et participent, au bout de la lecture, à la construction des représentations sociales et idéologiques du locuteur/récepteur.

Ainsi, dans le cadre de l'ADT, l'établissement d'isotopies est le résultat d'un processus sémantique interprétatif nourri et encadré par le traitement statistique des données textuelles du corpus susceptible d'objectiver la fabrique du sens.

D'une certaine manière, la recherche des « mondes lexicaux » par la Méthode Reinert a été et reste une proposition intéressante, toujours actuellement implémentée dans les logiciels Alceste et Iramuteq (Reinert, 1983). Mais ce sont les travaux de Viprey (1997, 2006) qui apparaissent les mieux aboutis linguistiquement, en reprenant explicitement les approches distributionnelles harrissiennes et les fondements du contextualisme de Firth ou de Halliday.

L'idée principale repose sur l'hypothèse selon laquelle l'organisation d'un texte, sa progression thématique, sa cohérence sémantique relèvent de la micro-distribution des unités lexicales tout le long du corpus, c'est-à-dire, pour l'ADT, relève de la compilation systématique et chiffrée des co(n)textes locaux d'utilisation des mots dans l'ensemble des textes. La forme minimale de ce co(n)texte d'utilisation des mots peut être définie par la cooccurrence (lorsque A et B apparaissent ensemble statistiquement (cooccurent), ils se contextualisent minimalement l'un l'autre) ; et la forme aboutie de ce co(n)texte d'utilisation d'un mot devient son profil cooccurentiel général, c'est-à-dire l'ensemble des mots qui cooccurrent avec lui à un moment donné. A noter que cette idée, déjà présente dans Benzécri (1973), est fort répandue aujourd'hui dans les *embeddings* et leurs traitements type Word2Vec (Mikolov *et al.*, 2013) : un mot peut être représenté par une suite numérique puis par un vecteur constitués de la fréquence de ses cooccurrences avec les autres mots du corpus.

Dès lors, charge à nous, et à des logiciels, ici comme Hyperbase (fonction *Corrélat*), de construire des matrices de cooccurrences, ramenées dans cette contribution, comme nous le verrons, aux seuls substantifs du texte, et à traiter ces matrices par des méthodes benzécriennes de représentation vectorielle pour voir apparaître sur le plan factoriel les isotopies du discours ; ce que Jean-Marie Viprey nomme plus factuellement, en regardant les graphiques, « isotopies » (iso/tropes : les mots qui ont la même direction vectorielle et apparaissent ainsi dans les mêmes quadrants. cf. section 4. *Résultats*).

2.2. Self-attention et distribution neuronale

Le renouveau du connexionisme ces dernières années se traduit aujourd'hui par le triomphe des modèles *Transformer* (Vaswani *et al.*, 2017 déjà cité, et nombre de publications sur le sujet comme, précédemment, Bahdanau *et al.*, 2015). Les architectures *Transformer* reposent sur des mécanismes de self-attention qui permettent d'apprendre et de pondérer dynamiquement, au sein du corpus d'apprentissage, les relations à courte distance (bi-gram par exemple) ou à plus forte distance entre les unités du texte. Nous avons proposé dans plusieurs travaux récents d'exploiter de manière heuristique et descriptive ces mécanismes d'attention (Vanni and Mayaffre, 2026a ; Mayaffre et Vanni, 2026b), en considérant l'attention et plus généralement le profil attentionnel des mots comme un nouveau chiffrage ou une nouvelle représentation de la distribution co(n)textuelle, des mots entre eux dans le corpus (Cf. également, ici même, Vanni, Longrée and Mayaffre, 2026).

Appliqués ici aux seuls substantifs du texte, susceptibles de porter directement du sens, nous postulons ainsi que l’attention apprise par les modèles révèle des rapports isotopiques – on pourrait dire également de manière générique « paradigmatiques » – qu’entretiennent entre eux les mots du discours. La forte cohérence sémantique que l’on peut constater aujourd’hui dans la génération de textes type ChatGPT, DeepSeek, Gemini, Claude ou Mistral, etc. relèverait de cette capacité à prédire les mots isotopes – on pourrait dire aussi des mots appartenant aux mêmes *classèmes* – nécessaires à la pertinence sémantique du propos ; le cas le plus trivial, jusqu’à la caricature, étant celui des synonymes. On lira nécessairement à ce propos, et dans le cadre de la sémantique interprétative des textes : (Cavazza, 2026).

C’est en tout cas cette perspective qui nous invite à comparer directement le traitement de la cooccurrence généralisée (ou *corrélats*) *supra* présentée et le traitement de l’attention (self-attention), en examinant les isotopies respectives que semblent mettre au jour – si nos hypothèses sont bonnes – les deux approches dans les sorties-logicielles.

3. Corpus et paramétrages

3.1. Corpus

Le corpus étudié, que l’on pourra appeler corpus de travail en ADT ou corpus d’apprentissage spécifié en IA, comprend 148 discours présidentiels d’Emmanuel Macron prononcés depuis son accession au pouvoir. Il totalise 501.817 occurrences après normalisation. Le choix d’un corpus non contrastif répond à notre objectif de recherche : analyser la structure sémantique interne du discours présidentiel sans la rapporter à d’autres productions politiques. Le choix du corpus Macron relève, lui, d’une question d’actualité et de l’histoire politique française : presque 10 ans après l’accession au pouvoir de Macron, le bilan d’un discours qui apparaît critique pour le bon fonctionnement démocratique semble désormais nécessaire (*MOTS* 2026 ; Viktorovitch, 2025 ; Mayaffre, 2021).

3.2. Etiquetage et filtrage morpho-syntaxique

Le corpus a été étiqueté morphosyntaxiquement avec SpaCy, afin de pouvoir le filtrer selon les parties du discours, et ne conserver pour l’analyse que les substantifs. Ce filtrage repose sur l’hypothèse, classique en ADT, selon laquelle les substantifs constituent des unités privilégiées pour accéder au sens (sous conditions de les co(n)textualiser évidemment), et repérer ainsi plus facilement les isotopies. Nous assumons donc la perte de certaines dimensions du discours notamment les phénomènes énonciatifs, classiquement abordés par les pronoms, les adverbes déictiques ou les verbes modaux qui structurent avec tout autant de pertinence un discours.

Plus précisément, sur une base fréquentielle, nous avons retenu seulement les 300 substantifs les plus fréquents qui constituent à eux seuls les 2/3 des occurrences substantivales du corpus (62,8%). Deux avantages majeurs justifient cette réduction fréquentielle. D’abord en se limitant à 300 mots, la visualisation des données dans les graphiques sera nettement supérieure, en ne perdant pourtant qu’un tiers de l’information générale (37,2%) ; traiter du TLE avec son nombre considérable d’hapax ou de basses fréquences aboutit hélas le plus souvent à des graphiques illisibles. Ensuite, sélectionner seulement les substantifs très fréquents permet de construire des matrices – notamment la matrice cooccurrentielle – pleines ; les matrices creuses, constituées de basses fréquences, étant toujours plus difficiles à traiter statistiquement.

La construction de la matrice cooccurentielle n'a pas besoin d'être rappelée. Elle croise classiquement en ligne et en colonne les mots du texte dans un tableau de contingence carré symétrique (ou triangulaire de part et d'autre de la diagonale) : dans chaque cellule, la fréquence de la cooccurrence de A et B, de A et C, de A et D..., de Y et Z ; chaque ligne représente le profil cooccurentiel complet des mots A, B, C, D,... Z.

Enfin, les deux matrices ont été soumises à un traitement factoriel (AFC) selon les canons de la statistique exploratoire multidimensionnelle (Benzécri, 1973 ; Lebart *et al.*, 1995), implémentée ici dans le logiciel Hyperbase.

[illegible]

Figure 1. AFC des profils cooccurentiels des 300 substantifs du discours d'Emmanuel Macron



L'organisation lexicale du corpus Macron via ses substantifs, telle que la donne à voir le traitement factoriel de la cooccurrence d'une part et le traitement factoriel de la self-attention d'autre part (figure 1 et 2) semble attester de 4 isotopies majeures dans le discours : les 4 quadrants des deux graphiques respectifs mettent en effet en évidence, avec une rare cohérence sémantique, un vocabulaire isotope (ou « isotrope ») régalien, une isotopie (ou « isotropie ») économique, un vocabulaire isotope social et une isotopie sociétale.

Reste que la répétition quasi à l'identique de l'information des deux graphiques, nous permet de conclure que la profondeur du discours macronien tient dans sa capacité à structurer 4 pôles lexicaux, dont la cohérence semble suffisante pour être interprétée comme des isotopies : le corpus Macron peut se lire comme un discours recouvrant en matière (i) de postures régaliennes, (ii) de politique économique, (iii) de réflexions sociales, et (iv) de considérations sociétales.

A la fois gardien des Institutions et chef des armées, le président de la République mobilise/construit dans ses discours l'isotopie régalienne (quadrant haut/gauche des figures). En articulant des mots des profils distributionnels comparables comme « peuple », « constitution », « unité », « histoire », « indépendance » ou encore « frontière », « continent », « paix » ou « démocratie », le président incarne l'autorité souveraine et vitale du pays, cherchant ainsi à s'échapper des contingences politiciennes du moment. Comme en témoigne l'approche quantitative, les extraits de discours qui prennent ainsi de la hauteur politique sont nombreux dans le corpus, comme par exemple, entre mille, ce passage consacré à l'Europe :

La première, c'est que si nous décidons d'abandonner notre attachement à la démocratie et tout ce qu'elle emporte en Europe, nous ferons fausse route. La deuxième, c'est que nous pouvons dans ce cadre, et nous devons construire, une nouvelle souveraineté européenne par laquelle nous apporterons la réponse claire, ferme à nos concitoyens que nous pouvons les protéger, apporter une réponse à ces désordres du monde. [Macron, 18 avril 2018]

Particulièrement, la question de la « défense », des « frontières », de la « puissance » de la « sécurité » polarise le discours d'un locuteur que les Français doivent lire, entendre et interpréter comme un chef de guerre notamment autour du conflit en Ukraine ou au Moyen-Orient.

4.2. Isotopie économique

Pourtant, le président de la République ne structure pas seulement son discours autour des prérogatives supérieures du chef de l'Etat. Il considère également, plus prosaïquement, la question économique et budgétaire avec récurrence. Rançon, peut-être, d'une époque où les crises économiques se succèdent et de ses convictions néo-libérales souvent affirmées, Macron parle aux Français en termes de « production », d'« impôt », « d'entreprise », d'« argent », de « taxe », de « prix », « d'innovation » qu'ils distribuent et tissent dans le discours au sein de co(n)textes cooccurentiels ou attentionnels identiques (quadrant haut/droit des figures). L'isotopie de l'économie permet de construire l'image d'un dirigeant compétent et apparaît comme un des principes de lecture pertinent du texte macronien. Ici aussi, les extraits que nous pourrions produire en masse sont éloquents. Par exemple :

Il est impossible de distribuer quand on ne produit pas assez. L'argent qui s'investit, qui circule, qui crée de l'emploi n'est pas celui qu'on thésaurise jalousement. C'est pourquoi ceux qui risquent, qui osent, qui entreprennent doivent nous trouver à leurs côtés. La taxe à 75 % n'a pas créé d'emplois ni amélioré la condition de qui que ce soit en France. [Macron, 9 juillet 2018]

4.3. Isotopie sociale

Jamais très loin de l'isotopie économique, mais néanmoins distincte dans le discours et sur les figures (quadrant bas/droit), l'isotopie sociale structure également les discours de Macron. Autour des mots « emploi » ou « chômage », « retraite », « hôpital », « soignants », « jeunes », Macron produit un discours à la tonalité sociale.

Ici, les deux graphiques permettent de remarquer au sein de ce quadrant, l'importance forte et particulière de la thématique de « l'enseignement », de la « formation », des « élèves », des « classes », des « parents », des « écoles », de la « rentrée ». Très critiqué sur sa politique sociale en faveur des classes les plus favorisées, Macron peut faire ainsi valoir sa fibre sociale grâce à la promotion de tous par l'éducation, dans la droite ligne de penseurs libéraux-sociaux comme John Rawls. Ainsi par exemple, la mesure phare qu'il fait souvent valoir est le dédoublement des classes de CP susceptible de corriger les inégalités sociales :

Car les enfants déscolarisés à cet âge ne rattrapent plus leur retard de socialisation et d'apprentissage. Or, ce sont les enfants des milieux modestes qui étaient le moins scolarisés ou qui ne l'étaient pas. Cela entrera en vigueur

dès la rentrée 2019. Le dédoublement des classes de CP et CE1 en zone d'éducation prioritaire dites REP et REP+ est une mesure de justice sociale plus efficace que tous les dispositifs de redistribution, parce qu'on y distribue de manière différenciée le savoir fondamental, celui de lire, d'écrire, de compter, de se comporter. [Macron, ibid.]

4.4. Isotopie sociétale

Enfin pour que la sémantisation du discours de Macron soit complète, nous pouvons examiner le quadrant bas/gauche des figures qui rassemble un vocabulaire au profil cooccurenciel et au profil attentionnel identiques : des mots qui semblent relever de l'isotopie sociétale.

Moins dense que les autres quadrants, le quadrant bas/gauche réunit moins de mots que les autres, mais souligne comment le président vient dans son discours sur le terrain des enjeux de la société contemporaine, souvent à tonalité morale ou éthique.

Dans ce cadre interprétatif, c'est principalement la question des « violences » faite aux « femmes » ou de « l'égalité » entre les « hommes » et les « femmes » qui tisse le discours et permet de l'interpréter. La considération empathique des « victimes » – mot central sur les deux figures – devient un des modes de lecture du discours *care* de Macron (Mayaffre et Vanni, 2023).

Au-delà de la question féministe, ce sont des mots axiologiques ou célébrant des valeurs qui harmonisent alors sémantiquement le discours : « l'égalité » homme-femme nous l'avons dit, mais aussi la « dignité » ou le « respect », la « peur » ou les « droits ».

En parler et dire à toutes les femmes qui nous écoutent ce soir et qui sont victimes de violence, d'humiliations, de dire il y a un numéro d'urgence qui existe de le dire, d'en parler, de leur dire que ça n'est pas une honte. Ce ne sont pas elles les honteuses. Ce sont eux. [Macron, 15 octobre 2017]

5. Conclusion

L'analyse conduite sur le corpus des discours présidentiels d'Emmanuel Macron avait un double objectif : identifier les grandes isotopies structurant ce discours et évaluer l'apport méthodologique des modèles neuronaux de type *Transformer* pour l'analyse de données textuelles. En confrontant l'analyse des profils cooccurenciels des substantifs d'Emmanuel Macron, issue de la tradition statistique de l'ADT, à l'analyse des profils attentionnels issus de l'Intelligence artificielle, nous avons cherché à mesurer dans quelle mesure ces deux approches conduisent à des lectures sémantiques similaires du corpus Macron.

Les résultats montrent en effet une forte convergence entre les deux méthodes. Les analyses factorielles des matrices de cooccurrences et d'attention révèlent des structurations lexicales similaires, organisées autour de quatre pôles isotopiques majeurs : régalien, économique, social et sociétal. Malgré leurs différences algorithmiques, les deux approches capturent donc largement les mêmes régularités distributionnelles du vocabulaire présidentiel. Cela relativise l'idée d'une rupture introduite par les modèles *Transformer* dans l'analyse des textes : les modèles neuronaux attentionnels apprennent et restituent des relations lexicales proches de celles mises en évidence par l'analyse cooccurentielle traditionnelle.

Sur le plan de l'analyse du discours, ces résultats confirment que le discours présidentiel d'Emmanuel Macron s'organise autour de champs sémantiques structurants l'écriture, la lecture et l'interprétation, soit ce que la sémantique interprétative nomme « isotopies ». L'isotopie régalienne renvoie à la fonction institutionnelle et souveraine du chef de l'État et constitue la première clef de lecture du texte présidentiel. L'isotopie économique met en avant les compétences de Macron à promouvoir la production de richesses et la compétitivité du pays. L'isotopie sociale mobilise les thèmes de l'emploi, de la santé et principalement de l'éducation pour donner une tonalité progressiste au discours. Tandis que l'isotopie sociétale articule des valeurs normatives telles que l'égalité entre les hommes et les femmes, la dignité ou les droits. Ensemble, ces pôles constituent l'ossature sémantique du discours présidentiel, ses grilles de lecture, sa régulation du sens ; et participent à la construction d'une image politique combinant autorité institutionnelle, compétence économique, attention sociale et engagement normatif. Ils mettent au jour l'économie du texte macronien, ses règles de fonctionnement ou régulations sémantiques fondamentales ; pour faire sens.

Enfin, cette étude ouvre plusieurs perspectives de recherche. D'une part, l'exploration d'autres parties du discours (verbes, adjectifs, pronoms) permettrait de compléter l'analyse isotopique en intégrant, par exemple, les dimensions énonciatives et argumentatives du discours présidentiel. D'autre part, l'analyse contrastive avec d'autres présidences ou d'autres fonctions politiques (maires, députés, sénateurs, etc.) pourrait éclairer la spécificité ou la banalité des structures sémantiques observées ici chez Emmanuel Macron. Enfin, plus largement, la rencontre entre ADT et Intelligence artificielle appelle à poursuivre une réflexion théorique sur les rapports entre distribution statistique des mots, réseaux de neurones profonds appliqués aux vocabulaire et construction du sens, afin de mieux comprendre comment les régularités et représentations numériques du corpus peuvent être interprétées comme des organisations sémantiques pertinentes.

Références

- Bahdanau D. *et al.* (2015). Neural Machine Translation by Jointly Learning to Align and Translate, *Conference paper at ICLR 2015*, arXiv.1409.0473v7.
- Benzécri J.-P. *et al.* (1973). *L'analyse des données. Tome2, L'analyse des correspondances*, Bordas.
- Cavazza M. (2026). Sémantique des textes et grands modèles de langage, *Cahiers de sémiotique des cultures*, n°5– 1, 67-86.
- Greimas A. J. (1966). *Sémantique structurale*. Larousse.
- Lebart L., Morineau A. et Piron M. (1995). *Statistique exploratoire multidimensionnelle*, Dunod.
- Mayaffre D. (2021). *Macron ou le mystère du verbe. Ses discours décryptés par la machine*, L'Aube.
- Mayaffre D. (2023). Sémantique de corpus numérique et deep learning. Macron un président thaumaturge (2017-2023), *Espaces linguistiques*, n°6, 2023. [hal-04353800].
- MOTS (2026). Le discours macroniste. Un président en scène, n°140.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pages 3111–3119.
- Rastier F. (1972). *Systématique des isotopies*. Larousse.
- Rastier F. (1987). *Sémantique interprétative*. PUF.
- Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Cahiers de l'analyse des données*. Volume 8, n°2.

- Siblot P. (1989), Isotopie et réglage du sens, *Cahiers de praxématique* [En ligne], 12 | 1989 <http://journals.openedition.org/praxématique/3453> ; DOI : <https://doi.org/10.4000/praxématique.3453>
- Vanni L. and Mayaffre D. (2026a). Explore political discourse with transformers. Emergent paradigmatic and syntagmatic representations. *15th edition of the Language Resources and Evaluation Conference (LREC 2026)*, Palma de Mallorca, Spain.
- Vanni L. and Mayaffre D. (2026b). Une autre Intelligence Artificielle est-elle possible ? *Cahiers de sémiotique des cultures*, n° 5– 1, 113-144.
- Vanni L., Mayaffre D. and Longrée D. (2026). Decoding Self-Attention : How frequency matrices drive representation learning ?, in *JADT 2026*, Palermo.
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., and Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pages 6000–6010, Red Hook, NY, USA. Curran Associates.
- Viktorovitch C. (2025). *Logocratie*. Seuil.
- Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*. Champion.
- Viprey J.-M. (2006). Structure non-séquentielle des textes. *Langages*, 163, 71-85.

Les formes du futur en planification urbaine. Analyse textométrique des temps verbaux en planification régionale (Île-de-France, 1919-2024)

Isabelle Chesneau

ENSA Paris-Malaquais - PSL – isabelle.chesneau@paris-malaquais.archi.fr

Abstract

Does the use of the future tense as a grammatical form in urban planning documents reflect the evolution of how urban futures have been envisioned across decades? Have the uncertainties surrounding urban futures – stemming from multiple crises (economic, social, sanitary) – and the urgency imposed by climate change led planners and policymakers, over the past two decades, to transform the way they express futurity, shifting from a *will* to act to an *obligation* to act? These questions, debated within the emerging field of urban futures studies (Dixon and Tewdwr-Jones, 2021), are most commonly approached through a lexical lens (Tulumello, 2025) or a narrative one (Ammel, 2023); yet the verbal dimension – that is, the way futurity is modalized and temporalized in planning discourse – remains comparatively understudied. This article examines the extent to which verb tenses can shed light on the future as a temporal horizon, drawing on a textometric analysis of long-term regional planning documents. The study focuses on the capital region of France, Île-de-France, through a corpus of eight regional planning documents applied between 1919 and 2024. Over the course of a century, these documents have shifted in form – laws, development plans, master plans – yet retain a structural commonality: all articulate diagnoses, policy orientations, and arbitrations among stakeholders. The findings reveal that urban futures are not expressed through the simple future tense alone, and that a plurality of enunciative regimes of futurity coexists across the corpus. Marked differences emerge across periods: the simple future characterizes texts from the early twentieth century and the 1990s, while the *Trente Glorieuses* — the golden age of planning in France — shows a greater reliance on the conditional (future-in-the-past). Since 2010, the prospective dimension has increasingly favored the infinitive, expressing a future objective stripped of both a grammatical subject and verbal inflection. Each verbal configuration thus corresponds to a specific way of positioning public action between the irreversibility of choices and the openness of possibles.

Keywords: textometrics, verb tenses, urban futures, regional planning, retrospective approach.

Résumé

L'usage du futur comme temps grammatical dans les plans d'urbanisme reflète-t-il l'évolution de la vision de l'avenir des villes au fil des décennies ? Les incertitudes entourant les futurs urbains en raison de multiples crises (économiques, sociales, sanitaires) et l'urgence liée au changement climatique ont-elles poussé depuis une vingtaine d'années les urbanistes et les décideurs à transformer leur manière d'exprimer l'avenir, passant d'une volonté à une obligation d'agir ? Ces questions, discutées au sein d'un courant émergent d'*urban futures studies* (Dixon et Tewdwr-Jones 2021), sont souvent abordées sous l'angle lexical (Tulumello 2025) ou narratif (Ammel 2023), mais la dimension verbale, c'est-à-dire la manière dont le futur est modalisé et temporalisé en planification, reste en revanche peu étudiée. Cet article propose d'examiner dans quelle mesure les temps verbaux peuvent renseigner sur le futur comme époque, à partir d'une analyse textométrique menée sur des documents de planification régionale à long terme. L'enquête porte sur la région capitale, l'Île-de-France, à travers un corpus de huit textes d'aménagement du territoire appliqués entre 1919 et 2024. Sur un siècle, ces documents ont changé de forme – lois, plans d'aménagement, schémas directeurs –, mais conservent une proximité structurelle : tous exposent diagnostics, orientations et arbitrages entre acteurs. Les résultats révèlent que l'on ne parle pas du futur de la ville au futur simple et qu'il existe une pluralité de régimes d'énonciation de l'avenir. Des différences marquées apparaissent ainsi selon les périodes : le futur simple caractérise les textes du début du XX^e siècle et des années 1990, tandis que la période des Trente Glorieuses, âge d'or de la planification, recourt davantage au futur

du passé. Depuis 2010, la dimension prospective privilégie l'infinitif, exprimant un objectif futur sans sujet grammatical ni flexion verbale. On voit ainsi que chaque configuration verbale renvoie à une manière spécifique de situer l'action publique entre irréversibilité des choix et ouverture des possibles.

Mots clés : textométrie, temps verbaux, futurs urbains, planification régionale, approche rétrospective.

1. Introduction

Les incertitudes entourant l'avenir des villes, en raison de multiples crises (économiques, sociales, sanitaires) et de l'urgence liée au changement climatique, ont-elles poussé les urbanistes et les décideurs à transformer leur manière d'exprimer l'avenir au cours du temps, en donnant des inflexions à leur volonté d'agir en fonction des périodes ? Ces questions, discutées au sein d'un courant émergent d'*urban futures studies* (Dixon et Tewdwr-Jones, 2021), sont souvent abordées sous l'angle lexical (Tulumello, 2025) ou narratif (Ammel, 2023), mais la dimension verbale reste en revanche peu étudiée. Or, « Les Temps sont des formes qui ajoutent, à l'idée fondamentale de la signification du Verbe, l'idée accessoire d'un rapport d'existence à une époque » (Beauzée, 1767 cité par Apothéloz, 2023). L'usage des temps verbaux dans les documents de planification urbaine est-il source de compréhension des manières d'envisager l'avenir des villes et change-t-il au fil des décennies ? Dans cet article, nous nous proposons d'examiner dans quelle mesure le temps grammatical peut être source d'enseignement sur le temps comme époque, à partir d'une analyse textométrique menée sur des documents de planification de la région Île-de-France.

Pour effectuer cette analyse diachronique et déceler les évolutions grammaticales recherchées, une période suffisamment longue devait être envisagée. Ainsi, nous avons rassemblé un corpus de textes allant de l'émergence du droit de l'urbanisme jusqu'à aujourd'hui, afin de disposer d'une série textuelle s'étendant sur plus d'un siècle (1919-2024). Cette série chronologique comprend : (1) les lois Cornudet (1919 et 1924) et Sarraut (1928) ; (2) le Plan d'aménagement de la région parisienne de 1939 (PARP, dit Plan Prost) ainsi que les révisions de 1950 et de 1956 ; (3) le Plan d'aménagement et d'organisation générale de la région parisienne (PADOG) de 1960 ; (4) le Schéma directeur d'aménagement et d'urbanisme de la région parisienne (SDAURP) de 1965 ; (5) le Schéma directeur d'aménagement et d'urbanisme de la région Île-de-France de 1976 (SDAURIF) ; et enfin, les trois Schémas directeurs de la région Île-de-France (SDRIF) de 1994 (6), de 2013 (7) et le dernier (8), dit environnemental (SDRIF-E), approuvé en 2024. Reconstituer cet ensemble de huit groupes de documents s'est appuyé en partie sur une enquête archivistique et documentaire qui a montré l'importance des textes préparatoires et des mises en débat que ces documents d'orientation avaient générés. La notion de « document d'urbanisme » ne devait donc pas être prise au pied de la lettre. Ces textes d'orientation font partie d'un contexte et sont à envisager comme appartenant à un ensemble plus vaste, composé de textes délibératifs et justificatifs, dont la fonction est de discuter les visions de l'avenir urbain énoncées ou proposées et des changements à opérer. Nous avons donc intégré un certain nombre de ces documents préparatoires à notre corpus, tentant de reconstituer un « réseau de textes » autour de chaque document de planification, contribuant fortement à la construction de leur sens. Procéder ainsi était reconnaître à ces textes leur dimension foncièrement « polyphonique », « émanant de différentes voix, de différents points de vue » (Perrin, 2004). En outre, au cours de ces cent années, ces documents ont changé de forme, passant de lois à des plans d'aménagement, puis à des schémas directeurs, et connu différents profils de rédacteurs (députés, hauts fonctionnaires, urbanistes, etc.). Malgré cette hétérogénéité, tous ces documents ont en commun de présenter les grandes orientations de l'avenir de la région parisienne à plus ou moins long terme (5, 10 ou 15 ans, selon les périodes),

en s'appuyant sur une évaluation de la situation en cours. Ainsi, malgré leurs différences formelles, ils présentent de nombreuses similitudes, formant une cohérence d'ensemble (Pincemin, 2012). Le corpus ainsi constitué contient, *in fine*, 54 textes et représente un volume de 1 278 893 mots. Pour analyser les temps verbaux, nous avons utilisé le logiciel TXM, sans procéder à une annotation manuelle préalable du corpus, dans l'objectif de tester notre hypothèse sur le caractère heuristique des temps verbaux, avant d'approfondir l'analyse de manière plus précise.

L'expression de la futurité – entendue comme « tout emploi d'un temps verbal désignant un moment postérieur à l'énonciation » (Apothéloz, 2023) – est une notion plus complexe qu'il n'y paraît. Elle ne se réduit pas au futur simple et peut s'exprimer, par exemple, par un futur antérieur, un futur périphrastique (*aller* + *inf*), un conditionnel ou bien encore un présent accompagné de localisateurs temporels (*demain*, *dans dix ans*). Même en se concentrant sur le seul futur simple, on considère généralement (*id*) que ses emplois se partagent entre une valeur temporelle, où le temps verbal situe le « procès » dans l'avenir, et une valeur modale qui porte sur le « moment » de l'énonciation lui-même plutôt que sur le procès. Non seulement cette dichotomie entre temporalité et modalité résiste parfois mal à l'analyse (Gosselin, 2005), mais les temps verbaux s'associent en outre à des verbes lexicaux qui « possèdent eux-mêmes leurs propres caractéristiques temporelles et aspectuelles » (Apothéloz, 2021).

Comme notre objectif est de voir si la fonction grammaticale des verbes est propre à traduire l'idée du futur comme époque, nous adopterons une définition de la futurité délibérément resserrée, considérant d'une part les formes verbales exprimant un « projeté dans l'avenir » et, d'autre part, les verbes lexicaux à valeur prospective. Nous excluons pour le moment le présent avec un sens prospectif, trop complexe à saisir dans une démarche diachronique. L'approche textométrique observée ne permettra pas non plus d'identifier systématiquement les valeurs modales que ces formes verbales prennent en contexte. Elle permettra néanmoins un repérage systématique des formes concernées et des périodes au cours desquelles leur distribution se reconfigure. C'est à partir de ce premier cadrage que nous analyserons, dans la suite de l'article, la façon dont les discours de planification urbaine ont, tout au long du siècle étudié, articulé la projection de l'avenir urbain à différents registres d'action, comme la nécessité, la possibilité ou l'obligation d'agir.

2. Un emploi des temps verbaux contrasté selon les périodes

Dans un premier temps, il est nécessaire de visualiser la distribution globale des temps verbaux dans notre corpus pour apprécier la fréquence d'utilisation du futur simple. Dans TXM, à partir d'une partition organisée autour de chaque groupe de textes de planification, nous avons ainsi recherché les temps simples de l'indicatif, du conditionnel et des modes impersonnels (infinitif, participe) – en excluant le subjonctif et l'impératif trop peu présents dans le corpus – de tous les verbes du corpus, à l'exception des auxiliaires *avoir* et *être*, pour éviter de prendre en compte les temps composés à ce stade¹. Dans un second temps, nous avons cherché à distinguer les emplois du futur simple, du futur composé et du futur périphrastique.

La figure 1 ci-après montre que, dans ces textes consacrés à énoncer les orientations d'avenir de l'Île-de-France, le futur simple (rouge) n'est pas le temps verbal le plus courant. Il est

¹ Les formes verbales ont été extraites à l'aide d'une requête CQP construite sur le modèle [frpos = "VER.*" & frpos = "VER:Tverb" & flemma != "avoir" & flemma != "être"], où la valeur « Tverb » de frpos varie dans chaque partie de la requête (*futu*, *cond*, etc.) et où les auxiliaires *avoir* et *être* sont exclus par la contrainte « flemma != ». Ces sept requêtes partielles sont reliées entre elles par disjonction « | ».

caractéristique des textes du début du XX^e siècle (gp 1) et des années 1990 (gp 6), tandis que le conditionnel (orange) constitue un temps plus spécifique des textes des années 1960 (gp 3 et 4). Aux autres périodes, comme dans les années 1970 ou depuis les années 2010, les temps verbaux liés au futur sont fortement sous-représentés, laissant place à d'autres structures grammaticales. Cela suggère que la projection vers l'avenir ne s'exprime pas nécessairement au futur et qu'il convient de s'intéresser à tous les temps verbaux présents dans le corpus.

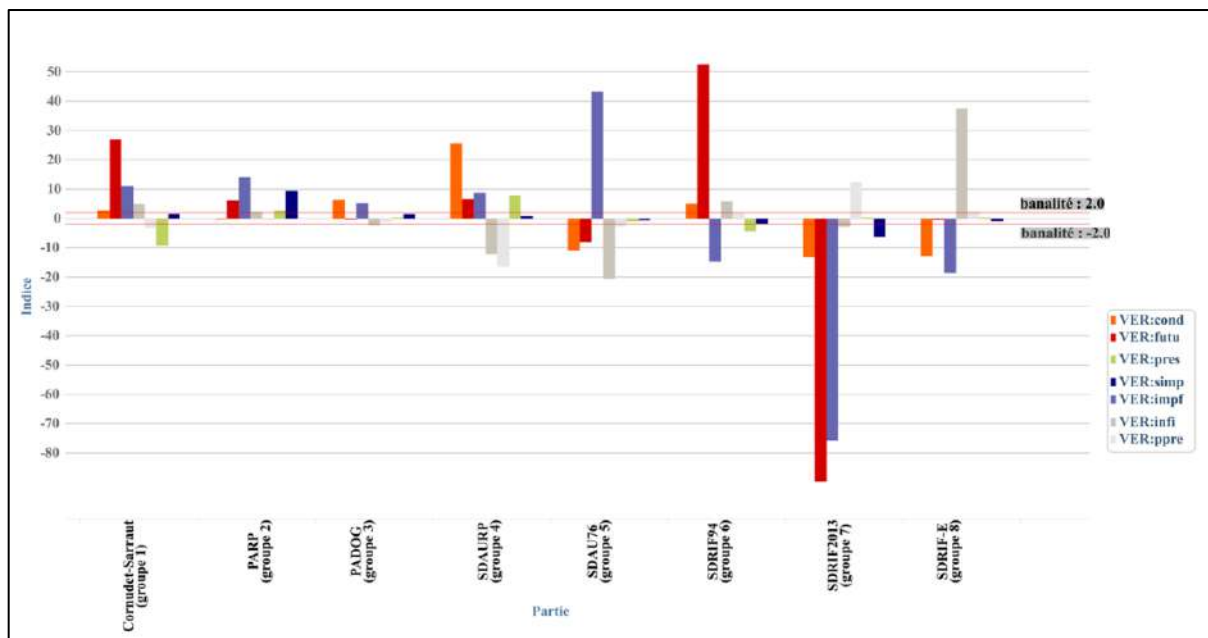


Fig. 1. Répartition des temps simples de l'indicatif, du conditionnel et des modes impersonnels (infinitif, participe) des verbes des huit groupes de textes du corpus planification (sans les auxiliaires avoir et être)

En comparant la spécificité de chaque temps verbal selon les périodes, on peut distinguer quatre familles d'emploi dans l'ensemble des textes. Les groupes 1, 2 et 6 favorisent principalement le futur (rouge), les groupes 3 et 4 le conditionnel (orange) en association avec d'autres temps, comme le futur périphrastique et le présent, le groupe 5 est dominé de manière surprenante par l'imparfait (bleu), tandis que les deux derniers (7 et 8) sont centrés sur l'infinitif (gris foncé) et le participe présent (gris clair). Pour examiner les conceptions du futur sous-jacentes à chacune de ces constructions grammaticales, l'utilisation d'un calcul de concordances dans TXM est indispensable pour passer des constats statistiques à une compréhension du contexte d'énonciation.

2.1. Le futur simple en lien avec le déroulement d'un procès à venir

Dans les groupes (1) et (6), le futur simple porte sur un petit nombre de verbes spécifiques (≥ 10 occ.). Ceux qui ont la spécificité la plus forte au cours de la première moitié du XX^e siècle (gp 1) sont *pouvoir*², *suivre*, *fixer*, *déterminer*, *exercer* et *avoir* (non auxiliaire) ; et ceux de la fin de ce même siècle (6) sont : *devoir*, *falloir*, *convenir*, *appuyer*, *relier*, *faire* et *traduire*. Dans les deux cas, ces verbes sont essentiellement associés à des actes directifs (injonctions, consignes, interdiction, autorisation, etc.), à destination soit de l'autorité publique, soit de ceux

² Nous n'avons pas exclu « *devoir* Vinf », « *pouvoir* Vinf », « *falloir* Vinf », *convenir* Vinf » souvent considérés comme périphrases modales (Gosselin, 2020), car leur sens modal ne semblait pas avéré dans ces textes de nature juridique.

qui veulent bâtir. C'est sans doute la raison pour laquelle ces verbes au futur simple entretiennent, dans le premier groupe de texte, un lien fréquent avec la négativité : « Les lotisseurs *ne* pourront *pas* user [...] de la faculté de délaissement prévu par l'article... ». Il est logique que ces textes expriment ce qui est permis et interdit, ou des obligations, car leur but est de normer les comportements à venir, afin de produire certains types d'actions et d'en éviter d'autres. Ainsi, dans les lois du début du XX^e siècle (lois Cornudet et Sarraut), on peut lire : « L'autorité expropriante *pourra* en revendant les terrains, inscrire dans le cahier des charges... », « Il est déclaré d'utilité publique par une loi qui *fixera* les mesures nécessaires à son application... », « Ce règlement *déterminera* notamment les modalités des contrats... » ; « L'interdiction de construire *s'exercera* sans réserve pendant trente ans... ».

Comme on le voit, « la futurité est ici bien réelle et concerne le procès proprement dit. Il s'agit donc d'effets pragmatiques accompagnant une interprétation littérale du Futur. » (Apothéloz, 2023). Toutefois, le déclenchement pragmatique de ces valeurs directives semble lié, outre l'expression de la futurité, au type de verbes employés. Au début du XX^e siècle, le verbe *pouvoir* sert principalement à exprimer le sens des actions à engager. Or, ce verbe est ambigu, il peut exprimer soit le « possible » – une potentialité qui détermine le réel (ce que l'on peut faire ou ne pas faire) – soit le « pouvoir » (commander, avoir autorité), à mettre en rapport avec le devoir et l'obligation. Ayant remarqué la polysémie de ce verbe, le député Pierre Dormoy souligne, en 1923, qu'il constitue l'un des points faibles de la loi Cornudet :

« [i]l ne faut d'ailleurs pas se faire d'illusions et avec ce mot “pourra”, les nouvelles dispositions courent grand risque de n'être pas plus appliquées que les anciennes, car à défaut de précision, dans l'arrêt préfectoral, les lotisseurs conservent toute leur liberté. »

En d'autres termes, sans affirmation claire de la puissance publique, le champ des possibles que le législateur cherche à contenir, voire à contrôler, reste entier.

Presque quarante ans plus tard, au milieu des années 1990, le caractère directif paraît tout aussi ambigu et relatif. *Devoir*, *falloir* et *convenir* ne sont en effet pas moins polysémiques que *pouvoir* et on lit ainsi dans le SDRIF 1994 (6) : « Les villes nouvelles *devront* conforter leur acquis de pôles régionaux », « Pour loger 2 775 000 Franciliens, il *faudra* construire 1 325 000 logements sur 25 ans », « il *conviendra* de valoriser le patrimoine architectural ». Dans ces textes à caractère institutionnel, l'ambiguïté semble découler d'une euphémisation de la modalité déontique de ces verbes. La source déontique reste indéterminée : le sujet est le plus souvent un « il » impersonnel qui ne désigne aucun responsable identifié (Buhler, Chesneau et Richeton, 2024), tandis que le futur atténue l'injonction en lui substituant une nécessité aléthique, comme si l'obligation découlait du contexte plutôt que d'une autorité.

2.2. Le futur antérieur en lien avec le résultat d'un procès à venir

Le futur simple est, dans les groupes 1, 2 et 6, souvent associé au futur antérieur et devance même le futur simple dans les groupes 1 et 2. On peut ainsi lire dans ces textes :

« Quand il *aura été voté* par le Parlement, les principales difficultés en présence desquelles se trouvent les municipalités *auront disparu* » ; « Les subventions de l'État ne *seront accordées* aux communes qu'après l'établissement de ce plan » ; « il *ne pourra* plus être édifié de constructions nouvelles [...] que suivant les alignements fixés » ; « si l'existence de construction est incompatible avec l'aménagement prévu par le plan et le programme approuvé, l'autorisation de construire ou de réparer *sera refusée* » ; « Les frais de ce projet *seront supportés* et les subventions de l'État *seront allouées* dans les conditions fixées par les lois... » ;

« Des actions très volontaires *devront* y être menées pour réduire les disparités » ; « Dans les espaces paysagers *pourront* toutefois être autorisés les golfs ».

Avec le futur antérieur, il s'agit moins d'expliquer un procès que les conditions de sa réalisation (*si, dans les conditions, dans le cas où, etc.*), le moment (*quand, après, etc.*) auquel il doit se produire ou bien encore le besoin de récapituler une série de mesures et d'en faire le bilan. L'emploi du futur antérieur est ainsi une manière de mettre en évidence le résultat attendu des nouvelles mesures imposées par les lois ou les plans d'aménagement.

2.3. Le futur périphrastique en lien avec un état actuel

Le futur périphrastique avec *aller* est peu présent dans le corpus : seuls les groupes 4 et 8 y recourir de manière spécifique.

On peut lire dans le SDAURP de 1965 (gp 4) : « Le “temps libre” – ainsi appelé faute d'un meilleur terme – *va* progressivement gagner sur le temps de travail professionnel » ; « Le nombre des voitures particulières *va* continuer à s'élever ; aujourd'hui en région de Paris, il en existe une pour cinq habitants, l'an 2000 en verra une pour trois habitants environ » ; « Les activités les plus librement choisies – acheter, se divertir – *vont* prendre une place croissante dans la vie urbaine ». Et, dans le dernier document de planification de l'Île-de-France (groupe 8) datant de 2024 : « Le SDRIF-E *va* créer 27 centralités et 117 polarités qui constitueront des pôles attractifs, autonomes et reliés entre eux, qui permettront de créer ce lien renforcé et vivant » ; « Nous *allons* ainsi accélérer les modes de déplacement de proximité ».

Ces énoncés montrent ce que nombre de travaux ont déjà mis en évidence (Rebotier, 2015 ; Apothéloz, 2023). Alors que le futur simple se pose en rupture par rapport au contexte de l'énonciation (demain sera différent d'aujourd'hui), le futur périphrastique s'inscrit plutôt dans une continuité, comme si les changements à venir avaient déjà commencé à affecter la situation présente. L'emploi de la forme future périphrastique correspond peut-être à une stratégie d'énonciation visant à atténuer la crainte du changement, en le présentant comme un processus en cours plutôt qu'une rupture brutale. Ce résultat corrobore celui d'un précédent travail portant sur ce même corpus (Chesneau, 2026), dans lequel nous montrions que la notion de changement urbain, si centrale dans ces textes de planification, était un motif (Rastier, 1996) largement passé sous silence, présenté comme s'il était le fruit d'un accord tacite.

2.4. Le conditionnel comme ultériorité du passé

Le conditionnel présent est un temps verbal caractéristique du SDAURP de 1965, le document de planification qui marque la période gaullienne. En grammaire, il existe un débat (Bres, 2024) pour savoir si ce temps verbal est un temps de l'époque future ou celui de l'époque passée. Dans le premier cas, on le considère comme un futur hypothétique et, dans le second, comme un « ultérieur du passé », sans aucune charge d'hypothèse. Si l'on examine certains passages du SDAURP (gp 4), on observe que les actes décrits sont plutôt situés dans le passé :

- (1) « La progression *devrait*, d'ailleurs, être sensiblement la même pour toutes celles des autres régions de France » ;
- (2) « Au demeurant, de telles prévisions ne *seraient* pas définitivement trop larges, mais seulement, fait rare, quelque peu en avance sur l'événement » ;
- (3) « dans les zones qui se construiront dans les prochaines années, sans axes de transport nouveaux, il n'y *aurait* pas de villes nouvelles »

Dans (1), la progression [des déplacements quotidiens] des Franciliens est bien située dans le passé par rapport au moment où cette phrase est écrite. De même, dans (2), les prévisions ont déjà été réalisées. Dans le cas de (3), l'emploi du conditionnel indique que la décision de créer des villes nouvelles a été prise dans le passé. Ainsi, parce que la création de ces villes nouvelles a été décidée politiquement, il est nécessaire de créer de nouveaux axes de transport. Ici, Paul Delouvrier énonce moins un programme d'actions futures qu'il ne rapporte une politique déjà décidée et ne décrit un futur présenté comme inéluctable.

2.5. L'imparfait pour se distancier du passé et ouvrir le futur

Le SDAU 1976 (gp 5) se présente comme le prolongement du SDAU de 1965 (gp 4). Toutefois, dix ans se sont écoulés entre ces deux textes, et le nouveau se doit de tenir compte des évolutions démographiques et économiques (chocs pétroliers) qui ont marqué la région depuis le milieu des années 1960 : « la nouvelle carte de synthèse demande à être explicitée dans les domaines où les dix années d'expérience écoulées ont révélé l'insuffisance du document initial ». Dans cette mise au point de 1976, on ne trouve quasiment pas de verbes conjugués au futur et, à l'inverse, c'est l'imparfait qui domine. La projection vers le futur y est pourtant très forte, mais s'y exprime de manière spécifique et plus radicale. On y lit des propos du type :

« le parti d'aménagement retenu en 1965 *mettait* l'accent principal sur la création des centres urbains nouveaux plutôt que sur l'évolution de la zone agglomérée » ; « l'explosion urbaine *imposait* un nouvel état d'esprit et *exigeait* de nouvelles solutions d'aménagement et d'urbanisme pour faire face aux dimensions de l'avenir » ; « les villes nouvelles *apparaissaient* à la fois comme [...]. Elles *constituaient* ainsi l'action clé ».

Dans ce document de 1976, le passé est sans cesse appelé en référence, mais il ne sert pas à ancrer les analyses en cours dans des tendances longues, comme c'était l'usage antérieurement. Il s'agit au contraire de se distancier des pratiques et, surtout, des récits hérités (rééquilibrage régional, politique de desserrement, optimisme de "l'an 2000", etc.), d'en faire la critique afin de rompre avec et de nourrir une réflexion prospective ouvrant sur de nouveaux récits. Ainsi, l'usage de l'imparfait du SDAU 1976 met en évidence cette logique rétrospective mise au service d'un exercice de prospective, alors assez nouveau à l'époque.

2.6. Les temps impersonnels comme futur anhistorique

Les temps impersonnels (infinitif et participe présent) qui caractérisent les deux derniers groupes de documents de notre corpus, en particulier le SDRIF-E de 2024, se composent le plus souvent d'un seul élément verbal, comme dans les extraits suivants :

« La Nature et la biodiversité sont fragiles, les *préserver* et les *transmettre* est un devoir moral qui nous guide. » ; « Face aux forces de centralisation, un nouveau mode d'aménagement recousant les déchirures entre centre et périphérie doit être encouragé afin *d'éviter* l'égoïsme territorial d'un côté, le sentiment de relégation de l'autre » ; « Pour *vivre* mieux, nous devons aussi *protéger* mieux. La Nature en ville est un levier puissant pour *lutter* contre les canicules, mais aussi *réduire* l'imperméabilisation des sols et donc *réduire* les phénomènes d'inondation » ; « l'ambition d'*atteindre* la neutralité carbone en 2050, qui implique, en matière de transport, de *se déplacer* mieux, mais aussi moins loin, rendant nécessaire de *structurer* des bassins de vie plus resserrés ».

Dépourvu de marque personnelle et temporelle, l'infinitif est généralement considéré comme la forme nominale du verbe (Krazem, 2007), et peut être interprété comme l'expression d'une injonction ou d'une prescription, comparable à l'impératif. Bien que le locuteur ne soit pas désigné, implicitement, ici, la puissance publique est aisément identifiable. En cela, le dernier

document de planification est très représentatif des schémas directeurs qui ont vocation à énoncer les principales finalités de l'action urbaine à l'échelle régionale et à s'en tenir à un haut niveau de généralités, laissant aux documents inférieurs (PDU, PLH, PLU/PLUi) le soin de préciser les choses localement. Ces textes sont semblables à des lois, ils énoncent des principes généraux et essentiels, valant pour tous les contextes. Pour cette raison, l'infinitif et le participe présent visent à accentuer la détemporalisation – au sens où elles font abstraction du temps – et la décontextualisation des propos, afin de gagner en généralité. Il porte implicitement l'intention d'une réalisation future, mais sans en préciser la temporalité.

3. Temps linéaire, temps ramifié et conceptions de l'avenir dans les textes de planification de l'Île-de-France

Selon Gosselin (1996), la « temporalité verbale » correspond à « un mode de manifestation du temps dans et par le langage » : elle ne se limite pas à une question de grammaire. En ce sens, l'analyse de ces textes de planification suggère qu'ils véhiculent une conception spécifique de l'avenir et une relation particulière avec le possible et le nécessaire, qui varient selon les époques historiques. Les résultats présentés, en montrant que le futur simple n'est pas le temps grammatical dominant dans ces textes prospectifs, invitent à reconsidérer ce que « parler d'avenir » signifie dans le discours aménagiste. L'hétérogénéité des formes verbales repérées – futur simple, futur antérieur, futur périphrastique, conditionnel, imparfait, infinitif – révèle autant de postures énonciatives différentes à l'égard de l'avenir, ouvrant ainsi de nouvelles pistes pour une analyse plus approfondie.

À la lumière de la distinction proposée par Robert Martin (1981) entre temps linéaire et temps ramifié, on constate que leurs logiques respectives sont loin d'être mécaniques en planification urbaine et contredisent souvent certaines idées reçues. Martin distingue deux manières de concevoir le futur linguistique : d'un côté, il y a un temps linéaire qui correspond à un avenir tracé comme une continuation déterminée du présent, sans bifurcation possible ; de l'autre, un temps ramifié (*branching time*) qui représente l'avenir comme un ensemble ouvert de possibles à partir d'un point de ramification situé au présent. Nous voudrions confronter nos résultats à cette grille de lecture afin d'ouvrir des pistes d'interprétation et d'élargir les critères analytiques.

L'analyse menée montre que la logique du temps linéaire, généralement fortement associée à la démarche de planification, n'est pas celle qui domine ces textes, même lorsque les rédacteurs recourent, comme dans les groupes 1 et 6, au futur simple. Dans ces deux groupes de textes, le futur est sous la dépendance des verbes *pouvoir*, *fixer*, *déterminer* pour les lois Cornudet et Sarraut et *devoir*, *falloir*, *convenir* dans le SDRIF 1994. Ils apparaissent comme des opérateurs d'actes directifs renvoyant à des injonctions, des prescriptions ou des interdictions. Le futur simple semble fonctionner ici comme la marque d'un avenir posé comme nécessaire, voire inéluctable, celui de l'exécution d'un programme d'action publique, dont la puissance normative est censée réduire le champ des possibles. Toutefois, la logique linéaire de ces différents verbes semble prise en tension entre une ouverture à un champ des possibles des actions autorisées et une affirmation d'autorité. Cette tension entre potentialité et nécessité laisse à penser que le temps prétendument linéaire de la loi peut, contrairement à sa visée perfective, réintroduire une certaine ramification : l'avenir prescrit reste conditionné à la volonté des acteurs, sur lesquels la puissance publique n'exerce qu'une emprise partielle. De même, dans le SDRIF 1994, les formulations comme « il *conviendra* de valoriser... » ou « les villes nouvelles *devront* conforter... » relèvent davantage d'objectifs futurs souhaités, dont les conditions de réalisation demeurent incertaines et plurielles. Autrement dit, il faut distinguer

dans ces textes la valeur sémantique du futur grammatical des effets pragmatiques que son emploi peut déclencher dans certains contextes. Ces effets pragmatiques, notamment directifs, n'infèrent pas l'incertitude du procès, mais présupposent un futur ouvert que la norme cherche précisément à refermer. Tout le paradoxe du futur simple dans les textes législatifs est là : son emploi anticipe un temps ramifié – un espace de possibles à contraindre — tout en affichant rhétoriquement un temps linéaire.

Le conditionnel des textes gaulliens des années 1960 s'inscrit en réalité davantage dans une conception du temps linéaire. Le PADOG et le SDAURP (groupes 3 et 4) expriment explicitement une logique de temps ramifié, mais le point de ramification se situe plus souvent dans le passé que dans le présent de l'énonciation. Comme le résume Robert Martin, ce conditionnel est le « futur du passé » : c'est un récit rétrospectif relatant une décision déjà prise, dont le conditionnel souligne le caractère à la fois décidé et non encore réalisé. Ainsi, dans le SDAURP de 1965, l'avenir se donne moins comme un champ ouvert de possibles à explorer que comme la conséquence différée d'orientations déjà arrêtées au niveau politique. Le temps ramifié y existe, mais le point de bifurcation est passé, la branche choisie est fixée et le procès lui-même appartient encore au futur. On pourrait parler d'un temps ramifié à ramification rétrospective, où l'ouverture du possible a déjà été résolue par la décision politique, et dont le conditionnel garde la trace énonciative.

L'imparfait du SDAU 1976 opère à l'inverse une distanciation affirmée par rapport aux récits hérités. Dans ce schéma directeur, son emploi – « l'explosion urbaine *imposait* un nouvel état d'esprit », « les villes nouvelles *apparaissaient* à la fois comme... » – n'est pas celui d'un passé de narration, mais d'un passé critique : il sert à mettre à distance les représentations et les pratiques antérieures, afin d'ouvrir l'espace discursif à une prospective renouvelée. On retrouve ici la logique du temps ramifié, mais dans une configuration particulière : le passé est mis en débat, ses certitudes sont contestées, et c'est cette mise en question rétrospective qui fonde la légitimité d'une bifurcation vers un avenir inédit.

Enfin, les modes impersonnels qui dominent les deux derniers groupes de textes, et en particulier le SDRIF-E de 2024 semblent brouiller cette grille de lecture entre temps linéaire et temps ramifié. L'infinitif, dépourvu de marque temporelle et personnelle, ne permet plus de repérer un point d'énonciation à partir duquel s'organise la représentation du temps. Comme le dit Gosselin « l'infinitif est aoristique par défaut » (2010). En outre, on observe que ces séries d'infinitifs sont aussi parfois subordonnées syntaxiquement à un présent de vérité générale : « La Nature en ville *est* un levier puissant pour *lutter* contre les canicules, mais aussi *réduire* l'imperméabilisation des sols et donc *réduire* les phénomènes d'inondation ». Dans ces énoncés, il ne s'agit pas de décrire une ligne temporelle ni d'explorer un arbre de possibles, mais d'envisager l'action par rapport à son point d'aboutissement, en considérant le procès dans son intégralité. Ce basculement vers l'infinitif semble traduire une évolution profonde du discours en planification et du rapport que celle-ci entretient avec l'avenir : elle n'est plus un programme d'actions à venir situées par rapport à un horizon précis, mais un ensemble de principes généraux à valeur quasi législative, dont la temporalité est celle de la règle plutôt que du projet.

4. Conclusion

L'analyse de la temporalité verbale dans les textes de planification régionale révèle, sous un jour nouveau, les relations au temps qui caractérisent chaque période de planification. En montrant que l'on ne parle pas du futur de la ville au futur simple, notre analyse met au jour une pluralité de régimes d'énonciation de l'avenir, chacun révélant des relations spécifiques

entre moment d'énonciation, moment de l'événement et horizon de référence. L'intérêt de cette approche est de revisiter l'histoire urbaine en montrant comment se construisent et se reconfigurent, d'un document à l'autre, les façons d'anticiper, de décider et de justifier l'avenir métropolitain.

Cette analyse conforte ainsi le bien-fondé d'une étude plus approfondie des modalités des temps verbaux pour étudier les représentations du temps en planification urbaine. L'annotation manuelle ou semi-automatique du corpus ne constitue cependant pas la méthode la plus adaptée, car les marqueurs modaux mobilisent avant tout des « éléments linguistiques divers, qui ne sont pas eux-mêmes spécifiquement modaux mais qui ont un aspect modal » (Gosselin, 2010), et leur interprétation exige de s'attacher avant tout au sens de chaque énoncé. TXM constitue néanmoins un appui indispensable pour repérer certaines régularités morphosyntaxiques que la diversité des formes d'énonciation n'invalide pas entièrement.

Références

- Ameel L., Martin Gurr J. and Buchenau B. (2023). *Narrative in Urban Planning: A Practical Field Guide*, Transcript Verlag.
- Apothéloz D. (2021). Temps verbaux : (II) les temps composés. In *Encyclopédie Grammaticale du Français*, encyclogram.fr.
- Apothéloz D. (2023). Le futur simple. In *Encyclopédie Grammaticale du Français*, encyclogram.fr.
- Ascher F. et al. (1993). *Les territoires du futur*. Éditions de l'aube.
- Bres J. (2024). Le conditionnel. In *Encyclopédie Grammaticale du Français*, encyclogram.fr.
- Buhler T., Chesneau I. and Richeton A. (2024). Since when has regional planning been vague? In AESOP Annual congress, SciencePo, *Book of proceedings*, hal.science/hal-04827225.
- Chesneau I. (2026). Planifier la ville, écrire le changement. In Eshkol-Taravella, I. and Ratouis, O. editors, *Déchiffrer et interpréter les discours urbains. Nouvelles méthodes pour l'analyse des textes d'urbanisme*. Presses universitaires de Nanterre.
- Dixon T. J. and Tewdwr-Jones M. (2021). *Urban Futures : Planning for City Foresight and City Visions*. Policy Press.
- Gosselin L. (1996). *Sémantique de la temporalité en français. Un modèle calculatoire et cognitif du temps et de l'aspect*. Duculot, Louvain-la-Neuve.
- Gosselin L. (2010). *Les modalités en français. La validation des représentations*, Amsterdam, Netherlands ; New York, Rodopi.
- Gosselin L. (2020). Périphrases aspectuelles. In *Encyclopédie Grammaticale du Français*, encyclogram.fr.
- Krazem M. (2007). Infinitif et nominalisation : une seule ou deux catégories ? *L'Information Grammaticale*, vol. 114(1) : 46-52.
- Martin R. (1981). Le futur linguistique : temps linéaire ou temps ramifié ? *Langages*, vol. 15(64) : 81-92.
- Perrin L. (2004). Polyphonie et autres formes d'hétérogénéité énonciative : Bakhtine, Bally, Ducrot, etc. *Pratiques*, vol. 123(1) : 7-26.
- Pincemin B. (2012). Hétérogénéité des corpus et textométrie. *Langages*, vol. 187(3) : 13-26.
- Rastier F. (1996). La sémantique des thèmes ou le voyage sentimental. *Revue Texto*, revue-texto.net/1996-2007/Inedits/Rastier/Rastier_Themes.html.
- Rebotier A. (2015). Le futur périphrastique français avec aller : un renvoi spécifique à l'avenir ou un temps en voie de grammaticalisation ? Une approche contrastive. *Revue de sémantique et pragmatique*, vol. 38(38), journals.openedition.org/rsp/502.
- Tulumello S. (2025). For a dialectic of planning pasts and futures: Theoretical courses and recourses in conversation with Patsy Healey. *plaNext. Next Generation Planning*, vol. 15.

Investigation de l'évolution des discours médiatiques à propos des réseaux socionumériques à partir de la distance de Labbé et de sa version améliorée*

Pierre Ratinaud¹, Brigitte Sebbah²

¹LERASS, Université de Toulouse 2 Jean Jaurès – pierre.ratinaud@univ-tlse2.fr

²LERASS, Université de Toulouse – brigitte.sebbah@iut-tlse3.fr

Abstract

In this study, we examine the relevance of using Labbé's distance (Labbé and Labbé, 2003) in both its original and "improved" versions (Cortelazzo et al., 2013) to contribute to the study of the evolution of media discourse regarding social networking sites (SNS). The corpora analyzed consist of press articles : the first includes mentions of at least one of the ten most popular SNS (n=109,262), while the second focuses on the terms "social networks" or "social network" without referencing the aforementioned platforms (n=48,528). These were extracted from seven major French national dailies and cover the entire lifespan of SNS, from 2003 to 2024. Based on these corpora, we conduct an experiment using the improved version of Labbé's distance to, on the one hand, replicate the results proposed by Cortelazzo et al. (2013) and, on the other, explore the use of inter-annual distance evolution as an indicator of the dynamics of social representations (Moliner, 2001).

Keywords : Labbé's distance, improved Labbé's distance, social network, social representations, newspapers

Résumé

Cet article examine la pertinence de l'utilisation de la distance de Labbé (Labbé and Labbé, 2003) dans sa version originale et dans sa version « améliorée » (Cortelazzo et al., 2013) pour participer à l'étude de l'évolution des discours médiatiques sur les réseaux socionumériques (RSN). Les corpus sur lesquels nous travaillons sont constitués d'articles de presse mentionnant au moins l'un des 10 RSN parmi les plus populaires (n=109262) pour le premier, et les notions de « réseaux sociaux » ou de « réseau social » sans faire référence aux RSN précédents (n=48528) pour le second. Ils sont extraits de 7 grands quotidiens nationaux français et couvrent l'entièreté de la période d'existence des RSN, de 2003 à 2024. A partir de ces corpus, nous menons une expérimentation en utilisant la version améliorée de la distance de Labbé pour, d'une part, répliquer les résultats proposés par Cortelazzo et al. (2013) et, d'autre part, envisager d'utiliser l'évolution des distances inter-années comme un indicateur de la dynamique des représentations sociales (Moliner, 2001).

Mots clés : distance de Labbé, distance de Labbé améliorée, réseaux sociaux, représentations sociales, presse quotidienne

1. Introduction

Nous avons probablement tous du mal à nous en souvenir, mais avant le début du 21^{ème} siècle, des marques comme Facebook, YouTube ou Twitter nous étaient parfaitement inconnues. Et

*Ce travail a été réalisé dans le cadre du Labex SMS portant la référence ANR-11-LABX-0066

pour cause, ni elles ni les référents lexicaux utilisés pour les désigner n'existaient. La notion de réseau social était à cette époque circonscrite à une branche de la sociologie qui s'intéresse aux réseaux relationnels des individus, et c'est cette signification dont nous retrouvons les traces dans la presse avant l'avènement des RSN. En l'espace d'un quart de siècle, non seulement le sens associé à l'expression "réseau social" va radicalement changer, mais ce changement va s'accompagner de l'entrée dans notre champ de connaissance de dizaines d'exemples de ces outils qui vont prendre une place prépondérante dans les vies quotidiennes des individus et plus largement dans la régulation de l'espace public, de la politique et du débat démocratique en général. Dans l'objectif d'étudier ce phénomène de naturalisation sur le temps long, nous avons, à l'image de l'étude princeps de Moscovici (1976) sur la psychanalyse, entrepris d'étudier les discours médiatiques de la presse quotidienne nationale (PQN) française qui ont accompagné et nourri la genèse des représentations sociales des réseaux socionumériques (RSN) en général et de leurs principaux représentants en particulier. Après avoir présenté les corpus que nous avons constitués dans cet objectif, nous rappellerons les propositions de Moliner (2001) sur l'évolution des discours sur les objets de représentation. Il décrit trois phases dans lesquelles peut se situer une représentation sociale (émergence, stabilité, transformation) et les caractéristiques des discours associés à ces phases. Dans ce travail, nous envisagerons l'évolution des distances inter-années de nos corpus comme un indicateur de la situation des représentations dans ces phases. Dans ce cadre, une baisse des distances lexicales peut être interprétée comme une phase de stabilisation des représentations et une augmentation comme le signe d'une diversification des discours, c'est à dire comme une phase d'émergence ou de transformation. Nous avons dans un premier temps envisagé d'étudier l'évolution des distances entre chacune des années d'indexation de nos corpus à partir des distances de Labbé (Labbé and Labbé, 2003). Les résultats montrent une diminution des distances inter-années de tous les corpus au fur et à mesure de l'avancée dans le temps. Mais les travaux de Cortelazzo et al. (2013) nous invitent à relativiser ce premier résultat. Ils soulignent en effet que la distance de Labbé est dépendante de la taille des corpus à comparer dans le sens où plus les corpus gagnent en importance, plus les distances entre deux corpus ont tendance à diminuer. Or, les corpus que nous manipulons ont presque tous tendance à augmenter de taille dans le temps (chaque année il y a plus d'articles mentionnant les RSN en général ou en particulier). Nous suivrons donc les recommandations de ces auteurs en changeant le mode de calcul des distances inter-années pour une approche par "morceaux". Nous décrirons donc cette approche, puis nous soumettrons nos corpus à une expérimentation qui consiste à faire varier la taille des morceaux utilisés, de façon à répliquer les résultats de Cortelazzo et al. (2013).

2. Description des corpus : quels réseaux dans quels journaux ?

Les RSN font partie des applications numériques les plus utilisées partout sur la planète. Nous avons sélectionné pour cette étude 10 plateformes : 9 d'entre elles sont dans la plupart des classements des RSN comptant le plus d'utilisateurs (Facebook, YouTube, Instagram, Twitter, Snapchat, LinkedIn, TikTok, Reddit et Pinterest). Nous avons ajouté à cette liste MySpace, un des précurseurs des RSN qui a connu un franc succès à ses débuts avant de périr sous l'effet de la vague Facebook. Concernant les journaux, nous avons retenu 7 des 8 représentants de la PQN les plus vendus en France : Le Monde, Le Figaro, Libération, L'Humanité, Aujourd'hui en France, La Croix et les Échos. Seul le journal sportif L'Equipe, qui complète cette liste, n'a pas été pris en compte. Nous avons retenu tous les articles qui mentionnent au moins une fois l'une des 10 RSN sur l'intégralité de leur période d'existence, de janvier 2003 à décembre 2024. Un onzième corpus, sur la même période, comprend les articles qui mentionnent les expressions "réseau social" ou "réseaux sociaux" sans faire intervenir les 10 RSN sélectionnés. Nous le

nommerons corpus "générique". Le tableau 1 présente les effectifs d'articles pour chacun des corpus :

	Fréquence des articles	Occurrences
Facebook	49528	35232811
Twitter	42854	30196244
Youtube	18934	14277738
Instagram	14208	12410920
Tiktok	4799	4248550
Snapchat	2990	2497282
Linkedin	2850	2112137
MySpace	1832	1085925
Reddit	673	652208
Pinterest	643	428376
Générique	48528	38837656

TABLE 1: Fréquence des articles et des occurrences dans chacun des corpus

Étant donné le peu d'articles qui abordent Reddit et Pinterest et la particularité de la distribution de MySpace, qui connaît son pic de mention en 2007 puis disparaît presque complètement, nous ne les retiendrons pas pour cette analyse¹.

L'étude de l'évolution chronologique de la fréquence des articles dans ces corpus n'est pas une tâche aisée. Il faut en effet prendre en compte la baisse continue du nombre d'articles publiés par les journaux sélectionnés entre 2003 et 2024. Le graphique 1 présente donc ces évolutions en proportion du nombre total d'articles publiés dans ces journaux :

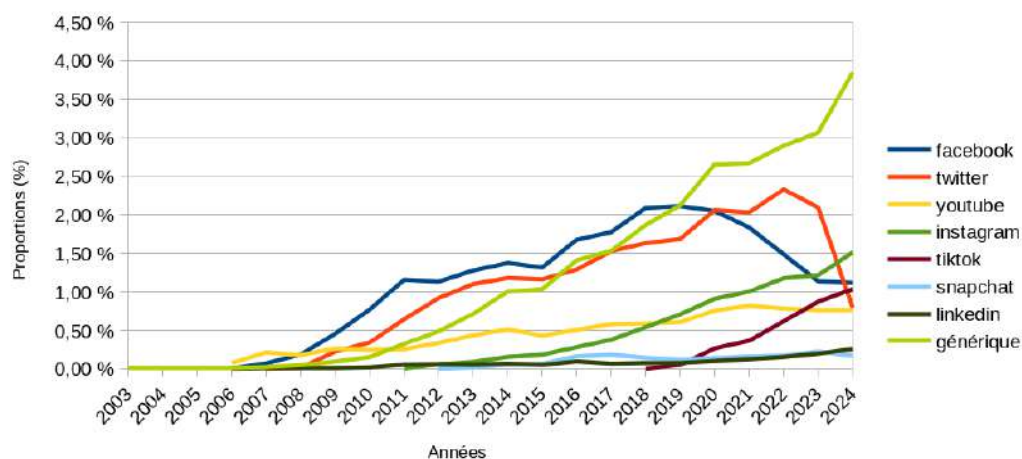


FIGURE 1: Évolution des fréquences d'articles dans les 8 corpus sélectionnés en proportion du nombre total d'articles publiés par année (2003-2024)

¹ Une analyse complète de ces corpus a été publiée dans un rapport de l'Observatoire des Pratiques Socio-Numérique (OPSN). Elle est disponible sur le site de l'OPSN : <https://opsn.univ-tlse3.fr/rapport-aux-origines-dun-monde>

A deux exceptions près, toutes ces courbes montrent une tendance à l'augmentation du nombre d'articles mentionnant les RSN. Cette augmentation est très forte pour le corpus générique, Instagram et TikTok. Elle est plus mesurée pour YouTube, Snapchat et LinkedIn. La forte baisse constatée pour Twitter est probablement en grande partie imputable au changement de nom de la plateforme (devenue X) qui intervient en 2023 suite à son rachat par Elon Musk en 2022. La baisse des mentions de Facebook, qui commence en 2018, semble plus structurelle. On peut penser qu'une partie peut être la conséquence du changement de nom de la maison mère de Facebook à Meta. Mais, d'une part, ce changement n'interviendra qu'en 2021, et d'autre part, une analyse complémentaire montre que l'intégration des articles mentionnant Meta ne compense pas la baisse constatée. Nous pensons qu'il s'agit plutôt de la conséquence du scandale de "Cambridge Analytica" (2018), des Facebook Files (2021) et d'un changement d'algorithme (2018) qui tend à pénaliser les contenus des médias traditionnels. Tous ces événements ont probablement participé à réduire l'intérêt des journalistes pour Facebook, qui, consciemment ou inconsciemment, le mentionnent moins dans les articles.

3. Mesurer l'évolution lexicale : représentations sociales et distance de Labbé

3.1. Dynamique des représentations sociales

Pendant les 22 ans de cette histoire, les journalistes vont donc de plus en plus participer à faire circuler des informations sur les réseaux sociaux en général et sur chacun de leur représentants. Ces informations vont participer à la structuration des représentations sociales de ces objets (Moscovici, 1976). Deux processus sont mobilisés dans cette phase :

- L'objectivation : le processus d'objectivation va permettre la sélection des informations à propos de l'objet, de façon à passer d'un flot informationnel complexe et abondant à une image simplifiée mais opérante.
- L'ancrage : le nouvel objet va être incorporé dans un ensemble pré-existant de connaissances et de valeurs, un système de représentations sociales, dont les orientations dépendantes des groupes sociaux concernés. Le nouvel objet doit trouver une place dans ce que Jodelet (1989) nomme "un déjà là pensé".

Dans un chapitre d'ouvrage consacré à la dynamique des représentations sociales, Moliner (2001) décrit trois phases dans lesquelles se présentent les représentations sociales. Il distingue une phase d'émergence, par laquelle toutes les représentations passent, une phase de stabilité, quand l'objet est complètement naturalisé dans les groupes sociaux concernés et qu'il a acquis le statut d'évidence de tous les objets qui nous sont familiers, et une éventuelle phase de transformation, qui intervient quand l'objet ou le groupe porteur de la représentation subit des transformations. Ces trois phases sont associées à des propriétés particulières des représentations ou des groupes sociaux. Parmi celles-ci, Moliner propose des indicateurs en lien avec la variabilité des discours concernant l'objet. Il décrit une phase d'émergence durant laquelle les communications collectives sont intenses et les discours diversifiés, une phase de stabilité avec moins de communications collectives et des discours plus consensuels et enfin une phase de transformation, qui intervient quand le groupe ou l'objet subit des modifications, associée de nouveau à une intensification et une diversification des discours. Nous partons donc du principe que le cheminement décrit par Moliner (2001) devrait laisser des traces dans les corpus de presse sous la forme d'une diminution de la distance lexicale entre les années d'indexation au fil du temps quand les représentations se stabilisent et une augmentation des distances quand elles apparaissent ou se transforment. L'idée que nous défendons est que si les discours perdent

en variabilité avec la naturalisation des objets, alors les distances lexicales entre deux années consécutives doivent diminuer avec le temps parce que les domaines d'ancrage et les éléments issus du processus d'objectivation tendent à se stabiliser. À l'inverse, l'incorporation ou la mise en rapport des objets avec de nouveaux domaines d'ancrage ou la reprise du processus d'objectivation doivent amener une plus grande variabilité dans les discours et se traduire par une augmentation des distances inter-années.

3.2. Evaluer les distances intertextuelles : la distance de Labbé

La distance intertextuelle de Labbé (Labbé and Labbé, 2003; Labbé and Monière, 2000) permet d'estimer les distances entre deux textes ou entre deux corpus sur la base de la fréquence de leur lexique. Nous résumons ici le calcul tel qu'il est présenté dans Labbé and Labbé (2003) et implémenté dans le logiciel IRaMuTeQ². La distance entre les textes A et B, notée $D(a,b)$ est obtenue par le calcul suivant :

$$D_{(a,b)} = \frac{\sum_{v \in A, v \in B} |F_{ia} - E_{ia(u)}|}{N_a + N'_b}$$

Nous supposons ici que le texte B a une taille supérieure au texte A. Dans ce calcul, les fréquences du lexique du texte B sont ramenées à ce qu'elles seraient si l'étendue du texte B était équivalente à celle du texte A (N_a). La fréquence attendue d'une forme i de B dans le texte A, notée $E_{ia(u)}$ dans la formule, est alors égale à la fréquence de la forme i dans B (F_{ib}) multiplié par le ratio de taille entre A et B :

$E_{ia(u)} = F_{ib} * U_{(a,b)}$ avec $U_{(a,b)} = N_a / N_b$

On note N'_b l'étendue du texte B', qui est la réduction de la longueur de B en fonction de celle de A. L'indice apprécie alors la distance entre la fréquence des formes dans A (F_{ia}) et celle dans B'. Ce calcul est appliqué à toutes les formes des textes aux exceptions suivantes :

- On ne considère qu'une seule fois les formes communes aux deux textes, indépendamment de leur fréquence.
- On applique le calcul à l'ensemble des formes de A, mais seules les formes de B ayant une fréquence attendue dans A supérieure ou égale à 1 seront considérées.

La distance obtenue est un score qui varie de 0 à 1 : 0 signale deux textes parfaitement identiques du point de vue de la fréquence de leur lexique et 1 deux textes qui n'ont aucun mot en commun.

3.2.1. Application de la distance de Labbé pour l'étude de l'évolution des distances inter-années

Nous avons donc appliqué le calcul de la distance de Labbé aux 8 corpus, en nous centrant sur les années successives et en limitant l'analyse aux lemmes des formes pleines de fréquence supérieure ou égale à 2. Cette analyse est réalisée avec l'implémentation de la distance de Labbé proposée dans IRaMuTeQ (Ratinaud and Marchand, 2016). Les résultats sont présentés dans le graphique 2 :

² Le code R utilisé pour tous les calculs de distance dans ce travail est disponible ici : https://gitlab.huma-num.fr/pratinaud/iramuteq/-/blob/3.0/Rscripts/distance-labbe.R?ref_type=heads

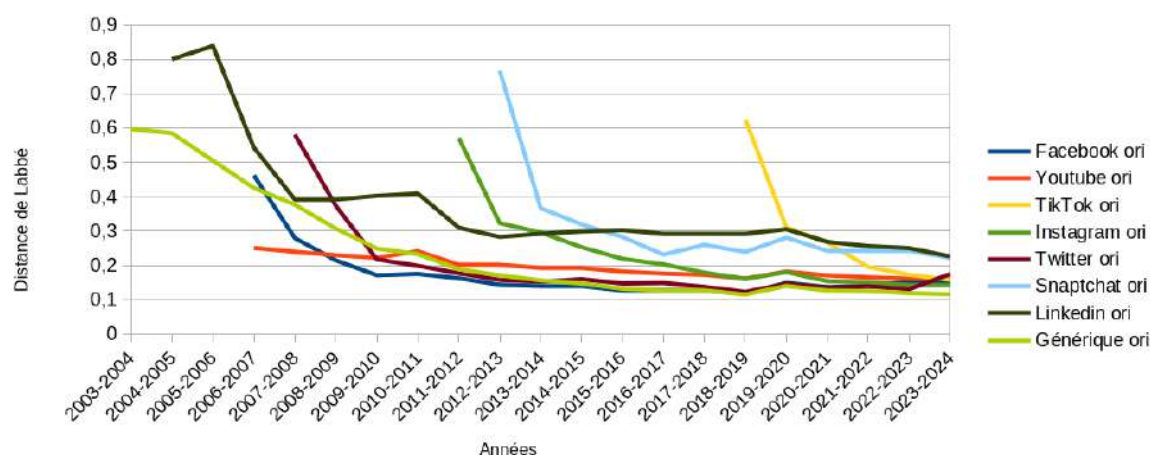


FIGURE 2: Évolution des distances inter-années (distance de Labbé) dans les 8 corpus sélectionnés

Trois phénomènes sont observables sur ce graphique :

- La plupart des corpus présente une première distance beaucoup plus importante que les suivantes ; ce phénomène est principalement la conséquence d'un tout petit nombre d'articles la première année d'indexation.
- Toutes les courbes présentent une tendance à la baisse : les distances inter-année diminuent au fur et à mesure de l'avancée dans le temps ;
- Les corpus les plus conséquents (Facebook, Twitter et le corpus Générique) présentent les distances les plus faibles à chaque étape ; les corpus les plus petits (Snapchat, LinkedIn) les distances les plus grandes.

Ces résultats soulignent une certaine dépendance des mesures de la distance de Labbé à la taille des corpus, phénomène déjà mis en évidence dans les travaux de Cortelazzo et al. (2013). En tenant compte de l'évolution des fréquences d'articles dans le temps dans ces corpus, orientée à la hausse pour la plupart, il devient difficile sur la base de ces données de conclure sur l'augmentation ou la diminution des distances inter-années.

4. Amélioration de la mesure des distances de Labbé (Cortelazzo et al., 2013)

Dans leur travail sur l'amélioration de la détermination des distances de Labbé, Cortelazzo et al. (2013) proposent de résoudre les problèmes des différences de taille entre texte ou corpus dans le calcul de la distance en substituant l'approche classique par une approche par "morceaux" (chunks). Ils suggèrent de découper les textes à comparer en morceaux de taille constante, de calculer la distance entre tous les couples de morceaux des deux textes et de déduire la distance finale en calculant la moyenne des distances des couples de morceaux. Ils ont appliqué cette procédure à un corpus de 160 romans italiens en faisant varier la taille des morceaux utilisés. La figure 3 présente un de leurs résultats :

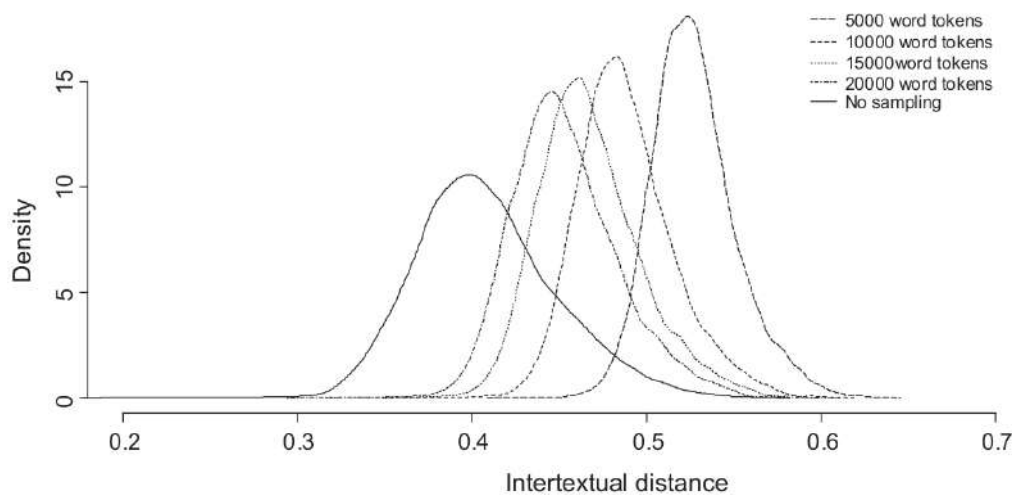


FIGURE 3: Distribution des distances intertextuelles obtenues avec et sans découpage. Calcul réalisé sur les formes. D'après Cortelazzo et al. (2013)

Ce premier résultat montre clairement que la taille des morceaux influence la distance dans le sens où plus les morceaux sont grands, plus la distance diminue. Il montre par ailleurs que la variabilité des distances est moindre dans les procédures par morceaux que dans celle qui utilise les textes comme unité. Des résultats équivalents sont obtenus en prenant les lemmes à la place des formes, avec le constat d'une légère baisse des distances moyennes. Cette procédure présente deux avantages pour notre analyse : d'une part, elle permet d'effacer l'effet généré par l'augmentation continue de la taille des ensembles comparés et d'autre part, elle permet la comparaison des valeurs des distances obtenues sur des corpus différents. Nous avons donc appliqué cette démarche sur nos corpus, avec quelques légères différences.

5. Application de l'approche par morceaux sur nos corpus

Dans leur expérimentation, Cortelazzo et al. (2013) comparent des romans qu'ils découpent en morceaux de 5000, 10000, 15000 ou 20000 occurrences. Pour nous approcher le plus possible d'un échantillonnage construit sur un nombre d'occurrences sans avoir à casser l'unité de notre analyse (l'article), nous avons constitué des ensembles qui accumulent autant d'articles que nécessaire pour atteindre 15000 occurrences pour la première mesure et 30000 occurrences pour la seconde³. La taille moyenne des articles de presse que nous manipulons étant de 788 occurrences, la variation de taille de nos morceaux est minime, mais non nul. Nous utiliserons donc la même procédure de calcul que précédemment. Les tailles des morceaux ont aussi été choisies sur un critère de faisabilité. La procédure nécessite en effet un grand nombre de calculs. Par exemple, la détermination de la distance de l'année 2018 à l'année 2019 dans le corpus de Facebook nécessite de calculer 8010 distances entre groupes de 15000 occurrences. Nous avons résumé les résultats obtenus dans la figure 4 :

³ Nous avons mené une première expérience en utilisant des morceaux basés sur un nombre d'articles (des groupes de 50 vs des groupes de 100 articles) : cette procédure s'est révélée inadéquate. Il semble en effet que la taille des articles augmente avec le temps, ce qui conduit à répliquer l'artefact de taille que nous cherchons à contrôler.

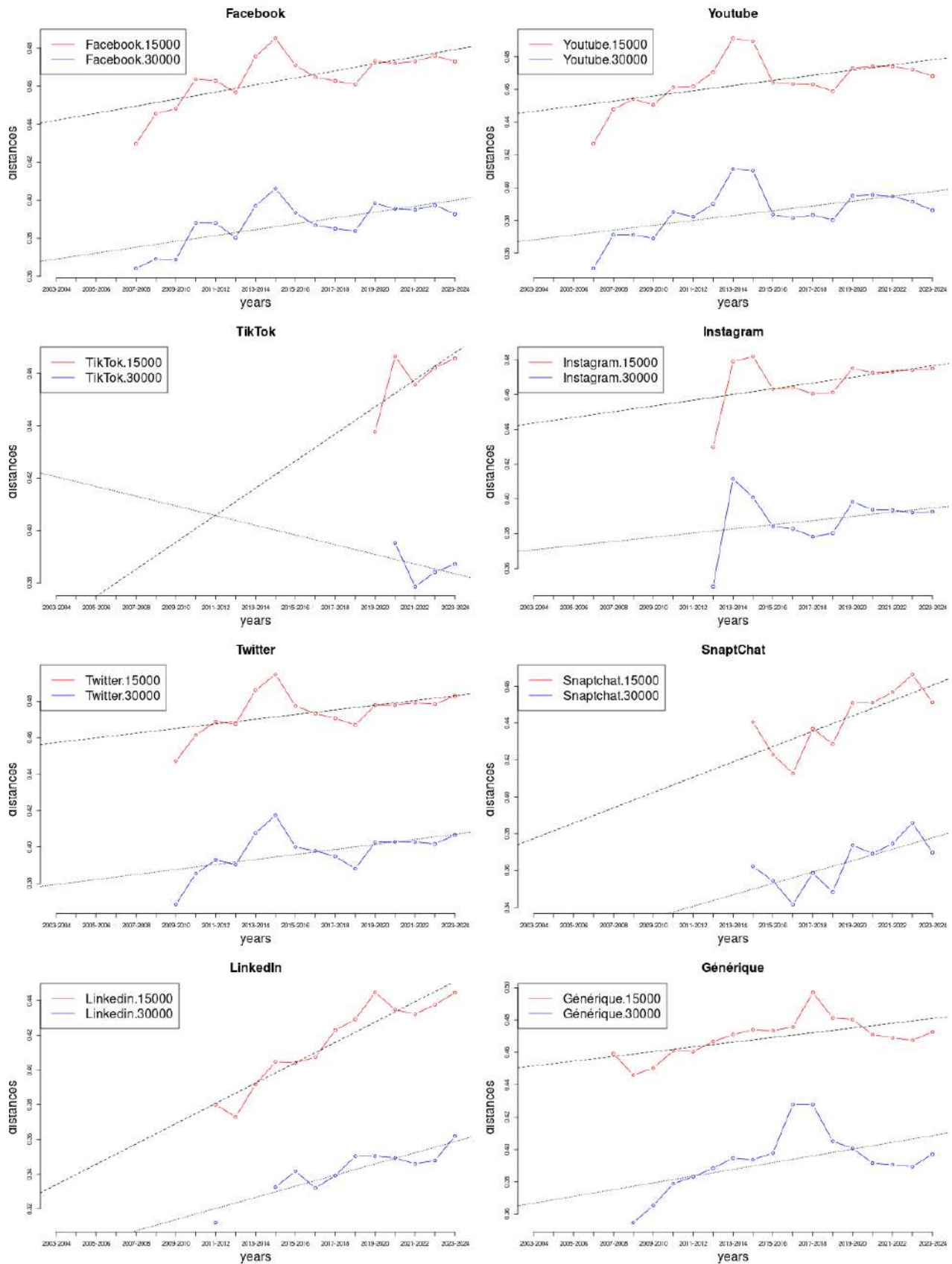


FIGURE 4: Distribution des distances inter-années sur chacun des 8 corpus avec la méthode par morceaux. Les lignes noires correspondent aux droites de régression de chacune des distributions

Ces résultats permettent plusieurs commentaires. Le premier consiste à constater que nous reproduisons ici les résultats proposés par Cortelazzo et al. (2013) : d'une part la taille des échantillons choisie a bien une incidence sur les distances calculées dans le sens d'une diminution des distances quand la taille des échantillons augmente. D'autre part, les ordres de grandeur des distances obtenues pour différents les corpus sont tout à fait comparables, contrairement à l'analyse avec l'approche classique. La démarche par morceaux annulant l'effet des variations de taille des ensembles comparés, on observe une tendance générale à la hausse, avec des inflexions selon les corpus, résultat inverse de celui obtenu avec l'approche classique. Nous constatons enfin que les distributions issues des deux tailles d'échantillon suivent des évolutions parfaitement synchronisées. Cette dimension est fondamentale de notre point de vue, elle vient renforcer l'intérêt, pour l'analyste, d'investiguer ces changements. Sur ce point, les résultats sur le corpus générique sont particulièrement intéressants : ils présentent un pic d'augmentation de la distance entre les années 2017 et 2018 avant d'entamer une phase de baisse. Nous avons effectivement constaté dans ce corpus un changement d'orientation des discours autour de l'année 2018 avec une diminution des thématiques liées à l'économie et à l'innovation par exemple, et une augmentation des thématiques sur la fiscalité des plateformes et sur leur régulation (Ratinaud et al., 2026). Il en est de même pour le pic de hausse commun aux corpus de Facebook, Twitter et YouTube entre 2014 et 2015, dont on peut faire l'hypothèse qu'il est la conséquence des attentats terroristes en France, période durant laquelle les RSN ont été soupçonnés de participer à la radicalisation religieuse d'une partie de la jeunesse.

6. Conclusion

Dans ce travail, nous avons utilisé une approche du calcul des distances de Labbé par morceaux proposée par Cortelazzo et al. (2013) pour déterminer l'évolution des distances inter-années sur des corpus massifs de presse à propos des RSN. Cette approche permet de remédier à la tendance des distances de Labbé de diminuer avec l'augmentation de la taille des corpus et facilite donc l'étude des distances entre corpus de tailles différentes. Elle permet également d'obtenir des distances dont les ordres de grandeur ne dépendent plus de la taille des corpus mais uniquement de la taille des morceaux sélectionnée, et offre donc la possibilité de comparaisons entre des distances obtenues sur des ensembles de tailles différentes.

La démarche entreprise ici permet de proposer un indicateur pour positionner les représentations sociales étudiées dans les phases énoncées par Moliner (2001). Le constat de l'augmentation des distances inter-années dans l'ensemble des corpus au fur et à mesure de l'avancée dans le temps, associé à l'augmentation des articles qui mentionnent les RSN correspond clairement aux phases d'émergence et de transformation décrites par cet auteur. Ces résultats suggèrent que les représentations sociales des RSN seraient toujours en phase de construction et de transformation. Les résultats nous semblent par ailleurs compatibles avec une lecture de l'évolution des thématiques dans ces corpus (Ratinaud et al., 2026), qui souligne la diversité des angles des articles qui mentionnent les RSN ou la notion de "réseaux sociaux" sur la période. De la description de dispositifs techniques et de ce à quoi ils pourraient servir, dans un début de siècle où les lecteurs ne disposent pas tous d'un accès à Internet, à la description du poids croissant qu'ils prennent dans la communication politique et dans une économie qui se remet de l'éclatement de la bulle internet, en passant par le rôle qu'on leur attribue dans les révolutions arabes (2011), dans la radicalisation religieuse (2014/2016) ou dans le mouvement des Gilets Jaunes (2018-2019), les RSN sont associées à une part croissante des dimensions de la vie sociale. Dans un monde post-covid, qui a acté la quasi-généralisation de leur usage, les journalistes rendent compte de leur inscription dans la vie quotidienne (cuisine, mode de vie, santé, éducation, sport, culture populaire, littérature...), et parallèlement, les mentionnent de plus en plus

dans des articles relevant des champs de la politique et des débats démocratiques. Ce sont les traces de cette variabilité renouvelée dont témoignent les analyses proposées.

Références

- Cortelazzo M. A., Nadalutti P., and Tuzzi A. (2013). Improving Labbé's Intertextual Distance : Testing a Revised Version on a Large Corpus of Italian Literature. *Journal of Quantitative Linguistics*, 20(2) :125–152.
- Jodelet D. (1989). Représentations sociales : un domaine en expansion. In Jodelet D., editor, *Les représentations sociales*, p. 424. PUF, Paris, France.
- Labbé C. and Labbé D. (2003). La distance intertextuelle. *Corpus*, (2).
- Labbé D. and Monière D. (2000). La connexion intertextuelle. Application au discours gouvernemental québécois. In Rajman M. and Chappelier J.-C., editors, *Actes Des 5èmes Journées Internationales d'Analyse Statistique Des Données Textuelles*, pp. 85–94. EPLF, Lausanne, Suisse.
- Moliner P. (2001). *La dynamique des représentations sociales : pourquoi et comment les représentations se transforment-elles ?* Vies sociales. Presses universitaires de Grenoble, Grenoble.
- Moscovici S. (1976). *La psychanalyse : son image et son public*. Bibliothèque de psychanalyse. Presses universitaires de France, Paris, 2e édition entièrement refondue edition.
- Ratinaud P. and Marchand P. (2016). Quelques méthodes pour l'étude des relations entre classifications lexicales de corpus hétérogènes : Application aux débats à l'assemblée nationale et aux sites web de partis politiques. In Mayaffre D., Poudat C., Vanni L., Magri V., and Follette P., editors, *Statistical Analysis of Textual Data*, pp. 193–202, Nice, France.
- Ratinaud P., Sebbah B., Cabanac G., Fraisier-Vanier O., and Milard B. (à paraître 2026). Aux origines d'un monde : Les réseaux socionumériques dans la presse quotidienne française (2003-2024). Rapport de recherche, Observatoire des Pratiques Socio-Numériques (OPSN), Toulouse.

Caractériser les commentaires en ligne : les *émojis* dans la loi « Duplomb »

Pascal Marchand¹, Pierre Ratinaud²

¹Université de Toulouse – pascal.marchand@univ-tlse3.fr

²Université de Toulouse – pierre.ratinaud@univ-tlse2.fr

Abstract

The “Duplomb Law” is an agricultural law debated in the French Parliament in the summer of 2025. For the first time in France, a petition on the National Assembly’s website surpassed two million signatures in three weeks. We are using textometrics to (1) attempt to define its scope and (2) raise new questions about the methodology for analyzing online comments. From a theoretical perspective, the results confirm Kahneman’s “*System 1 and System 2 thinking*” model. From a methodological standpoint, the question is whether it is appropriate to include emojis and, if so, how. Issues such as tokenization (indexing) and tagging (object or emotion semantics), as well as statistical status (active or supplementary), are discussed. Comparative analyses are presented, which reveal no differences in usage between denotative and connotative *emojis*.

Keywords: Lexical classification; emojis; social cognition; *Iramuteq*

Résumé

La « Loi Duplomb » est une loi agricole débattue au Parlement français à l’été 2025. Pour la première fois en France, une pétition sur le site de l’Assemblée nationale a dépassé les deux millions de signatures en trois semaines. Nous mobilisons la textométrie pour (1) tâcher d’en cerner les contours et (2) introduire de nouvelles questions sur la méthodologie d’analyse de commentaires en ligne. Du point de vue théorique, les résultats confirment le modèles des « deux systèmes de pensée » de Kahneman (2011). Du point de vue méthodologique, on se demande s’il est pertinent de prendre en compte les *émojis* et comment. On évoque les questions de tokenisation (indexation) et étiquetage (sémantique d’objet ou d’émotion), statut statistique (actif ou supplémentaire) ? Des analyses comparatives sont présentées, qui ne révèlent pas d’usages différentiels entre les *émojis* dénotatifs et connotatifs.

Mots clés : Classification lexicale ; *émojis* ; cognition sociale ; *Iramuteq*.

Cette publication a bénéficié d'une aide du Labex « Structuration des Mondes sociaux » (ANR-11-LABX-0066).

1. Introduction

Apparus au Japon en 1997, les *émojis* ont connu une évolution fulgurante. En 2015, Facebook en révélait cinq milliards par jour. En 2015, on estimait que 4 % des tweets contenaient des *émojis* (Kralj Novak, 2015). On considère aujourd’hui que, sur des plateformes telles que *Twitter*, *Facebook*, *Instagram*..., les *émojis* peuvent représenter près de la moitié du contenu.

Les *émojis* sont souvent associés à des formes d’expression sémantique émotionnelle et abordés sous l’angle de l’analyse des sentiments pour les classer selon leur valence positive, négative ou neutre (Tang & al., 2024). Robertson et al. (2021) ont montré que les *émojis* étaient relativement stables et que seul un petit nombre semble subir des changements sémantiques significatifs. Plus précisément, moins un *emoji* est abstrait, plus il est susceptible de subir des changements sémantiques importants (voir aussi McGillivray, 2024).

La prise en compte des *émojis* dans l'analyse des données textuelles massives (textométrie) pose les questions de leur tokenisation (reconnaissance des unicodes simples et complexes), de leur lemmatisation (lorsque plusieurs formes similaires ont des codes différents), de leur indexation morphosyntaxique (nom, adjectif, onomatopée...) et de leur statut statistique (actif / supplémentaire).

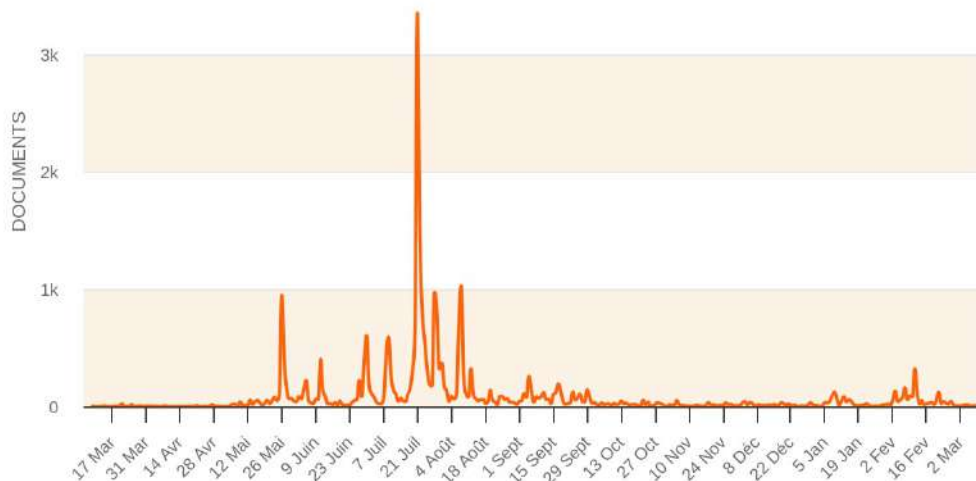
2. Corpus : les commentaires sur la « loi Duplomb »

La loi dite « Duplomb » est une loi agricole débattue durant l'été 2025. Après avoir été examinée par plusieurs commissions (affaires économiques, mixte paritaire), elle a été adoptée par le Sénat le 2 juillet 2025, par l'Assemblée nationale le 8 juillet 2025, et a été promulguée le 11 août 2025.

Elle contient plusieurs articles sur l'évolution d'institutions (*Agence nationale de sécurité sanitaire ; Office français de la biodiversité*), la requalification juridique des réserves d'eau (« mégabassines »), les relèvements des seuils d'élevages... Mais le volet le plus controversé concernait la réintroduction dérogatoire de pesticides et notamment les néonicotinoïdes comme l'acétamipride, interdit depuis 2018.

La mobilisation de l'opinion publique a été exceptionnelle : une pétition déposée sur le site de l'Assemblée nationale, par une étudiante en santé environnementale sans affiliation politique, a recueilli 500 000 signatures en neuf jours et a dépassé les deux millions en trois semaines. Ce résultat implique, pour la première fois, en France, que la loi soit rediscutée à l'Assemblée nationale, mais sans examen sur le fond, ni vote, ni possibilité d'abrogation. Initialement prévue le 7 janvier, cette discussion parlementaire a été reportée en raison d'un déplacement de la ministre de l'Agriculture à Bruxelles. Elle a finalement eu lieu le 11 février 2026.

Le graphe suivant montre l'évolution de la « loi Duplomb » dans la presse nationale française.



Graphique 1 : évolution de la fréquence des articles mentionnant la « loi Duplomb » dans la presse française (source : *Europresse*).

Nous nous intéressons ici à une expression particulière de l'opinion publique, telle qu'elle s'exprime dans les commentaires postés sur les principaux sites d'information en ligne du 12 mai 2025 au 11 mars 2026. Plus précisément, il s'agissait de repérer les chaînes d'information générale ayant les vidéos les plus vues avec la requête « loi Duplomb ». Les données ont été récupérées à partir de l'API *Youtube*¹, en effectuant une requête "loi Duplomb" pour les 17

¹ Google. (s.d.). YouTube Data API (v3). Google for Developers. <https://developers.google.com/youtube/v3>

chaînes remontant les vidéos les plus populaires. Une fois les vidéos identifiées, nous avons pour chacune d'entre elles récolté les commentaires retournés par la même API en annotant la date de création du commentaire et la chaîne *youtube* autrice de la vidéo commentée (*chanel).

	Effectifs	pourcentage
*chanel_20_Minutes_France	1611	4.26 %
*chanel_28_minutes_-_ARTE	1898	5.02 %
*chanel_BFMTV	5344	14.13 %
*chanel_BLAST_Le_souffle_de_l'info	3128	8.27 %
*chanel_C_dans_l'air_-_France_Télévisions	751	1.99 %
*chanel_Europe_1	5844	15.45 %
*chanel_France_3_Grand_Est	1112	2.94 %
*chanel_franceinfo	4637	12.26 %
*chanel_LCI	266	0.7 %
*chanel_Le_Figaro	184	0.49 %
*chanel_Le_Parisien	1698	4.49 %
*chanel_Le_Point	1278	3.38 %
*chanel_LeHuffPost	1411	3.73 %
*chanel_RMC	2188	5.78 %
*chanel_RTL__	719	1.9 %
*chanel_Sud_Radio	990	2.62 %
*chanel_TF1_INFO	4768	12.6 %
	37827	100 %

Tableau 1 : Effectifs de commentaires sur chacune des chaînes

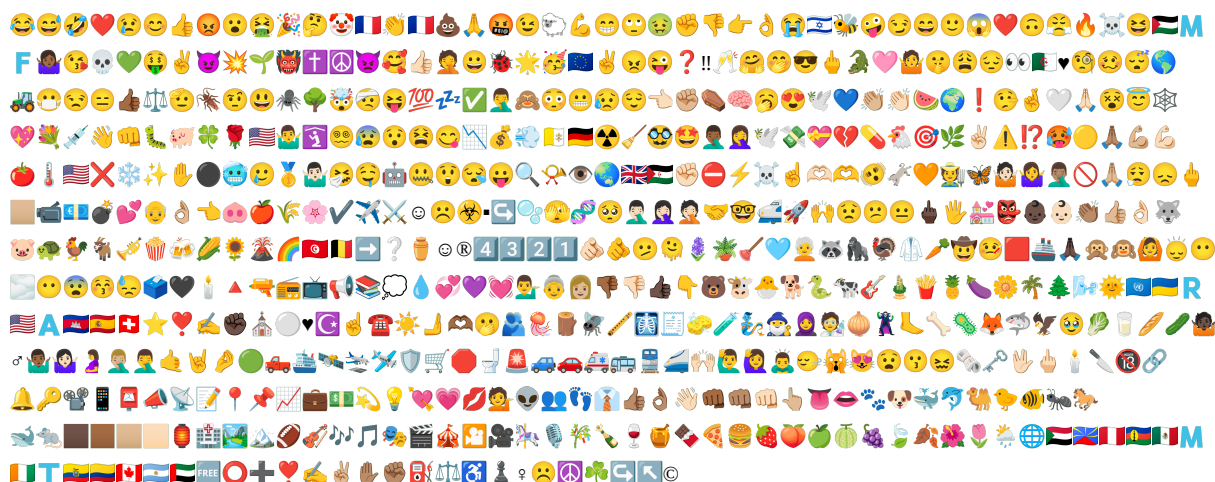
3. Formatage

3.1. Tokenisation

Le corpus textuel est formaté selon les consignes usuelles pour le logiciel libre *Iramuteq* (Pierre Ratinaud, Lerass, Université de Toulouse)². Rappelons que la tokenisation par *Iramuteq* définit une liste de caractères alphanumériques et accentués à conserver pour l'analyse. Tout caractère absent de cette liste est donc rejeté. Ainsi, les *émojis* sont habituellement rejetés comme n'appartenant pas aux caractères conservés. La tokenisation des *émojis* révèle que le code de certains n'est lié qu'à un seul caractère, mais que pour d'autres le code est lié à une composition de plusieurs caractères³. Un inventaire des *émojis* du corpus est donc effectué et donne la liste suivante (fréquences décroissantes) :

² <https://pratinaud.gitpages.huma-num.fr/iramuteq-website/category/documentation.html>

³ <https://unicode.org/emoji/charts/full-emoji-list.html>



Ces caractères sont ajoutés à la liste des caractères conservés pour l'analyse.

3.2. Lemmatisation

On observe que certains *emojis* peuvent apparaître plusieurs fois dans la liste, car ils ont des codes différents. Dans l'analyse présente, nous ne les avons pas réduits, mais cela pourrait faire l'objet d'un travail ultérieur.

3.3. Statuts

L'indexation des *emojis* pose la question de leurs statuts syntaxique et statistique.

Du point de vue syntaxique, ils sont souvent considérés comme des marqueurs connotatifs, particulièrement émotionnels. Mais certains semblent plus dénotatifs, renvoyant à des objets et pouvant même se substituer à des formes plus référentielles (des drapeaux ou animaux, par exemple). En première analyse, nous avons réservé, dans le dictionnaire d'analyse d'*Iramuteq*, une catégorie spécifique pour tous les *emojis*. Mais des études à venir devront éventuellement établir des différences.

Du point de vue statistique, la catégorie des *emojis* peut être considérée comme donnée textuelle *active* ou *supplémentaire*. Le propos sera ici de tester ces deux choix d'analyse.

4. Analyses et résultats

L'analyse repose sur une classification hiérarchique descendante (CHD) selon la méthode Reinert (1983), implémentée dans le logiciel *Iramuteq* (Ratinaud, 2018), qui permet de décrire la structure du vocabulaire. Les 37 827 commentaires (1 289 509 occurrences) représentant 30 093 lemmes dont 29 095 actifs) ont été découpés en 52 487 segments de texte (de 24,6 formes en moyenne). Sur la base d'une matrice binaire codant la présence (1) ou l'absence (0) d'une forme lexicale (lemmes en colonnes) dans un segment de texte (en lignes), on regroupe les segments dont les profils sont semblables. 47 941 segments peuvent ainsi être classés (91.34%) dans une arborisation hiérarchique (*dendrogramme*).

Comme le montre le tableau 2, après lemmatisation, reconnaissance des locutions et indexation morphosyntaxique, les *emojis*, ici indexés comme « adjectifs supplémentaires », ne composent qu'1,16 % des types grammaticaux, ce qui est très loin des proportions qui peuvent être observées dans les analyses de réseaux socionumériques.

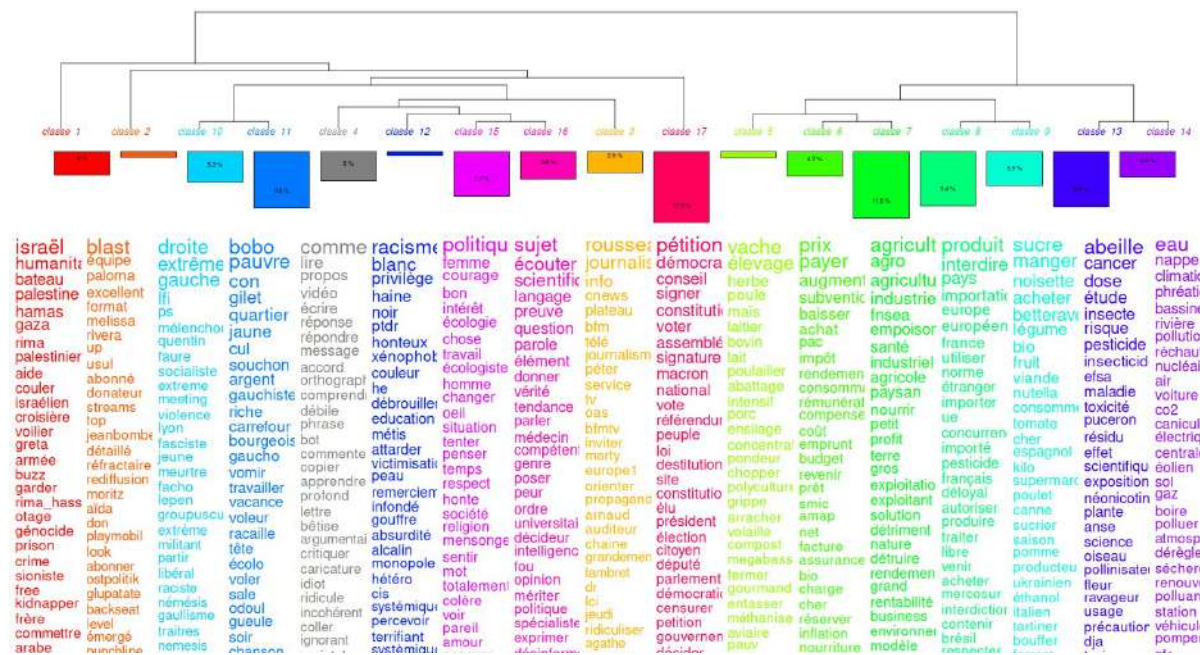
Type	Effectif	Pourcent
nom	225075	17,45
pre	148005	11,48
art_def	122631	9,51
ver	110133	8,54
adv_sup	83184	6,45
pro_per	79246	6,15
aux	73038	5,66
adj	70917	5,50
con	67353	5,22
nr	56044	4,35
art_ind	42460	3,29
pro_rel	34004	2,64
ver_sup	31741	2,46
pro_dem	31643	2,45
pro_ind	20386	1,58
nom_sup	15273	1,18
adj_pos	14940	1,16
adj_sup	14916	1,16
adv	14498	1,12
num	11575	0,90
adj_dem	6983	0,54
ono	6477	0,50
adj_ind	5332	0,41
adj_num	2088	0,16
adj_int	1367	0,11
pro_pos	147	0,01
art_	53	0,00
	1 289 509	100

Tableau 2 : Effectifs et pourcentages des types grammaticaux

4.1. Analyse standard

On effectue tout d'abord une analyse selon un plan relativement standard, dans lequel les catégories analysées sont essentiellement les mots pleins (noms, verbes, adjectifs, adverbes, formes non-reconnues), les autres (articles, connecteurs, auxiliaires, chiffres... ainsi que les *émojis*) étant placés en variables supplémentaires.

Le dendrogramme de la CDH est le suivant :



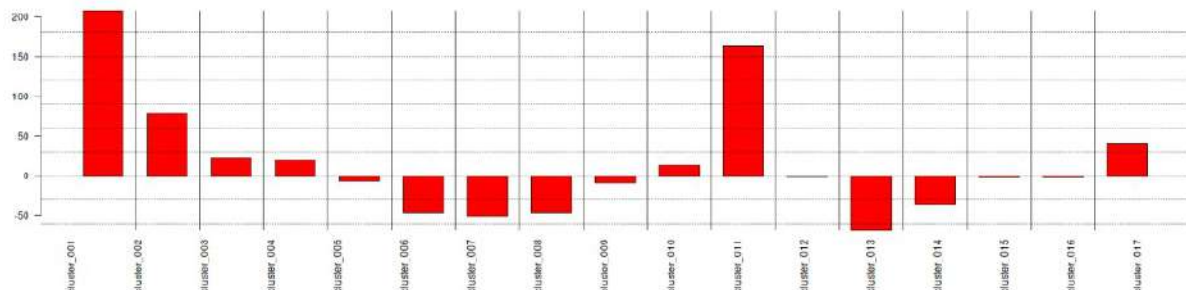
Graphique 2 : Dendrogramme de la classification, taille de classes et premières formes caractéristiques

La CDH permet de classer 91,26% des segments de texte en 17 classes lexicales, qui se répartissent en deux grands « blocs » définissant des *narratifs* différents de la « loi Duplomb ». A la suite de P. Marchand et P. Ratinaud (2024), ces résultats peuvent être interprétés selon le modèle sociocognitif des « systèmes de pensée » de D. Kahneman (2011).

Un premier « bloc » (classes 5, 6, 7, 8, 9, 13, 14) regroupe des lexiques directement liés à la thématique de la « loi Duplomb » : agriculture, élevage, environnement, climat, alimentation... On peut interpréter l'usage de ces lexiques comme mobilisant le système 2 de pensée, *systématique* et analytique, focalisé sur l'objet du débat lui-même.

Un second « bloc » (classes 1, 2, 17, 10, 11, 3, 4, 12, 15, 16) regroupe des lexiques non directement liés à la thématique de la « loi Duplomb » : politique, journalisme, société, événements d'actualité... mais également des invectives, voire des propos insultants. On peut interpréter l'usage de ces lexiques comme mobilisant le système 1 de pensée, *heuristique* et intuitif, qui excède l'objet du débat lui-même et actualise des rapports conflictuels.

La distribution des *emojis* par classes de la CDH permet d'obtenir le graphique suivant :



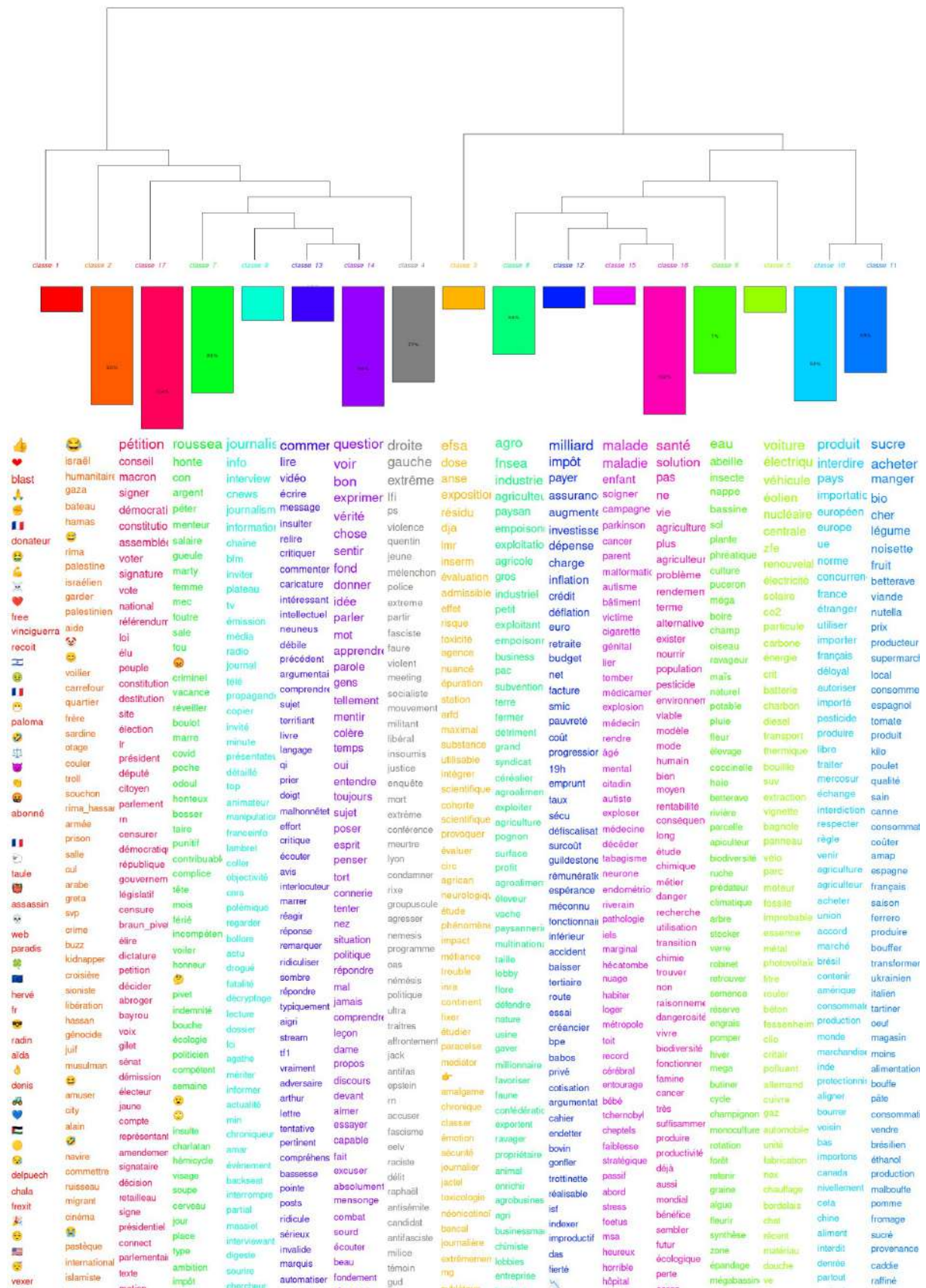
Graphique 3 : chi2 de liaison de la catégorie « emojis » aux différentes classes de la CDH

[illegible]

1. dans quelle rêve a t elle cru qu **israël** allait **laisser** ce **bateau humanitaire** accoster à **gaza**
2. l **aide humanitaire** est détournée par le **hamas** qui en fait commerce le **hamas** attise la **haine** contre **israël** et se sert de la population comme **bouclier** humain si le **hamas** avait **libéré** les **otages** il n y aurait jamais eu cette **guerre**
3. **rima_hassan** vous avez refusez la **libération** de **boualem_sansal** **israël** apporte déjà sont **aide humanitaire** au **palestiniens**
4. les **projecteurs** sur les **crimes** et le **blocus humanitaire** d **israël** fait **subir** à **gaza** obliger la france pays des **droits** de l **homme** a réagir c est déjà pas mal
5. paraît il des **camions humanitaires** ont pu enfin suite à l **arrivée** en **bateau** de grands écologistes européens **entrer** dans la **bande** à **gaza** en **palestine** pour donner à aux **civils palestiniens** journalistes **affamés** moi pas journaliste à vérifier 11 juin 2025 ici département 69
6. on est tous **rima** ❤️ ❤️ ❤️ ❤️ ❤️ ❤️ ❤️ ❤️ free **palestine** resistance free les **combattants palestiniens** ❤️ ❤️ ❤️ ❤️ ❤️ ❤️ ❤️ abat **israël** teoristes
7. mais tout cela elle le savait encore une **dinguerie** de **rima** qui pourra encore **craché** sur **israël** on est pas des sauvages elle critique **israël** tout le temps et elle pensait qu on la **laisserai** avec son **voilier humanitaire** 😂😂😂
8. **israël** 🇮🇱 veut **colonisé** **gaza** 🇵🇸 ça toujours été la **palestine** 🇵🇸 **israël** n existe que depuis 1948 **créé** par l **onu** 🇺🇳 1947 **arrivée** des injustes **juifs** en **palestine** 🇵🇸
9. **israël** coule un **bateau** avec **greta thunberg** et **rima_hassan** à son **bord** qui s en sort

342

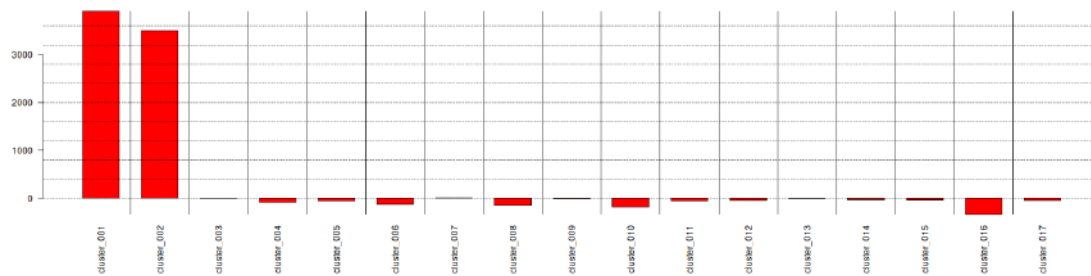
4.2. Analyse intégrant les emojis



Graphique 4 : dendrogramme de la classification avec les émojis en formes actives

La CDH intégrant les *emojis* en variables actives permet de classer 91,34% des segments de texte en 17 classes lexicales, à nouveau, qui restituent également les deux grands « blocs » renvoyant toujours aux *narratifs systématiques* (3, 10, 11, 5, 6, 8, 12, 15, 16) et *heuristiques* (1, 2, 17, 4, 7, 9, 13, 14).

On observe, sur le dendrogramme précédent, une première classe (2 % des segments classés) fortement saturée en *emojis*. Le graphe de distribution suivant confirme l'emploi significatif de la catégorie globale des *emojis* dans les classes 1 et 2.



Graphique 5 : chi2 de liaison de la catégorie « emojis » aux différentes classes de la CDH

Les classes qui rejettent le plus significativement les *emojis* apparaissent moins lisiblement à cause de la saturation des deux premières, mais sont, ici encore, celles qui révèlent une pensée *systématique* focalisée sur l'objet (16, 10, 6, 8, 11).

Les segments les plus caractéristiques de cette classe 1 le confirment :

1. jardin est un **écrivain humaniste gratitude** 🙏 à lui pour son regard pour les gueux 🇫🇷 ❤️ 🇵🇷 véritable **leader** pour le coup 👍 👏
2. 🙄 😊 le projet malthus continu et plus le temps passe plus les riches **exploiteurs** les lobbys industrielles m en donnent la preuve dommage on ne se trompe pas sur le déshumanisation voulue et orchestrée merci **blast** pour votre **travail exceptionnelle** ❤️ ❤️ 👍 👍 👍 👍 👍
3. ❤️ 👍 👏 👏 👏 👏 👏 🙏
4. merci à **alexandre** jardin 👍 🙏 et vive les gueux !! 🇫🇷 ❤️ 🇵🇷 🙌 🙌 🙌 aider pas punir 🌟
5. 👍 ❤️
6. mon futur président de la république ❤️ 👍
7. un **entretien d utilité** publique ❤️ sans vouloir diminuer la contribution de pierre **sujobert** j ai trouvé le témoignage de **fleur breteau** vraiment **exceptionnel** 😊 merci au deux merci **paloma** et merci **blast** 🙏
8. il n est jamais clair faure 🤔 c est dingue 😊 quand c est **flou** il y a un loup 🐺 🤔 🤔 🤔 🤔 🤔 🤔 merci mme **aubry** 🙌 🙌 🙌 🙌 🙌 vive lfi 👍 👍 🙏 👏 🎉

5. Conclusion

On rappellera d'abord que le taux d'*emojis* dans les commentaires postés sur les vidéos en ligne de grands médias d'information est très inférieur à ce qui a pu être observé sur les réseaux socionumériques. Et il est rare qu'une séquence d'*emojis* constitue l'intégralité d'un segment de texte. L'usage le plus fréquent ici est bien celui d'une connotation par ajout de marqueurs émotionnels. Des analyses devront donc s'appliquer à d'autres terrains pour tester la méthode proposée.

Néanmoins, l'analyse textométrique des commentaires permet de valider la pertinence du modèle des deux systèmes de pensée de Kahneman (2011). L'intégration des *emojis* dans le plan d'analyse montre qu'ils sont de bons révélateurs de la pensée heuristique par rapport à la

pensée analytique. En revanche, la nature des *emojis* mobilisés dans les classes ne semble pas différencier un usage spécifiquement dénotatif ou connotatif. On note que les *emojis* dénotatifs apparaissent parfois, soit pour doubler des formes pleines, soit pour s'y substituer. Dans ce dernier cas, les ignorer revient à perdre une occurrence lexicale importante.

Partant du principe selon lequel la pensée heuristique est plus sensible aux biais cognitifs que la pensée systématique (Marchand & Ratinaud, o.c.), nos résultats peuvent avoir des conséquences en cybersécurité pour la détection de tels biais, voire de tentatives de désinformation. On pourra, par exemple, s'attacher à différencier les *emojis connotatifs*, qui marquent une émotion (❤️, 👍, 🙌, 🙏, 🙄, 🙇, ...) et révèlent une attitude *syntagmatique* (Beauvois & Ghiglione, 1981), et les *emojis dénotatifs*, qui mobilisent une référence (🌽, 🌻, 🚗, 🐄, 🇫🇷, 🇪🇸, 🇺🇸, ...) et marquent une attitude *paradigmatique* (idem). Ce dernier cas pourra distinguer des références explicites, dont le rapport au contenu lexical est manifeste, et des références plus ou moins implicites, certaines relevant de la métaphore compréhensible par un public général (🐛, 🐜, 🐞, ...), d'autres plus latentes, qui révèlent un traitement plus ciblé pour un public spécifique (*dog whistle*) et/ou destiné à contourner les procédures de régulation numérique (🍉, 🍷, ...).

D'autres recherches devront approfondir ces premiers résultats en intégrant, éventuellement, des ressources de *Natural Language Processing* ou de *Sentiment Analysis*.

References

- Beauvois J-L., Ghiglione R. (1981). *L'homme et son langage - Attitudes et enjeux sociaux*. Paris : PUF.
- Fernández-Gavilanes M., Juncal-Martínez J., García-Méndez S., Costa-Montenegro E., González-Castaño F.J. (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. *Expert Systems with Applications*, 103 : 74-91.
- Kahneman D. (2012). *Système 1 / Système 2 : Les deux vitesses de pensée*. Paris : Flammarion.
- McGillivray B. (2024). Quantitative analyses of diachronic semantics: corpus annotation, disambiguation, and change detection. *JADT 2024 - 17th International Conference on Statistical Analysis of Textual Data*, Bruxelles.
- Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. *PLOS ONE* 10(12): e0144296. <https://doi.org/10.1371/journal.pone.0144296>
- Marchand P., Ratinaud P. (2024). Croiser ADT et NLP pour caractériser les commentaires en ligne et détecter les tendances complotistes : le cas des vaccins. *Mots comptés, textes déchiffrés (JADT 2024)*, Anne Dister; Dominique Longrée, Jun 2024, Louvain (BE), pp.593-602.
- Ratinaud P. (2018). Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In Iezzi D. F., Celardo L. & Misuraca M. (Eds.), *JADT' 2018, Proceedings of the 14th international conference on statistical analysis of textual data*. Vol. 2. Rome, Italie : Universitalia, pp. 616-625.
- Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8 (2), 187-198.
- Robertson A., Ferdousi Liza F, Nguyen D., McGillivray B., Hale S. (2021). Semantic Journeys: Quantifying change in emoji meaning from 2012-2018. *Emoji 2021 Workshop*
- Tang H, Tang W, Zhu D, Wang S, Wang Y, et al. (2024) EMFSA: Emoji-based multifeature fusion sentiment analysis. *PLOS ONE* 19(9): e0310715. <https://doi.org/10.1371/journal.pone.0310715>

Étudier les vidéos TikTok : comparaison lexicométrique des plus-values et des limites des retranscriptions automatiques versus les descriptions des vidéos

Panos Tsimpoukis¹, Lucie Loubere², Pierre Ratinaud³, Charis Papaevangelou⁴, Nikos Smyrnaio⁵

¹Université de Toulouse – panagiotis.tsimpoukis@utoulouse.fr

²Université de Toulouse – lucie.loubere@univ-tlse2.fr

³Université de Toulouse – pierre.ratinaud@univ-tlse2.fr

⁴Université de Toulouse – charilaos.papaevangelou@utoulouse.fr

⁵Université de Toulouse – nikolaos.smyrnaio@utoulouse.fr

Abstract

Digital social media has now become a distinct field of social science research, and the discourse found there provides material for investigation across numerous disciplines. However, until recently, the video format of certain platforms posed challenges for processing these spoken texts, and the widespread adoption of LLMs promises to change this. While data access platforms, combined with advances in transcription tools, have made it possible to quickly and easily transform existing spoken discourse, they raise questions about the time and, above all, the energy costs involved.

The work presented here examines the relationship between the descriptions of videos available on TikTok and their automatically transcribed audio content. We will show that these shorts, easily accessible clips already provide numerous elements that can be used to narrow down the video samples to be collected.

FORSEE is funded by the framework of the Horizon Europe Research and Innovation Programme under grant agreement N° 101177579

Keywords : TikTok, digital social media, video discourse, CHD Reinert, Analysis of specific features

Résumé

Les réseaux sociaux numériques constituent aujourd'hui un terrain à part entière de la recherche en sciences sociales et les discours qui y sont présents offrent une matière à investiguer dans de nombreuses disciplines. Cependant, le format vidéo de certaines plateformes constituait jusqu'à peu un frein à l'analyse de ces données, et l'arrivée des LLM promet aux chercheurs de nouvelles possibilités. En effet, si les plateformes d'accès aux données, associées aux progrès des outils de retranscription, permettent de transformer aisément les discours oraux en texte, elles soulèvent des questions sur le coût temporel, mais également énergétique, qu'elles impliquent. Ces réflexions nous ont amenés à comparer la plus-value des discours oraux retranscrits face à des données moindres, mais plus facilement accessibles (les descriptions des vidéos par exemple).

Le travail présenté ici interroge la relation entre la description des vidéos disponibles sur TikTok et leurs contenus audio retranscrits de façon automatique. Nous montrerons que les courtes présentations, facilement accessibles, éclairent sur les contenus des vidéos, permettant de restreindre les échantillons de vidéos à récolter.

Le projet FORSEE est financé dans le cadre du programme «Horizon Europe» pour la recherche et l'innovation, au titre de la convention de subvention n° 101177579

Mots clés : TikTok, réseaux sociaux numériques, discours vidéo, CHD Reinert, Analyse de spécificités

1. Introduction

Les réseaux sociaux numériques (RSN) occupent une place de plus en plus importante dans notre société. Comme l'annoncent les derniers résultats de l'enquête Digital News Report (2025) du Reuters Institute, ces outils ont, pour la première fois, dépassé les autres médias (télévision, sites web, radio, etc.) dans les pratiques informationnelles. Ces résultats confirment les recherches en sciences humaines et sociales, qui étudient depuis longtemps l'impact de ces dispositifs sur les pratiques de diverses populations. On retrouve, de façon non exhaustive, la politisation des individus (Cervi et al., 2021 ; Hindarto, 2022), la circulation de l'information (Costes, 2023 ; Mercier et al., 2017), notamment dans le domaine de la santé (Ostrovsky & Chen, 2020), ou encore la consommation en général (Stenger & Coutant, 2009).

Dans cette agora, les plateformes spécialisées dans les vidéos, comme TikTok ou YouTube, représentent un enjeu pour la recherche en sciences humaines et sociales. Consultées par des populations plus jeunes et aux profils sociométriques différents de ceux des autres plateformes, les contenus qui y sont véhiculés sont plus complexes à appréhender dans leur ensemble, notamment en raison de leur format vidéo. Les discours qui y sont présents ont pourtant un impact sociétal important, qu'il s'agisse de politique interne (Vasconcellos et al., 2023) ou d'autres sujets sociétaux, comme les impacts sur la santé mentale (Chartrand, 2022 ; Jehel, 2025).

Pour être exploités comme corpus textuel, les discours de ces productions nécessitent un traitement de retranscription. Si YouTube propose des sous-titres générés automatiquement dans ses exports, la plateforme TikTok n'en fait pas de même, et l'usage d'un outil supplémentaire est donc nécessaire. Cette tâche peut aujourd'hui être aisément automatisée, mais le temps et le coût énergétique que demande cette manipulation interrogent sa pertinence, en comparaison avec l'étude de la simple description des vidéos (accessible via l'API).

Nous proposons, dans cette étude, une comparaison lexicométrique d'un corpus constitué des retranscriptions automatisées de vidéos et d'un corpus constitué des descriptions de ces mêmes vidéos. L'objectif est de mettre en perspective les éléments communs et les éléments saillants d'une méthode à l'autre.

2. Recueil des données

2.1. Les données exploitables sur l'API TikTok

Les contenus diffusés sur TikTok sont des vidéos auxquelles s'ajoute une description directement renseignée par le compte émetteur. Il est également possible pour les utilisateurs de soutenir les vidéos vues et de les commenter (mais cette partie ne sera pas exploitée dans cette analyse). La documentation de la plateforme indique la possibilité d'enregistrer des vidéos jusqu'à 10 minutes, et il est également possible d'importer des vidéos jusqu'à 60 minutes. La description qu'il est possible de saisir par l'émetteur de la vidéo est limitée à 4 000 caractères et peut contenir du texte, des hyperliens et des hashtags.

L'accès aux données de cette plateforme est rendu possible par une API (2024) mise à disposition des chercheurs académiques. Cet outil permet, entre autres, de collecter des informations sur les vidéos à partir de requêtes basées sur des mots-clés, des auteurs, des périodes et la géolocalisation. Parmi ces données, nous retrouvons les identifiants de la vidéo, l'auteur, leur date de création, les indicateurs d'engagement (likes et partages) et la description textuelle du contenu.

2.2. Le recueil des données

Les corpus sur lesquels repose cette étude proviennent du projet européen FORSE (Forging Successful AI Applications for European Economy and Society), dont le WP4 étudie les discours véhiculés sur l'intelligence artificielle à travers les médias et les réseaux sociaux numériques.

Les exports via l'API étant restreints, nous avons échelonné l'extraction des données au mois d'octobre 2025, à partir d'une requête basée sur les mots-clés « intelligence artificielle » ou « ChatGPT », localisée en France, en utilisant des séquences temporelles de 7 jours consécutifs, allant du 01/01/2022 au 31/06/2025. Ces extractions ont retourné 31 806 vidéos correspondant à notre recherche, pour lesquelles nous avons récupéré les descriptions.

La seconde étape a consisté à télécharger les vidéos sélectionnées, mais seules 20 755 restaient disponibles, les autres étant soit en accès privé, soit supprimées de la plateforme. Nous avons effectué une retranscription automatique sur cet ensemble de vidéos.

La retranscription repose sur l'outil Whisper, développé par OpenAI. Pour cette expérimentation, nous avons utilisé le niveau large, qui permet d'obtenir un résultat plus proche des locutions (Bertin & Quellec, 2025), notamment au niveau de l'extrapolation de la ponctuation à partir de la prosodie. Ce critère nous est apparu important, puisqu'il est utilisé par la suite pour la segmentation des verbatims.

Afin d'être cohérents dans notre comparaison entre les descriptions et les contenus retranscrits des vidéos, nous avons choisi de limiter le corpus de descriptions aux vidéos présentes dans celui des retranscriptions. Les deux corpus contiendront donc 20 755 textes.

3. Descriptions des corpus

Les éléments comparatifs des deux corpus sont repris dans le tableau 1.

TABLE 1 : *Éléments comparatifs des deux corpus*

	Descriptions	Retranscriptions
nombre de segments de texte	47 465	98 836
nombre d'occurrences	1 493 444	3 246 921
nombre de mots	80 607	62 699
nombre d'hapax	43 434	26 654

Ainsi, la retranscription des vidéos a généré deux fois plus de données textuelles, et la présence de 17 908 mots supplémentaires dans le corpus des descriptions s'explique en très grande partie par le nombre d'hapax (16 780 de plus). Ces derniers proviennent, pour la majorité, de hashtags.

4. Calcul de spécificités

Afin d'observer les variations entre le discours écrit, que sont les descriptions, et les retranscriptions des vidéos, nous avons constitué un corpus rassemblant et identifiant ces deux types de sources. Nous y avons appliqué un calcul de spécificité, qui compare, pour chaque mot, sa distribution entre ces deux types de données (Lebart et al., 2019).

Les résultats montrent des marqueurs spécifiques au langage oral (« aller », « voir », « chose ») dans les retranscriptions des vidéos, tandis que les descriptions sont caractérisées par la présence de hashtags et de liens hypertextes (« https », « com », etc.).

Au-delà de ce formalisme, nous observons un biais dans l’encodage des vidéos sans paroles, qui sont balisées par le mot unique « musique ». Il arrive également que les longs passages silencieux soient mentionnés de la même façon.

Pour le reste du lexique, la différence n’est pas tant basée sur les thématiques que sur la formulation : les descriptions utilisent davantage de noms, tandis que les vidéos sont plus axées sur les verbes. Ainsi, le mot « création » est spécifique aux descriptions, alors que les verbes « créer » et « générer » sont spécifiques aux retranscriptions.

À la lecture des contenus, nous trouvons une première explication dans l’usage de la mention « créé par l’IA » dans les descriptions de nombreuses vidéos ne contenant qu’une musique.

La seconde explication repose sur un décalage entre une dénomination du contenu de la vidéo dans les descriptions, qui utilise plus facilement des noms, alors que nous retrouvons un discours commentant l’action dans les vidéos, qui exploitera plus de verbes. Nous citerons ici quelques extraits illustrant ces tendances :

TABLE 2 : *Extraits de retranscriptions et descriptions de vidéos*

id vidéo	retranscription	vidéo
740082612993589584	[...] une formation complète, de la création des modèles à [...]	[...] j’ai créé un guide complet [...]
746464237400269135	[...] utiliser l’intelligence artificielle, et faire le plein de créativité [...]	[...] L’ia peut te proposer de te créer une accroche textuelle et de générer ta description [...]
74114001238271460	[...] rejoints mois sur cette aventure et accéder à la formation [...]	[...] Pour ne pas rester à la traîne, il est impératif de se former dès maintenant [...]

Cette analyse indique donc bien une concordance des thématiques entre les deux corpus, ainsi qu’une différenciation du lexique qui tient davantage du format proposé par le dispositif que d’une ligne éditoriale.

5. Classification hiérarchique descendante

L’objectif de cette étape est de mettre en évidence les thématiques structurantes des deux corpus. Pour cela, nous utilisons la CHD de type Reinert, telle qu’elle est proposée dans le logiciel Iramuteq (Ratinaud, 2024).

Cette analyse mise au point par Max Reinert (1993) repose sur une segmentation du discours, puis un classement de ces segments par une analyse factorielle par bipartition d’un tableau lexical reprenant les mots actifs présents ou absents dans les segments. Le classement opéré par ce processus se focalise donc non plus sur la spécificité des mots, mais sur leurs cooccurrences au sein des mêmes segments. Ce classement permet de mettre au jour les « mondes lexicaux », qui sont les traces d’une « activité cohérente » entre le point de vue de l’auteur et le contexte d’énonciation (Reinert, 1993, p. 11).

Comme nous l’avons vu précédemment, les deux sources de données sont hétérogènes dans leurs formes énonciatives. Étudier les mondes lexicaux dans ce contexte ne serait pas pertinent, puisque chaque monde lexical se forme en rapport avec l’ensemble du corpus étudié. Afin d’observer les différentes thématiques présentes dans chaque ensemble, nous reviendrons donc à une analyse de chaque corpus indépendamment de l’autre.

5.1. Le corpus de descriptions

La classification des descriptions proposée pour les descriptions de vidéos est issue d'une CHD de 125 classes en phase 1, avec un minimum de 500 segments de texte (ST) par classe. Elle reprend 88 % du corpus en 11 classes, présentées dans la figure 1.

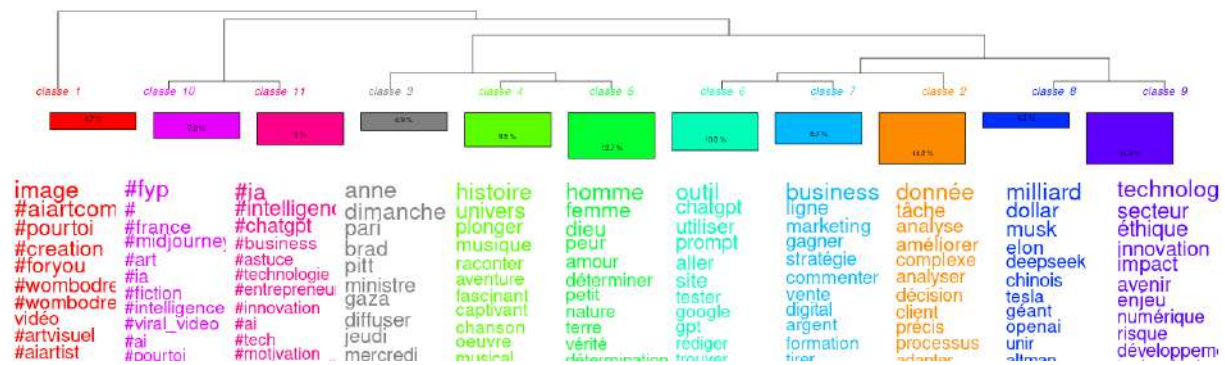


FIGURE 1 : Dendrogramme et profil de la classification des descriptions.

Nous observons une première rupture séparant les classes portant plusieurs hashtags (classes 1, 10, 11). Après plusieurs essais, nous avons décidé de conserver le caractère # et ainsi garder l'intégralité de ces formes de discours. Ce choix est motivé par la persistance de mots non reconnus issus de ces mots-clés, même sans le hashtag (par exemple « pourquoi », « aiartcom »...), alors que d'autres, comme « France » ou « art », se mêleraient aux énonciations de ces mots dans le reste du discours.

Le second bloc de classes (3, 6 et 5) porte sur la création artistique et humaniste. Nous y retrouvons l'actualité des personnalités publiques, ainsi que l'IA au service de la création numérique, qu'elle soit narrative (histoires mises en vidéo) ou graphique (classe 4). Le dernier discours de ce groupe est centré sur l'homme et la femme : nous y trouvons un mélange de propos allant de la création divine aux réflexions sur la cohabitation des hommes et des femmes, voire leur séduction.

Le troisième bloc est composé des classes 6, 7 et 2. Il porte un discours technique abordant l'utilisation de l'IA, que ce soit à travers des retours d'expérience avec un discours général (classe 6), son usage dans l'activité économique (classe 7) ou encore dans la prise de décision (classe 2).

Enfin, les deux dernières classes sont composées d'un discours économique sur le secteur de l'IA.

5.2. Le corpus de retranscriptions

La classification des retranscriptions a été réalisée avec 90 classes en phase 1 et un minimum de 3 000 segments de texte (ST) par classe. Elle présente, en 9 classes, 85 % du corpus conservé.

Le premier groupe de classes à se détacher est composé d'une classe portant sur l'actualité internationale (classe 1), d'une autre sur le discours concernant la famille (classe 9), et enfin d'une classe sur les relations entre humains, sur fond de religion (classe 5).

Dans le second bloc, nous retrouvons les classes techniques, orientées à la manière de tutoriels ou de tests (classes 6 et 7), ou encore sous un angle marketing. En revanche, une classe marquée par des formules engageantes (« commentez », « partagez », « abonnez-vous »...) apparaît (classe 2).

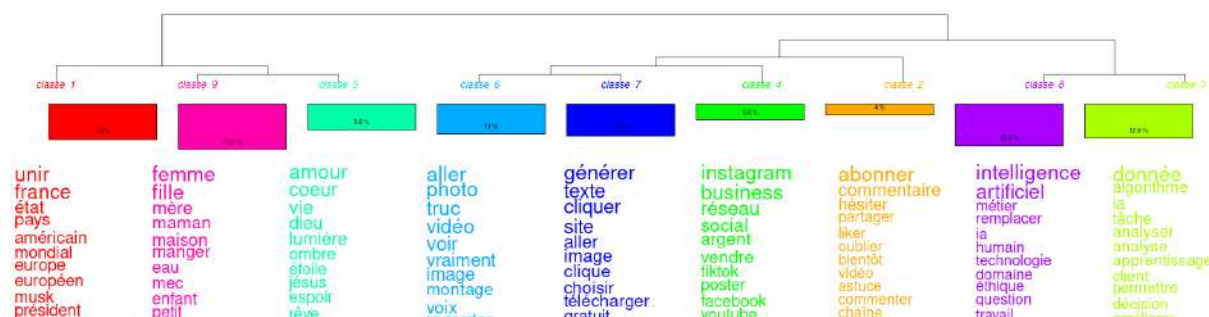


FIGURE 2 : Dendrogramme et profils sur le corpus des retranscriptions vidéos.

Enfin, les deux dernières classes portent sur l'impact social et économique de l'IA (classe 8) et son usage pour la prise de décision (classe 3).

Nous observons bien la présence des thématiques principales dans les deux corpus, indiquant ainsi une relation entre les descriptions et les contenus oraux des vidéos.

6. Persévérance des vidéos sources à travers les deux corpus

Afin de mesurer si les sources de contenus (vidéos) se répartissent de façon homogène et similaire dans les deux classifications, nous avons effectué un test de corrélation. Pour cela, nous avons réalisé un test du χ^2 pour chaque couple de classes (corpus des retranscriptions vs corpus des descriptions). Ce test a été calculé sur un tableau de contingence reprenant pour chaque couple de classes le nombre de vidéos présentes ou absentes dans les classes.

TABLE 3 : Exemple de tableau de contingence pour calcul du χ^2 de liaison interclasse des deux classifications

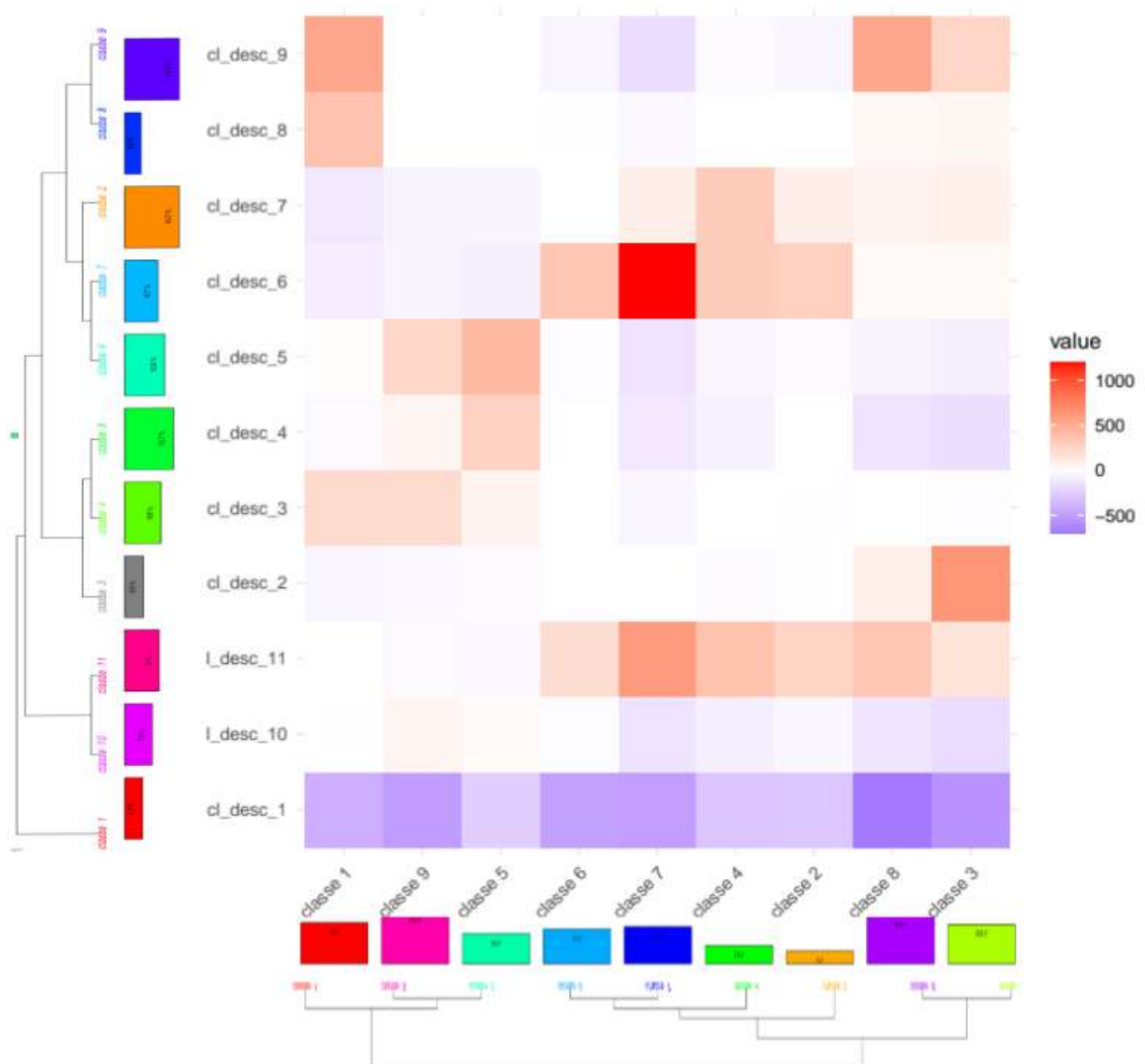
	nombre de vidéos présentes dans la classe du corpus de description	nombre de vidéos absentes dans la classe du corpus de description
nombre de vidéos présentes dans la classe du corpus de retranscription	xx	xx
nombre de vidéos absentes dans la classe du corpus de retranscription	xx	xx

Les résultats sont représentés dans la figure 3, où plus la valeur du χ^2 est élevée (indiquant une surreprésentation des vidéos présentes dans les 2 classes), plus la case est rouge. À l'inverse, plus il y a sous-représentation, plus la cellule est bleue.

Nous observons ici que les classes portant sur les mêmes thématiques sont également centrées sur les mêmes vidéos.

7. Conclusion

Si la description des corpus issus des descriptions ou des retranscriptions des vidéos nous a montré une plus grande quantité de texte dans les vidéos que dans les descriptions (plus du double); l'étude comparative que nous avons menée nous permet de mettre en perspective une concordance entre ces deux sources de données. Si les deux ensembles montrent des particularités

FIGURE 3 : *Heatmap sur la corrélation des sources entre les deux corpus.*

formelles comme les hashtags pour les descriptions et les appels à engagement pour les vidéos, les thématiques structurantes identiques ressortent des classifications. Cette proximité thématique est complétée par la présence des mêmes vidéos dans les classes de discours portant sur les thématiques similaires à travers les deux corpus.

Ces résultats nous amènent à reposer notre question de départ sur l'utilité d'investiguer les contenus des vidéos compte tenu de la masse de données disponibles sur la plateforme, et du temps et ressources nécessaires pour le téléchargement et la retranscription des vidéos.

Les deux approches semblent complémentaires, la fouille des descriptions par sa facilité d'accès et sa frugalité de traitement nous permettrait d'identifier les vidéos portant sur les thématiques pertinentes pour notre étude. Ce travail préalable permettrait une forte réduction du nombre des contenus à télécharger et à retranscrire. Cela offrirait à la fois une investigation plus fine des vidéos sélectionnées et une économie de temps, mais surtout d'énergie, dans la pratique de ce type de recherche.

Références

- Bertin, T., & Quéllec, G. (2025). Transcription automatique des interactions verbales. Limites observées et perspectives envisagées à partir d'un corpus de consultations médicales. *Corpus*, (26). <https://doi.org/10.4000/1364r>
- Cervi, L., Tejedor, S., & Lladó, C. M. (2021). TikTok and the new language of political communication. *Cultura, lenguaje y representación*, 26, 267-287.
- Chartrand, A. (2022). *Relations entre l'utilisation de TikTok et le bien-être chez les jeunes adultes*.
- Costes, M. (2023). " *S'informer quand on est étudiant en INP : Tendances et pratiques informationnelles* ", *Rapport de recherche, Université Toulouse III Paul Sabatier, LERASS, juillet 2023*. LERASS. <https://hal.science/hal-04854039>
- Hindarto, I. H. (2022). Tiktok and political communication of youth : A systematic review. *Jurnal Review Politik*, 12(2), 146-176.
- Jehel, S. (2025). Audition du 23 avril 2025 Commission d'enquête sur les effets psychologiques de Tiktok sur les mineurs. *Santé mentale et Tiktok*.
- Lebart, L., Pincemin, B., & Poudat, C. (2019). *Analyse des données textuelles*. PUQ.
- Mercier, A., Ouakrat, A., & Pignard-Cheynel, N. (2017). Facebook pour s' informer ? Actualité et usages de la plateforme par les jeunes. # *info. Commenter et partager l'actualité sur Twitter et Facebook*, 169-197.
- Ostrovsky, A. M., & Chen, J. R. (2020). TikTok and its role in COVID-19 information propagation. *The Journal of Adolescent Health*, 67(5), 730.
- Ratinaud, P. (2024). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. *Téléchargeable à l'adresse : <http://www.iramuteq.org>*.
- Reinert, M. (1993). Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, 66, 5-39.
- Stenger, T., & Coutant, A. (2009). La prescription ordinaire de la consommation sur les réseaux socionumériques : De la sociabilité en ligne à la consommation ? *14e Journées de recherche en marketing de Bourgogne*, 12-13.
- TikTok. (2024). *TikTok Research API Documentation* [Logiciel]. <https://developers.tiktok.com/doc/overview>
- Vasconcellos, P. H. S., Lara, P. D. de A., & Marques-Neto, H. T. (2023). Analyzing polarization and toxicity on political debate in Brazilian TikTok videos transcriptions. *Proceedings of the 15th ACM Web Science Conference 2023*, 33-42.

C02

French NLP, lexicon and contrastive neology

09-lug
15:00 - 16:30
Aula Multimediale C

Vers une correction automatique des formes inconnues : application du Masked Language Modeling de CamemBERT au corpus océrisé des Cahiers citoyens

Sami Guembour, Solomiia Korol, Catherine Dominguès

Univ Gustave Eiffel, Géodata Paris, IGN, LASTIG, F-77420 Champs-sur-Marne
prénom.nom@ign.fr

Abstract (in English)

In 2019, le Grand Débat national led to make the digital corpus *Cahiers citoyens*. Preliminary analyzes showed that the OCR process produced poor results and several types of errors were identified. This paper describes the process of resolving one of them related to unknown forms. It relies on semantic index (the MLM task provided by CamemBERT-base suggest probable words that could replace the unknown form) and on a typographic index (the best word is that one that most closely matches the typographic form of the unknown form). Tests were conducted to select the best similarity measure (Jaro-Winckler) and to define the minimum similarity threshold to select the best probable word. The process was applied to the entire corpus and made it possible to correct 41% of the unknown forms.

Keywords: Cahiers citoyens, CamemBERT, MLM, typographic similarity, OCR correction, unknown form

Résumé

En 2019, la consultation du Grand Débat national a permis la construction du corpus numérisé des Cahiers citoyens. Des premières analyses sur ce corpus ont montré la mauvaise qualité de la transcription et ont conduit à identifier différents types d'erreur dont des formes inconnues. Cet article décrit le processus mis en place pour les corriger, qui s'appuie sur des indices sémantiques (la prédiction d'un mot probable obtenue par la tâche MLM du modèle de langue CamemBERT-base affiné) et typographiques (qui permettent de choisir parmi les prédictions la plus similaire au mot à corriger). Des tests ont été menés afin de choisir la mesure de similarité entre chaînes de caractères (Jaro-Winckler a été choisie) et de définir le seuil de similarité minimale qui permet d'optimiser le choix entre les différents mots prédits. Le processus a été appliqué à l'ensemble du corpus et a permis de corriger 41% des formes inconnues.

Mots clés : Cahiers citoyens, CamemBERT, MLM, similarité typographique, correction d'OCR, forme inconnue

1. Introduction

En janvier 2019, en réponse au mouvement des Gilets jaunes, le président de la République lance le Grand Débat national¹ (GDN). Il s'adresse aux citoyen·ne·s pour leur proposer de participer à des débats au travers de deux modalités principales : une plate-forme numérique dématérialisée, où les participant·e·s peuvent répondre à des questions, et des cahiers physiques mis à disposition principalement dans les mairies pour une expression libre, sans contrainte de taille ni de sujet, appelés les *Cahiers citoyens*. Par le contexte de la consultation et sa temporalité fondée sur celle d'événements sociaux et politiques, par la diversité des acteurs impliqués et par le volume des contributions collectées, l'ensemble des contributions constitue

¹ <https://granddebat.fr/>

un matériau à la fois singulier et riche. Des travaux ont été engagés : sur le contexte de la collecte Duguet (2018), Pengam (2024), Pottier (2025), sur les contributions de la plate-forme Ploux et al. (2021), ou sur certains cahiers collectés dans des territoires spécifiques : Bendinelli (2023) à Dole, Noguera (2023) en Gironde, Baciocchi et al. (2024) pour la Somme. D'autres travaux se sont intéressés à la totalité des Cahiers citoyens pour en fournir une analyse diplomatique Chabin and Vasseur (2021), sémantique Authors (2024), Authors (2025), ou concernant certaines thématiques : l'éolien Jolivet et al. (2024), l'immigration Ploux and Dominguès (2025). Cependant peu de travaux se sont, à notre connaissance, interrogés sur la qualité du corpus numérisé des Cahiers citoyens. Le corpus sous sa forme numérique initiale est constitué de l'ensemble des images des pages des 20 152 cahiers. Ces images ont été océrisées pour produire un corpus textuel numérique de 225 224 contributions qui n'a été que partiellement vérifié, la vérification concernant à la fois la segmentation des pages des cahiers en contributions et la fidélité à la graphie originale des contributeur·rice·s. L'océrisation a achoppé sur la diversité des contributions. En effet, une contribution peut avoir été écrite directement sur une page d'un cahier, ou bien collée dans le cahier après avoir été écrite ou imprimée au préalable sur un autre support papier. Elle peut prendre la forme d'une lettre, d'une liste à puces, d'une suite de paragraphes. Des procédés graphiques ont été utilisés pour mettre en valeur un élément de la contribution : capitales, souligné, encadré, un seul mot sur une page, etc. Dans Dominguès and Jolivet (2024), la comparaison sur des critères textométriques des contributions provenant des cahiers océrisés et de celles de la plate-forme² (désigné par *cPF*) a montré un taux sensiblement plus important de formes non analysables, et donc d'information perdue, dans les premières que dans les secondes. L'influence de ces formes non analysables sur la qualité des analyses ultérieures ne peut être négligée, aussi différents travaux ont tenté de diminuer leur nombre.

Le processus décrit dans ce papier vise à corriger les formes inconnues. Le corpus et les types d'erreur identifiés sont présentés en section 2. Pour proposer des corrections, la méthode s'appuie sur le contexte sémantique des formes inconnues (donné par un modèle de langue) et la typographie de la forme inconnue ; elle est décrite et testée en 3. Après application à l'ensemble du corpus (en 4.), les résultats sont discutés en 5. qui se termine par une brève conclusion.

2. Corpus et erreurs

La table 6 récapitule les caractéristiques numériques du corpus des Cahiers citoyens (dorénavant *cCC*). Celles-ci diffèrent légèrement de celles présentées initialement dans Dominguès and Jolivet (2024), ligne V0 de la table 6, parce que des éléments comme des adresses postales et électroniques ont été automatiquement supprimés des contributions, réduisant aussi le nombre de formes et d'occurrences inconnues, qui reste important et témoigne de la mauvaise qualité de la transcription. Des analyses ont conduit à une typologie en trois types d'erreur :

- **mention *ILLISIBLE*** : lorsqu'une séquence de lettres n'est pas reconnue par l'opérateur humain, la mention manuelle "*ILLISIBLE*" lui est substituée. Ce type d'erreur ne concerne que les contributions vérifiées manuellement, mais le fait qu'une contribution ait été ou non vérifiée manuellement n'est pas noté dans les métadonnées des contributions ;
- **mot inconnu** : les mots inconnus sont ceux qui ne sont pas étiquetés (lemme, étiquette grammaticale, informations de flexion) lorsque le corpus est soumis à une analyse morpho-lexicale fondée sur des dictionnaires électroniques (réalisée ici à l'aide de l'outil Unitex

² Pour que les comparaisons soient légitimes, seules les réponses rédigées en texte libre ont été retenues ; les réponses aux questions à choix multiples, déjà rédigées, n'ont pas été prise en compte.

Paumier et al. (2021)). Les erreurs sont très majoritairement des erreurs d'océrisation ayant transformé un mot initialement valide, ou des coquilles, plus rarement des néologismes pour lesquels des dictionnaires spécifiques ont aussi été ajoutés ;

- **mot échangé** : un mot est dit échangé lorsqu'il est reconnu par les dictionnaires, mais considéré comme erroné dans le contexte de la phrase où il figure. Ce type d'erreur se produit lorsque la vérification manuelle n'a pas été réalisée ou n'a pas repéré ni corrigé l'erreur produite par l'outil d'océrisation. Son identification relève d'une interprétation humaine, par exemple *miches* et *riches* associés à *fiscales* pour produire les expressions *miches fiscales* et *riches fiscales* sont considérés comme des mots transcrits à la place de *niches* (dans l'expression *niches fiscales*), et donc échangés.

3. Méthode de correction des erreurs de type "mot inconnu"

La détection des mots inconnus s'appuie sur les dictionnaires d'Unitex ; leur correction vise à proposer un mot valide pour chaque forme inconnue. La méthode est fondée sur la fonction offerte par les modèles de langue qui permet de prédire un mot masqué dans une unité textuelle en fonction de son contexte, et sur une mesure de similarité typographique entre la forme inconnue et les prédictions fournies par le modèle. Des tâches préalables ont été réalisées afin d'ajuster la méthode. Puisque la détection des mots inconnus repose sur les dictionnaires d'Unitex, ceux-ci ont été adaptés au contexte de la consultation et aux thématiques traitées dans le GDN, en particulier l'écologie, en 3.1. Le contexte est essentiel pour la prédiction d'un mot masqué, aussi une expérimentation préalable a été conduite afin de déterminer le meilleur contexte pour la prédiction du mot inconnu en 3.2. Différentes mesures de similarité entre chaînes de caractères ont été testées en contexte afin de choisir la plus adaptée, en 3.3. Ainsi ajustée, la méthode de correction a été validée sur un échantillon en 3.4.

3.1. Enrichissement des dictionnaires

L'objectif est de diminuer le nombre de mots qui sont non analysés – ils n'ont alors pas d'étiquette grammaticale et donc ne peuvent être repérés par un patron morpho-syntaxique – ou mal analysés – à l'étape de calcul du vecteur de contexte par le modèle de langue, ces mots peu fréquents (éventuellement parce qu'ils sont mal formés) peuvent être segmentés, conduisant à la dilution de leur valeur sémantique.

Aussi les dictionnaires de base ont été enrichis pour tenir compte des thématiques abordées dans le GDN et du contexte politique. Il s'agit d'établir un compromis entre l'adaptation au contexte, et la multiplication d'analyses trop spécifiques. Certains ajouts peuvent être considérés comme des simples mises à jour des dictionnaires d'Unitex et sont réutilisables pour l'analyse de corpus contemporains. Il s'agit de formes lexicalisées déjà entrées dans des dictionnaires généraux (ici, Le Robert en ligne : <https://dictionnaire.lerobert.com>) : *altermondialiste*, *anonymiser*, *ubériser*, d'emprunts étrangers ou à des langues de spécialité : *agribashing*, *agroforesterie*, *anonymiser*, d'abréviations ou des squelettes consonantiques courants dans le discours médié par ordinateur Panckhurst (2009) : *agglo*, *alloc*, *interco*, *bcp*, *pdt*, *nbe*. D'autres sont ambigus ou trop spécifiques pour être utilisables hors contexte comme des acronymes et des marques : *AMAP*, *AMIF*, *AMRF*, *APL*, *UBER*, *ZAD*, *DVG*, *RN*, ou bien liés au contexte du GDN ou de validité temporelle limitée comme des entités nommées de personnes : *Cahuzac*, *Jancovici* ou bien des néologismes sur le nom du président de la République, obtenus par des schémas productifs : *macronie*, *macronien*, etc. ou plus rares : *macroneur*, *macronologue*. Les lemmes ajoutés sont en majorité des noms, et dans une moindre mesure des adjectifs et des verbes. Au total, près de 3 000 formes fléchies ont été ajoutées aux dictionnaires d'Unitex.

En outre, des corrections systématiques ont été faites pour tenir compte des variations orthographiques répétées observées dans le corpus: par exemple les séquences *acceuil* au lieu d'*accueil* (et aussi dans les formes fléchies du verbe *accueillir*), ou *renumération* (admettant aussi des variations sur les diacritiques) au lieu de *rémunération*. Certains objets ayant cristallisé les doléances au moment du GDN étant désignés par des acronymes, ceux-ci ont été recherchés et normalisés dans le corpus : *ISF*, *RIC* et *RIP*, *ENA*, etc. Au total près de 120 expressions régulières ont été appliquées pour corriger systématiquement, et donc hors contexte, ces formes initialement non reconnues, ou dispersées sur différentes orthographes. Cette version du corpus correspond à la ligne V2 dans la table 6.

3.2. Choix du modèle de langue et de l'unité de traitement

Une expérimentation a été menée afin de choisir le modèle de langue et l'unité de contexte des formes inconnues les plus adaptés à cette tâche. Deux modèles ont été testés : CamemBERT-large Martin et al. (2019) et CamemBERT-base affiné sur le corpus CC. Deux unités ont été testées : la phrase dans laquelle se trouve la forme inconnue ou bien la contribution complète. L'expérimentation a été conduite sur 150 contributions contenant chacune une seule forme inconnue (qui n'est pas une entité nommée de lieu ou de personne) ; le mot correct (la vérité terrain) a été déterminé par une prédiction manuelle à partir du contexte. Pour chaque forme inconnue, la prédiction propose une liste de k mots classés par probabilité décroissante ; le résultat est considéré positif si la vérité terrain est présente parmi les 20 premières prédictions. La table 1 montre que les meilleurs résultats sont obtenus avec CamemBERT affiné et la contribution comme unité de contexte de la prédiction.

Table 1: Résultats de l'expérimentation du choix du modèle et de l'unité de traitement

	phrase (en %)	contribution (en %)
CamemBERT-large	95 (63 %)	109 (73 %)
CamemBERT-base affiné	118 (79 %)	124 (83 %)

3.3. Choix de la similarité typographique et définition des seuils de similarité

La forme typographique inconnue constitue un indice pour sélectionner le mot prédit : celui qui est le plus proche typographiquement de la forme inconnue doit être préféré³. Plusieurs similarités (Jaro, Jaccard, MLIPNS, Jaro-Winkler, Needleman-Wunsch, Overlap, Cosine) ont été comparées entre 14 formes inconnues et les mots prédits correspondants selon trois critères : 1/ bien séparer les chaînes typographiquement proches *accueill/acceuil* de celles qui sont éloignées comme *accueill/esclavage* ; 2/ être sensible à l'ordre des lettres : la distance entre *accueil* et *acceuil* ne doit pas être nulle ; 3/ ne pas être (trop) sensible à la casse : les usages des majuscules et minuscules non conformes aux règles orthographiques étant très courants dans le corpus, ils ne doivent pas être sanctionnés. La similarité MLIPNS s'est révélée trop stricte, en donnant souvent une similarité nulle pour des mots très proches. Needleman-Wunsch et Overlap ont été éliminées parce que les résultats n'étaient pas facilement interprétables, Jaccard et Cosine parce qu'elles n'étaient pas sensibles à l'inversion des lettres (*accueil/acceuil* ont une similarité de 1). Finalement, Jaro-Winkler a été préférée à Jaro, parce qu'elle minimise la différence entre minuscule et majuscule (*acceuil/Acceuil*).

³ Le processus d'océrisation qui a permis d'obtenir la version textuelle du corpus guide la recherche d'une similarité typographique, et écarte celle d'une similarité phonétique.

La définition d'un seuil de similarité minimale permet d'éliminer les prédictions peu plausibles parce que typographiquement trop éloignées de la forme inconnue. Le seuil a été déterminé grâce à une expérimentation sur 100 formes inconnues dont la vérité terrain a été définie d'après le contexte de la contribution ; il s'agit d'observer expérimentalement quelle valeur du seuil permet d'éliminer les prédictions qui ne sont pas identiques à la vérité terrain. D'après la table 2, les valeurs de similarité entre forme inconnue et vérité terrain correspondante (souligné) obtiennent le meilleur score, quand la vérité terrain figure parmi les prédictions (ce qui n'est pas le cas pour *feroutage* ni *cicative*). Pour éliminer les prédictions erronées, le seuil minimal devrait être de 0,83. Cette première valeur de seuil a été affinée pour optimiser le ratio de corrections valides. La table 3 montre que le meilleur ratio de corrections valides est obtenu avec un seuil minimal de 0,87. Un seuil maximal de similarité a aussi été ajouté afin d'éliminer les prédictions identiques à la forme inconnue. Le modèle ayant été affiné sur le corpus, ce cas s'est produit, ouvrant ainsi d'autres pistes de corrections systématiques des contributions.

Table 2: Valeurs de similarité Jaro-Winkler entre formes inconnues et prédictions associées. Par définition, toutes les similarités donnent une mesure entre 0 et 1, ce qui permet de les comparer.

forme inconnue	mot1	mot2	mot3	mot4	mot5	mot6	mot7	mot8	mot9
<i>aniamux</i>	<u>0.93</u>	0.55	0.79	0.70	0.65	0.54	0.72	0.71	0.62
<i>feroutage</i>	0.64	0.45	0.64	0.81	0.62	0.45	0.69	0.65	0.65
<i>cicative</i>	0.37	0.53	0.62	0.53	0.67	0.33	0.61	0.54	0.45
<i>emmigré</i>	0.63	<u>0.83</u>	0.66	0.51	0.45	0.33	0.66	0.57	0.54

Table 3: Variation des taux de corrections valides selon les seuils de similarité minimale

Seuil	Nbe de formes corrigées	Nbe de corrections valides	en %
0.91	42	36	86%
0.87	51	45	88%
0.84	57	48	84%
0.81	62	49	79%

3.4. Validation de la méthode sur un échantillon

La méthode de correction consiste donc à masquer la forme inconnue, à la faire prédire par la fonction de masquage (MLM) du modèle de langue (CamemBERT-base affiné sur le corpus CC) et à sélectionner la forme prédite la plus plausible par sa ressemblance typographique avec la forme inconnue (évaluée par la similarité Jaro-Winkler). Elle a été mise en œuvre sur un échantillon de 188 contributions (unité de contexte de la prédiction) tirées au hasard parmi celles contenant une seule forme inconnue.

La validation est fondée sur l'indicateur d'accuracy : sur les 188 formes inconnues, 63 ont été corrigées et parmi elles 57 correspondent à la vérité terrain. Le taux d'intervention est donc de 34%, et celui d'accuracy plus de 90%, ce qui permet d'envisager d'utiliser cette méthode pour la totalité du corpus. La table 4 montre des exemples de correction de formes inconnues. Plusieurs types d'erreurs sont corrigés : la transformation d'une lettre (*pilliard/milliard*), l'omission de lettres ou de diacritiques (*administ/administratif*, *debat/débat*), l'inversion de plusieurs lettres (*renumération/rémunération*), l'omission de la majuscule sur l'initiale des noms propres (*yvelines/Yvelines*), et aussi la combinaison de plusieurs erreurs : par exemple l'omission de la majuscule initiale et remplacement d'une lettre comme dans *cherac/Chirac*.

Table 4: Exemples de corrections proposées. Tous les remplacements sont valides, sauf pour *ferrostage* pour lequel la meilleure proposition est : *ferroviaire*, et la vérité terrain : *ferroustage*.

Forme inconnue	Mot prédit	Forme inconnue	Mot prédit
<i>ferrostage</i>	<i>ferroviaire</i>	<i>industrielle</i>	<i>industrielle</i>
<i>pilliard</i>	<i>milliard</i>	<i>vontants</i>	<i>votant</i>
<i>administ</i>	<i>administratif</i>	<i>cherac</i>	<i>Chirac</i>
<i>servera</i>	<i>servira</i>	<i>yvelines</i>	<i>Yvelines</i>
<i>debat</i>	<i>débat</i>	<i>leius</i>	<i>leurs</i>

Les formes inconnues qui n’ont pas été corrigées sont celles pour lesquelles la similarité avec la forme prédite n’atteint pas le seuil, et aussi les toponymes et les sigles souvent mal prédits parce que le modèle de langue, même fine-tuné, n’a pas une connaissance exhaustive de tous les éléments référentiels du corpus.

4. Correction des mots inconnus sur l’ensemble du corpus

La méthode validée précédemment a été appliquée à l’ensemble du corpus, avec les mêmes paramètres. La sélection des contributions candidates à la correction constitue un enjeu et est précisée en 4.1. ; les résultats de la prédiction sont présentés en 4.2.

4.1. Sélection des contributions à corriger

La sélection des contributions à corriger repose sur un compromis entre l’intérêt de corriger le plus de contributions possibles, et la précision des corrections prédites par la fonction MLM. Cette précision est fondée sur le contexte (ici, l’unité de contexte est la phrase) du mot masqué (c’est-à-dire la forme inconnue). La prédiction est meilleure lorsque le contexte est la contribution tout entière (3.2.).

Cependant, certaines contributions peuvent contenir plusieurs mots inconnus et plusieurs mentions *ILLISIBLE*, ce qui obère la qualité de leur vecteur de contexte. La question se pose alors de définir des critères de sélection des contributions à corriger qui permettent de garantir la qualité du contexte de prédiction du mot inconnu.

Les contributions qui contiennent une seule forme inconnue et aucune mention *ILLISIBLE* sont des candidates évidentes. Cependant, elles ne couvrent que 76% des contributions contenant une forme inconnue. Aussi, ont été ajoutées à la sélection les contributions contenant une seule forme inconnue et une ou plusieurs mentions *ILLISIBLE*. En prétraitement, cette mention a été supprimée de la contribution pour ne pas en altérer le vecteur de contexte. En effet, toutes les mentions auraient alors été considérées comme des occurrences de la même forme – ce qui n’est évidemment pas le cas – avec un vecteur de contexte dérivé de cette même forme.

4.2. Résultats de la prédiction des mots inconnus

La table 5 récapitule la répartition des contributions en fonction de leurs nombres de mots inconnus et de mentions *ILLISIBLE*. Parmi les 38 106 contributions qui ne contiennent qu’une seule forme inconnue, le taux de correction le plus élevé (41%) concerne celles qui n’ont pas d’autre forme non analysable. Plusieurs hypothèses peuvent expliquer ce taux relativement faible : le champ lexical et les associations sémantiques du *cCC* sont particulièrement originaux, ou l’océrisation éloigne sensiblement les formes originales de la transcription, ou les erreurs d’océrisation ne se résument pas à des transformations de lettres. D’autre part, le taux

Table 5: Répartition des contributions corrigées par rapport aux nombres de mots inconnus et de mentions *ILLISIBLE*

#contri- butions	#contributions sans forme inconnue	#contributions contenant au moins 1 forme inconnue			
225 224	147 260	77 964			
		1 forme inconnue			n formes inconnues
		38 106			39 858
		0 <i>ILLISIBLE</i>	1 <i>ILLISIBLE</i>	n <i>ILLISIBLE</i>	
		29 018	4 014	5 074	
		corrigées	corrigées	corrigées	corrigées
		11 912 (41%)	1 585 (39%)	1 938 (38%)	0
Avec modèle CamemBERT-base affiné sur <i>cPF</i>					
		6 689 (23%)	893 (22%)	1 099 (21%)	0

de correction décroît avec l’augmentation du nombre de mots inutilisables du contexte, ce qui est cohérent avec la perte de précision globale du contexte de la proposition due à cette augmentation. De manière corollaire, ce résultat conforte l’idée que le contexte sémantique modélisé par les vecteurs de contexte constitue une base solide pour la prédiction des formes inconnues.

Le ratio du nombre d’occurrences inconnues par rapport au nombre total de mots (dernière colonne de la table 6) montre l’intérêt de la méthode proposée. Initialement à 8 % (V0), il s’améliore avec les corrections sans contexte (V2), puis avec celles en contexte (V3). Après la correction des mots inconnus en contexte, il passe à 5 %. Cette amélioration permet de rapprocher le taux de mots non analysables dans *cCC* de celui de *cPF* (6 %) qui, nativement numérique, n’a pas subi d’océrisation. Cependant, le réemploi moyen des formes inconnues (ratio [3]/[2]) varie du simple au triple entre les deux corpus. Une explication pourrait être que les formes inconnues de *cPF* sont plutôt des néologismes liés au contexte lexical des questions qui pourraient disparaître en élargissant la couverture des dictionnaires, alors que celles du *cCC* seraient des coquilles, majoritairement dues à l’océrisation, qu’il faudrait corriger.

5. Discussion et perspectives

Les résultats de la méthode proposée donnent matière à discussion sur plusieurs points : l’amélioration du modèle de langue (5.1., les formes inconnues dues à des segmentations fautives (5.2.), et la sélection des contributions à corriger (5.3.). Des conclusions sont proposées en fin de section.

5.1. Amélioration du modèle de langue

L’hypothèse a été formulée qu’un affinage du modèle adapté au contexte (le GDN) et sans erreur d’océrisation fournirait des vecteurs de contexte de meilleure qualité. Un test a donc été effectué en utilisant ceux fournis par le même modèle de langue (CamemBERT-base) mais affiné sur *cPF*. Les taux de corrections réalisées (table 5) sont dans toutes les situations inférieurs à ceux obtenus avec le modèle affiné sur *cCC*. L’interprétation serait que les thèmes abordés dans les Cahiers citoyens ainsi que les idiosyncrasies de leurs contributeur·rice·s sont différents de ceux de la plate-forme, en particulier parce que la participation aux premiers est plutôt rurale, et au second serait plus urbaine et plus jeune.

Table 6: Description numérique des différentes version du *cCC* et du (*cPF*). Le décompte a été réalisé à l’aide de l’outil Unitex Paumier et al. (2021) ; un token correspond à un élément unitaire du texte : mot, nombre (formé d’un ou plusieurs chiffres) et séparateurs dont les signes de ponctuation (le symbole # vaut pour "nombre de").

versions	#tokens	mots simples		formes inconnues			ratio
		#formes	#occrr. [1]	#formes [2]	#occrr. [3]	[3]/[2]	
V0			50 080 638		396 574		0,008
V2	114 829 705	455 503	49 885 072	144 608	286 068	1,96	0,006
V3	114 829 705	455 503	49 885 072	138 207	271 207	1,96	0,005
<i>cPF</i>	308 157 080	366 500	138 768 691	132 451	813 614	6,14	0,006

5.2. Segmentation en mots fautive

L’observation en contexte des mots inconnus non remplacés montre que l’océrisation peut transformer des lettres mais aussi la segmentation en mots. Des séparateurs peuvent avoir été ignorés conduisant à concaténer deux mots en une seule forme : *àdormir* (au lieu d’*à dormir*), en particulier dans les noms des contributeur-*rice-s* : *gillesetmonique*, *mariedupont*. Un signe de ponctuation peut être transformé en lettre comme dans *hollywoodiensivégétaliser* (*hollywoodiens : végétaliser*) ou la boucle d’une lettre en séparateur comme dans *observ’ation* (*observation*). A l’inverse, l’espacement entre lettres a pu être interprété comme un espacement entre mots ; par exemple, *annuel* s’est transformé en *an nuel* où *nuel* est une forme inconnue⁴. L’identification de ces mots inconnus par erreur de segmentation reste difficile – en outre, l’erreur d’océrisation a pu conduire à produire deux mots valides ou deux mots inconnus. Un indice pourrait être la présence d’une lettre majuscule à l’intérieur d’un mot quand l’initiale est minuscule, mais il ne couvre que peu de cas.

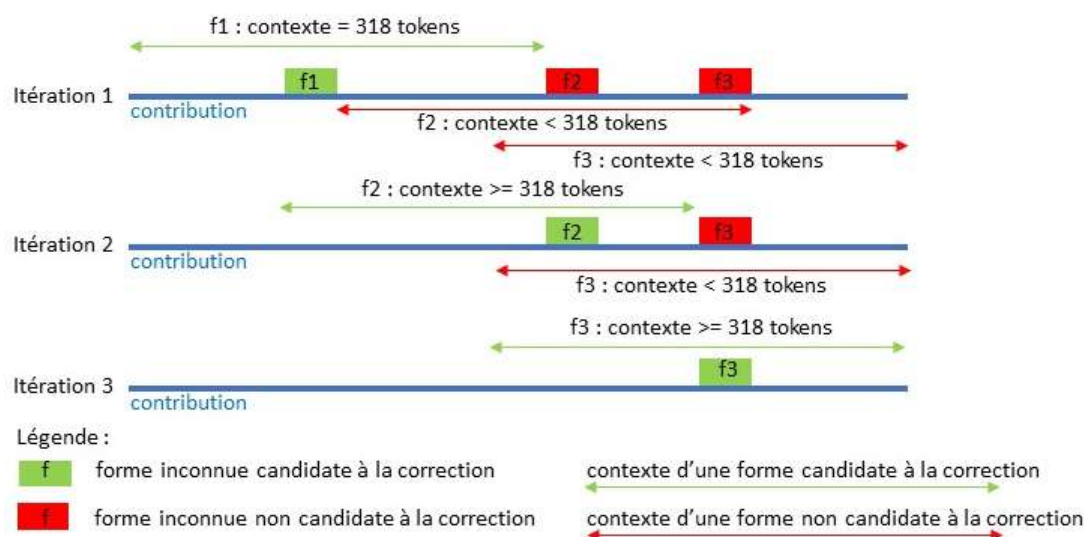
5.3. Définition d’un nouveau procédé de sélection des contributions à corriger

Le nombre de formes inconnues pourrait encore être diminué : en modifiant la sélection des contributions candidates à la correction, et en réitérant la prédiction afin de tirer profit de l’amélioration du contexte due à la prédiction de formes précédemment non analysables.

Augmentation de la sélection Il s’agit d’augmenter le nombre de contributions à corriger sans que le vecteur de contexte du mot à prédire y soit dégradé. La fonction MLM s’exerce dans une fenêtre d’attention limitée à 512 tokens. Cependant, la taille moyenne d’une contribution contenant une seule forme inconnue est de 318 tokens (au sens de CamemBERT). Ce n’est donc pas la limitation du modèle qui contraint le contexte de la prédiction mais la taille même de la contribution. En conséquence, la correction pourrait s’appliquer à des contributions contenant plusieurs mots inconnus mais dont au moins un peut disposer d’un contexte de caractéristiques comparables à celui des autres formes inconnues du corpus – c’est-à-dire d’au moins 318 tokens et sans forme inconnue ni mention *ILLISIBLE*.

⁴ Cette remarque a conduit à éliminer des contributions à corriger celles dont la seule forme inconnue était de longueur inférieure à 4 lettres, d’une part parce que la segmentation fautive d’un mot conduit à créer deux mots plus courts, d’autre part parce que les mots courts sont souvent des mots grammaticaux ou des auxiliaires verbaux dont le poids dans le vecteur de contexte de la contribution peut plus facilement être ignoré que celui des noms ou des verbes.

Figure 1: Evolution du nombre de formes à corriger par des itérations successives



Réitération de la méthode Un mode de sélection itératif (comme le montre le schéma 1) devrait permettre d'obtenir après chaque itération de nouvelles contributions candidates à la correction. En effet, si une contribution contient plus d'une forme inconnue, elle n'est pas sélectionnable par le processus décrit en 4.1. Cependant, si une des formes inconnues de la contribution dispose d'un contexte suffisant (f1 dans le schéma), elle pourrait être candidate à la correction, puis corrigée, par une première itération. Elle disparaîtrait alors en tant que forme inconnue, rendant ainsi la contribution candidate pour la correction de f2 dans l'itération suivante (Itération 2). La correction des différentes formes inconnues pourrait ainsi être itérée (Itération 3 pour f3). Enfin, même si la méthode de correction des mentions *ILLISIBLE* est différente de celle concernant les formes inconnues, le procédé de sélection des contributions candidates pourrait s'appliquer de la même manière pour ces deux types de correction.

La correction des formes inconnues du *cCC* s'intègre à l'objectif global d'analyse de ce corpus qui vise à repérer les thématiques abordées dans les contributions. Leur identification repose sur des méthodes de clustering des phrases du corpus et l'hypothèse est que le clustering sera de meilleure qualité si le nombre de mots non analysables est plus faible. Cependant, une interrogation demeure quant au principe de correction des formes inconnues. L'utilisation d'un LLM pour la prédiction revient à remplacer une forme inconnue, dont le fait qu'elle provienne d'une erreur d'océrisation ou de la volonté du contributeur-riche n'est pas tranché, par le mot le plus probable selon l'entraînement préalable sur des corpus non maîtrisés, voire non identifiés. La question se pose alors de la légitimité de ce remplacement ; quand il est systématique, il revient à renforcer les associations majoritaires, et donc à affaiblir encore plus les associations quantitativement peu présentes.

Remerciements

Ces travaux ont bénéficié de ressources de calcul en IA et de stockage à l'IDRIS au travers de l'allocation de ressources 2025-AD011015849 attribuée par GENCI sur la partition A100 du calculateur Jean Zay.

References

Authors (2024). Title hidden for anonymity.

Authors (2025). Title hidden for anonymity.

Baciocchi, S., Benoit-Lévy, R., Castanié, S., Cerutti, S., Sanchez, P.-L., and Schijman, E. (2024). Déchiffrer les cahiers de doléances: source, contextes et propositions d'économie politique (département de la Somme, 2018-2019). In *Annales. Histoire, sciences sociales*, volume 79, pages 7–56. Éditions de l'EHESS.

Bendinelli, M. (2023). Sens et matérialités d'un cahier citoyen: le cas de la ville de Dole (jura). *Mots. Les langages du politique*, 131(1):145–169.

Chabin, M.-A. and Vasseur, É. (2021). Exploitation des contenus versus analyse des documents: exemple des cahiers citoyens 2019. *La Gazette des archives*, 262(2):103–115.

Dominguès, C. and Jolivet, L. (2024). Analyse textométrique et spatialisée des Cahiers citoyens. In Dister, A. and Longrée, D., editors, *JADT 2024 Mots comptés, textes déchiffrés. Tome 1*, pages 309–318, Bruxelles, Belgique. Presses universitaires de Louvain.

Duguet, R. (2018). Les Cahiers de Doléances. *Gilets Jaunes: Des clés pour comprendre*.

Jolivet, L., Dominguès, C., Monnier, M., and Ploux, S. (2024). Construction de la notion de proximité : le thème de l'éolien dans les Cahiers citoyens. *GéoProximitéS*, 4.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2019). CamemBERT: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Noguera, S. (2023). Les cahiers citoyens et de doléances girondins (2018-2019): sociogenèse d'un dispositif pluriel. *7èmes Journées doctorales sur les expérimentations démocratiques et la démocratie participative, Groupement d'intérêt scientifique Démocratie & Participation/Université de La Rochelle, La Rochelle*, pages 23–24.

Panckhurst, R. (2009). Short Message Service (SMS) : typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin*, pages 33–52. Université Paul-Valéry Montpellier 3.

Paumier, S., Flury, W., Guenther, F., Laporte, E., Malchok, F., Marschner, C., Martineau, C., Martínez, C., Maurel, D., Nagel, S., Neme, A., Petit, M., Stiehler, J., and Vollant, G. (2021). UNITEX 3.3 User Manual. working paper or preprint.

Pengam, M. (2024). Les cahiers citoyens du grand débat national (2019) : d'un geste présidentiel dépolitisant à une (re) politisation citoyenne. *Mots. Les langages du politique*, 134(1):119–136.

Ploux, S. and Dominguès, C. (2025). L'immigration dans les Cahiers citoyens. *Plein droit*, 145(2):38–43.

Ploux, S., Genay, M., and Ploux-Chillès, L. (2021). Les mots du Grand débat national : les réseaux lexicaux des contributions déposées sur trois plateformes. *Humanités numériques*, 4.

Pottier, J.-M. (2025). Des gilets jaunes aux Cahiers citoyens. In *Les Grands Dossiers des Sciences Humaines*, volume 78, pages 26–27. Éditions Sciences Humaines.

Séjour à l'étranger et marqueurs discursifs en FLE : étude sur corpus d'apprenants japonophones¹

Kaori Sugiyama

Seinan Gakuin University – sugiyama@seinan-gakuin.jp

Abstract

This study examines inferred developmental patterns of discourse markers (DMs) in the oral production of Japanese-speaking learners of French, using contrastive interlanguage analysis (CIA) as a heuristic tool in a cross-sectional design. IPFC-J learners are grouped by length of study abroad (0–6 months, 7–12 months, over 12 months) and compared with the Aix reference corpus, treated as a Reference Language Variety rather than as a native-speaker norm. Log-likelihood, False Discovery Rate (FDR) correction and log ratio identify 11 DMs with significant and substantial differences in at least one adjacent comparison: *donc*, *d'abord*, *enfin*, *c'est-à-dire*, *puis*, *en fait*, *voilà*, *ben*, *tu vois*, *quoi* and *hein*. These items are described through four inferred configurations: non-convergent overuse, apparent or punctual frequency rapprochement, non-convergent underuse with partial diffusion, and strongly marked interpersonal underuse. Combining frequency, speaker-user rate, within-group concentration and intra-utterance position shows that study-abroad length does not lead to linear convergence toward the Aix profile. Sequential and ideational DMs appear relatively accessible but may remain overused or only apparently convergent, whereas strongly interpersonal DMs show limited diffusion even after longer stays. The pattern is compatible with contact-dependent accounts of DM acquisition (Sankoff et al., 1997; Deng, 2022).

Keywords: discourse markers, study abroad, contrastive interlanguage analysis, learner corpus, intra-utterance position

Résumé

Cette étude analyse les patrons développementaux inférés des marqueurs discursifs (MD) dans la production orale d'apprenants japonophones en français, en mobilisant l'analyse contrastive d'interlangue (ACI ; Contrastive Interlanguage Analysis, CIA) comme outil heuristique dans un design transversal. Les apprenants du corpus IPFC-J sont classés selon la durée du séjour à l'étranger (0–6 mois, 7–12 mois, plus de 12 mois) et comparés au corpus de référence Aix, considéré comme une variété de langue de référence (Reference Language Variety, RLV) plutôt que comme une norme native. La log-vraisemblance, la correction du taux de fausses découvertes (False Discovery Rate, FDR) et le *log ratio* identifient 11 MD présentant des différences statistiquement significatives et substantielles dans au moins une comparaison adjacente : *donc*, *d'abord*, *enfin*, *c'est-à-dire*, *puis*, *en fait*, *voilà*, *ben*, *tu vois*, *quoi* et *hein*. Ces items sont décrits selon quatre configurations inférées : suremploi non convergent, rapprochement fréquentiel apparent ou ponctuel, sous-emploi non convergent avec diffusion partielle, et sous-emploi interpersonnel fortement marqué. Le croisement de la fréquence, du taux de locuteurs utilisateurs, de la concentration intra-groupe et de la position intra-énoncé montre que la durée de séjour ne conduit pas à une convergence linéaire vers le profil d'Aix. Les MD séquentiels et idéationnels semblent relativement accessibles, mais peuvent rester suremployés ou seulement apparemment convergents, tandis que les MD fortement interpersonnels demeurent peu diffusés même après un séjour long. Ce patron est compatible avec les modèles d'acquisition contact-dépendante des MD (Sankoff et al., 1997 ; Deng, 2022).

Mots clés : marqueurs discursifs, séjour à l'étranger, analyse contrastive d'interlangue, corpus d'apprenants, position intra-énoncé

¹ Cette recherche a bénéficié du soutien de la JSPS KAKENHI, subventions nos JP25K00472 et JP24K00093.

1. Introduction

De nombreuses recherches ont rendu compte de l'impact positif de l'expérience d'études à l'étranger sur le développement de la langue seconde. Ces dernières années, l'intérêt s'est toutefois déplacé d'une question binaire, « le séjour à l'étranger est-il ou non efficace ? », vers une question plus précise : « quelles formes linguistiques se développent, dans quelle mesure et grâce à quel type de séjour ? » Or, les marqueurs discursifs (désormais MD) ne participent pas directement au contenu propositionnel ; ils sont donc difficiles à saisir dans le cadre de l'analyse des erreurs ou de l'exactitude grammaticale, et les études quantitatives comparatives restent peu nombreuses dans ce domaine.

La présente étude classe les données d'apprenants japonophones du corpus IPFC-J selon la durée de séjour et décrit, dans une perspective transversale, les patrons développementaux inférés à partir d'un ensemble de MD mis en évidence par l'Analyse Contrastive de l'Interlangue (Contrastive Interlanguage Analysis ; CIA ; Granger, 2015). Les courbes d'estimation de densité par noyau (EDN) complètent l'analyse fréquentielle en visualisant les tendances continues de distribution positionnelle intra-énoncé.

2. Travaux antérieurs

2.1. Définition des MD du français et position intra-énoncé

Les MD se définissent comme des formes linguistiques morphologiquement invariables et syntaxiquement facultatives, qui ne participent pas directement au contenu propositionnel et qui assurent la structuration du discours ainsi que la régulation interactionnelle entre locuteurs (Hansen, 1998 ; Fraser, 1999 ; Crible & Degand, 2019). Crible & Degand (2019) ont affiné ce cadre en décrivant les informations sémantico-pragmatiques selon deux couches indépendantes : le « domaine » (idéationnel, rhétorique, séquentiel et interpersonnel) et la « fonction » (addition, cause, etc., 15 types au total).

Par ailleurs, la position intra-énoncé constitue un indice distributionnel important pour l'interprétation fonctionnelle des MD. Les travaux sur la périphérie gauche et la périphérie droite en français parlé suggèrent que la périphérie gauche (PG) tend à accueillir des marqueurs assurant le cadrage du discours et la connexion logique, tandis que la périphérie droite (PD) est fréquemment associée à des marqueurs de régulation interactionnelle ou de sollicitation de confirmation (Degand, 2014).

Concernant les marqueurs individuels, les travaux antérieurs ont décrit la multifonctionnalité pragmatique de *donc* (Hansen, 1997 ; Pellet, 2009), la fonction de recherche d'accord associée à *hein* en fin d'énoncé (Reali, 2011), le fonctionnement colloquatif de *ben* en combinaison avec d'autres marqueurs discursifs (Dostie, 2012), la fonction de clôture d'énoncé de *quoi* (Lefeuve, 2011) et la conventionnalisation de *tu vois* comme marqueur pragmatique intersubjectif (Bolly, 2009).

2.2. Acquisition des MD par les apprenants

Les MD sont rarement enseignés explicitement en classe et semblent dépendre fortement du contact direct avec la communauté des locuteurs natifs francophones (Sankoff et al., 1997). Deng (2022) a montré, auprès de 40 locuteurs sinophones de français L2 résidant à Paris (séjours d'un à plus de 7 ans), que seul le contact extracurriculaire avec les locuteurs natifs s'est révélé statistiquement significatif pour *bon*, *ben* et *fin*, la durée de séjour n'atteignant pas la significativité. En ce qui concerne les japonophones, des recherches fondées sur IPFC-J h

montrent un sous-emploi global des MD (Sugiyama, 2017), mais aussi des effets liés au séjour et aux usages quasi-discursifs (Sugiyama, 2023). L'étude de *bon* (Seimiya & Kawaguchi, 2025) montre en outre que des fréquences élevées peuvent masquer des différences de cooccurrence et de distribution fonctionnelle. Ces travaux invitent donc à dépasser le simple comptage global des MD et à comparer plusieurs marqueurs dans un même cadre d'analyse.

Les études qui comparent quantitativement plusieurs MD dans un même cadre en prenant la durée de séjour comme variable restent peu nombreuses ; la présente étude vise à combler ce vide. Sur la base de cette discussion, nous formulons les trois questions de recherche (QR) suivantes :

(QR1) Comment les mots à fonction de MD émergent-ils statistiquement dans l'analyse des mots-clés ?

(QR2) Quel patron développemental la fréquence d'emploi des MD présente-t-elle selon la durée du séjour à l'étranger, inféré à partir d'une comparaison transversale ?

(QR3) Observe-t-on une différence systématique dans les tendances développementales entre les marqueurs des domaines séquentiel et idéationnel, typiquement associés à la périphérie gauche, et les marqueurs du domaine interpersonnel, typiquement associés à la périphérie droite ?

3. Méthode

3.1. Corpus et données analysées

Les données d'apprenants sont issues de la partie conversation libre du corpus IPFC-J (Detey & Kawaguchi, 2008) et classées en trois groupes selon la durée de séjour (Tableau 1). Le corpus de référence est le corpus Aix, recueilli selon le même format de tâche, à savoir une conversation libre (Koga et al., 2011). L'étude adopte le cadre de l'analyse contrastive d'interlangue (Contrastive Interlanguage Analysis, CIA ; Granger, 1996, 2015) : le corpus Aix est considéré comme une variété de langue de référence (Reference Language Variety, RLV), comparable par le format de tâche, et non comme une norme générale du français natif. Les différences observées sont donc interprétées comme des contrastes entre corpus.

Tableau 1 : Présentation du corpus

Groupe	Locuteurs	Tokens	Durée de séjour	Abréviation
≤ un semestre	50	31427	0-6 mois	US
> un semestre à un an	77	53278	7-12 mois	UA
> un an	39	35326	plus de 12 mois	PUA
Aix (Natif)	72	429149	-	Aix

Le design est transversal : les groupes US/UA/PUA réunissent des individus distincts comparés synchroniquement. L'expression « patrons développementaux inférés » désigne donc une reconstruction inter-groupes selon la durée de séjour, cohérente avec des études comparables sur les MD et le séjour à l'étranger (Magliacane & Howard, 2019). Les groupes d'apprenants présentent un rapport tokens/locuteur croissant (US : 628 ; UA : 692 ; PUA : 906) ; cet effet est pris en compte au niveau agrégé par la normalisation pm (fréquences) et $c_i/(n-1)$ (positions ; voir §3.2).

3.2. Méthode d'analyse, identification des MD et position intra-énoncé

L'analyse repose sur trois comparaisons adjacentes : $US \rightarrow UA$, $UA \rightarrow PUA$ et $PUA \rightarrow Aix$. La log-vraisemblance (Dunning, 1993) a été calculée sur les fréquences d'occurrence, avec correction du taux de fausses découvertes (False Discovery Rate, FDR) par la méthode de Benjamini & Hochberg (1995) ($q < 0,05$). Plutôt que la spécificité de Lafon (1980, fondée sur la distribution hypergéométrique), plus classique dans la lexicométrie française, nous retenons la log-vraisemblance (fondée sur les lois bi- et multinomiales) pour sa combinaison naturelle avec le *log ratio* (Hardie, 2014), qui permet de dissocier la significativité statistique de l'ampleur de l'écart. Les fréquences ont été normalisées en occurrences par million (pm). Afin de ne retenir que des écarts statistiquement significatifs et substantiels, un seuil de *log ratio* a été ajouté (Hardie, 2014 ; $|\log ratio| \geq 1$, soit environ un facteur 2). Par exemple, sur Aix (plus de 400 000 tokens), un écart aussi faible que $pm = 1\,500$ contre $1\,520$ peut atteindre la significativité statistique sans porter d'effet linguistique substantiel ; le *log ratio* permet d'écarter ce type de cas. Les mots satisfaisant ces critères ont été retenus comme mots-clés.

Afin d'évaluer l'effet des locuteurs fortement contributeurs, une analyse de sensibilité de type *leave-one-out* a été appliquée aux cellules $MD \times$ groupe où un locuteur concentrait plus de 50 % des occurrences, avec un total d'au moins 10 occurrences. Les fréquences ont alors été recalculées après exclusion de ce locuteur dominant.

Les mots mis en évidence statistiquement ont ensuite été vérifiés a posteriori au regard des critères de Crible & Degand (2019). Toutes les occurrences extraites par requêtes lexicales ont fait l'objet d'une vérification manuelle afin de distinguer les emplois discursifs des emplois non discursifs, notamment les emplois pronominaux ou interrogatifs de *quoi* et les emplois de succession temporelle de *puis*. Les marqueurs polylexicaux (*tu vois*, *c'est-à-dire*, *en fait*) ont été identifiés et comptabilisés comme unités lexicales uniques ; leur position est définie par celle du premier token.

L'énoncé est défini comme un tour de parole non interrompu par un changement de locuteur, segmenté secondairement aux frontières signalées par les conventions de transcription. La position d'un token i est définie par $c_i/(n - 1) \in [0, 1]$, où c_i est l'indice du token i (0-indexé) et n le nombre total de tokens de l'énoncé considéré. Cette normalisation permet de comparer les positions indépendamment de la longueur des énoncés.

La distribution des positions est visualisée au moyen de courbes d'estimation de densité par noyau (EDN), calculées avec un noyau gaussien et une largeur de bande déterminée selon la règle de Scott (Scott, 1992). Les courbes EDN étant sensibles aux faibles effectifs et aux valeurs aberrantes (Silverman, 1986), elles sont interprétées comme un complément descriptif à l'analyse fréquentielle.

4. Résultats et discussion

Le nombre total de mots-clés identifiés dans au moins une des trois comparaisons adjacentes s'élève à 286 ($US-UA$: 64, $UA-PUA$: 63, $PUA-Aix$: 216). Parmi eux, 11 items à fonction de MD ont été recensés (Tableau 2).

Les 11 items mis en évidence par la CIA présentent des patrons hétérogènes de fréquence, de diffusion interindividuelle et de distribution positionnelle. Leur organisation empirique repose sur trois dimensions descriptives : (i) le patron fréquentiel et diffusionnel inféré à partir de la CIA, (ii) le domaine sémantico-pragmatique, retenu ici sans annotation fine des fonctions en contexte (Crible & Degand, 2019), et (iii) le profil positionnel intra-énoncé, interprété en lien

avec l'opposition entre périphérie gauche et périphérie droite (Degand, 2014). Les quatre regroupements présentés ci-dessous ne résultent pas d'un croisement systématique a priori de ces dimensions, qui produirait un nombre plus élevé de cases théoriques, mais des configurations effectivement observées dans les données, sans visée de taxonomie exhaustive. Ils ne constituent donc pas une typologie catégorielle stricte des marqueurs, mais une mise en relation progressive entre domaines fonctionnels, tendances positionnelles et patrons acquisitionnels. Ils sont présentés selon une gradation allant des emplois de balisage explicite, plus accessibles et susceptibles de donner lieu à un suremploi, vers les emplois interpersonnels en périphérie droite, plus faiblement diffusés chez les apprenants.

Tableau 2 : Fréquences (pm), taux de locuteurs utilisateurs (n %), concentration maximale intra-groupe, pm LOO et patrons inférés des 11 MD.²

lemme	pm US	pm UA	pm PUA	pm Aix	max conc.	pm LOO	patron inféré
donc	8559.5 (32%)	7207.1 (36%)	11693.6 (31%)	5044.6 (99%)	PUA 27 %	—	suremploi non convergent
d'abord	222.8 (8%)	638.3 (14%)	424.7 (15%)	83.9 (33%)	US 43 %	—	suremploi non convergent
enfin	127.3 (6%)	638.3 (4%)	5209.0 (21%)	3728.8 (96%)	PUA 79 %	UA 244 ; PUA 1104	rapprochement ponctuel
c'est-à-dire	31.8 (2%)	619.5 (5%)	169.9 (8%)	221.4 (38%)	UA 67 %	UA 207	rapprochement ponctuel
puis	413.7 (10%)	1577.0 (14%)	877.8 (15%)	1407.6 (96%)	UA 62 %	US 159 ; UA 601	rapprochement ponctuel
en fait	381.8 (12%)	2590.6 (22%)	1047.6 (21%)	2828.8 (93%)	PUA 60 %	PUA 425	rapprochement ponctuel
voilà	95.5 (6%)	488.1 (8%)	2350.5 (21%)	3413.6 (100%)	UA 31 %	—	sous-emploi partiel
ben	95.5 (4%)	300.3 (8%)	906.0 (21%)	3439.8 (69%)	US 67 %	—	sous-emploi partiel
tu vois	0.0 (0%)	56.3 (1%)	28.3 (3%)	1821.9 (81%)	—	—	sous-emploi marqué
quoi	0.0 (0%)	262.8 (21%)	510.0 (13%)	4129.1 (96%)	PUA 67 %	PUA 170	sous-emploi marqué
hein	31.8 (2%)	450.5 (12%)	594.5 (18%)	3572.5 (92%)	PUA 38 %	—	sous-emploi marqué

(1) Pôle de balisage explicite : idéationnel/séquentiel : suremploi non convergent (*donc*, *d'abord*)

² (n%) = taux de locuteurs utilisateurs (US/50, UA/77, PUA/39, Aix/72). Max conc. = part des occurrences du locuteur le plus actif. pm LOO = fréquence recalculée après exclusion du locuteur dominant, lorsque celui-ci concentre plus de 50 % des occurrences et que le total est ≥ 10 ; — = non applicable.

Donc et *d'abord* fonctionnent principalement dans les domaines séquentiel et idéationnel, en marquant des relations logiques, inférentielles ou temporelles entre contenus propositionnels (Crible & Degand, 2019). Lorsqu'ils sont mobilisés comme marqueurs de balisage explicite en périphérie gauche (PG), où se concentrent les fonctions d'organisation discursive et de cadrage informationnel (Degand, 2014), ces emplois semblent relativement accessibles aux apprenants ; dans les données, cette accessibilité se traduit toutefois par un suremploi durable plutôt que par un ajustement progressif au profil d'Aix.

Donc présente une fréquence très élevée dès US (pm = 8559,5), maintenue en PUA (pm = 11693,6), contre 5044,6 pm dans Aix. Sa distribution est surtout initiale, ce qui suggère un emploi de cadrage ou de connexion explicite plutôt qu'une stabilisation des usages inférentiels plus variés décrits dans la littérature (Hansen, 1997 ; Pellet, 2009). *D'abord* est également suremployé, notamment en UA (pm = 638,3, contre 83,9 dans Aix). Sur le plan positionnel, US et Aix partagent une concentration principale dans la zone médiane, tandis que UA et PUA présentent des distributions plus étalées. Il n'apparaît donc pas de convergence positionnelle progressive selon la durée de séjour.

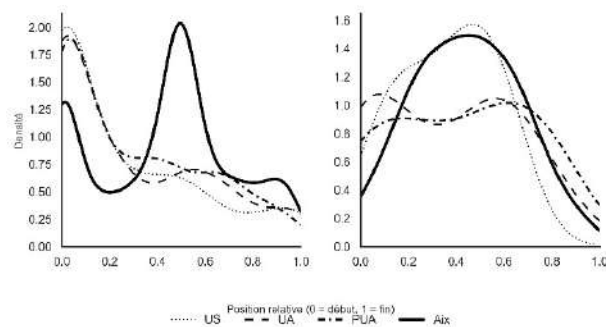


Fig. 1. Distribution positionnelle EDN de *donc* (gauche) et *d'abord* (droite).

(2) Zone intermédiaire I : polyfonctionnalité sans spécialisation positionnelle : rapprochement fréquentiel apparent ou ponctuel (*puis*, *c'est-à-dire*, *enfin*, *en fait*)

Enfin, *c'est-à-dire*, *puis* et *en fait* relèvent de domaines séquentiel et/ou rhétorique, dans la mesure où ils participent à l'organisation du discours, à la reformulation ou à l'ajustement du contenu énoncé (Crible & Degand, 2019 ; Eshkol-Taravella & Grabar, 2014 ; Ferré, 2011). Ne présentant pas de spécialisation positionnelle nette ni en PG ni en PD, ils occupent une zone intermédiaire entre les deux pôles, ce qui se reflète dans un rapprochement fréquentiel avec Aix à un stade donné, sans diffusion large ni stabilisation durable dans les groupes apprenants.

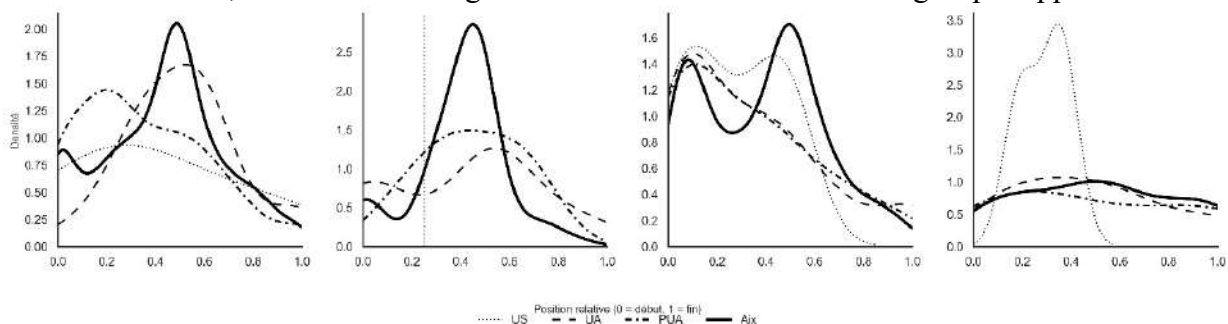


Fig. 2 : Distribution positionnelle EDN d'*enfin*, *c'est-à-dire*, *puis* et *en fait* (de gauche à droite).

Sur le plan fréquentiel, *enfin* dépasse Aix en PUA (pm = 5209,0 contre 3728,8), mais le pm LOO chute à 1104, ce qui indique un effet de concentration individuelle. *C'est-à-dire* dépasse

Aix en UA (pm = 619,5 contre 221,4), avec un faible taux de locuteurs utilisateurs (5 %) et une concentration maximale élevée (67 %). *Puis* atteint en UA une fréquence proche de celle d'Aix (pm = 1577,0 contre 1407,6), mais le taux de locuteurs utilisateurs reste faible (14 % en UA, contre 96 % dans Aix) et le pm LOO diminue nettement (601). Quant à *en fait*, il se rapproche d'Aix en UA (pm = 2590,6 contre 2828,8), mais ce rapprochement n'est pas maintenu en PUA (pm = 1047,6). Ces quatre cas relèvent donc d'un rapprochement fréquentiel apparent ou ponctuel plutôt que d'une convergence stabilisée.

Sur le plan positionnel, ces rapprochements fréquentiels ne s'accompagnent pas d'une stabilisation nette des profils d'emploi. Pour *enfin*, *c'est-à-dire* et *puis*, les courbes apprenantes restent distinctes de celles d'Aix ou doivent être interprétées avec prudence en raison de la concentration individuelle. Pour *en fait*, les courbes suggèrent également un profil encore instable, d'autant que le rapprochement observé en UA ne se prolonge pas en PUA.

(3) Zone intermédiaire II : composante PG-interactive : sous-emploi avec diffusion partielle (*ben*, *voilà*)

Voilà et *ben* présentent dans Aix un profil positionnel marqué associé à des fonctions distinctes, mais leur trajectoire chez les apprenants est comparable : leur fréquence augmente, surtout en PUA, sans atteindre celle du corpus Aix, et leur diffusion semble se restreindre aux composantes les plus accessibles en périphérie gauche. Ce regroupement correspond donc à un sous-emploi non convergent avec diffusion partielle.

Sur le plan fréquentiel, *voilà* passe de 95,5 pm en US à 2350,5 pm en PUA, mais reste inférieur à Aix (3413,6 pm). Son taux de locuteurs utilisateurs demeure aussi faible en PUA (21 %, contre 100 % dans Aix). *Ben* suit une progression comparable, de 95,5 pm en US à 906,0 pm en PUA, mais l'écart avec Aix reste important (3439,8 pm), malgré une hausse du taux de locuteurs utilisateurs (21 % en PUA, contre 69 % dans Aix).

Sur le plan positionnel, *voilà* présente dans Aix un profil pluriel, avec des emplois initiaux, médians et finaux. Chez les apprenants, les emplois initiaux de confirmation ou d'acceptation, proches du monitoring interactionnel, semblent plus accessibles, tandis que la clôture de séquence thématique en position médiane (Lee et al., 2019) reste peu stabilisée, même en PUA. *Ben*, de son côté, présente dans Aix une forte concentration initiale, notamment autour de la position 0, qui n'est pas reproduite avec la même netteté par les apprenants. Ces deux cas montrent ainsi une diffusion partielle, limitée à un sous-ensemble des fonctions de PG attestées dans Aix.

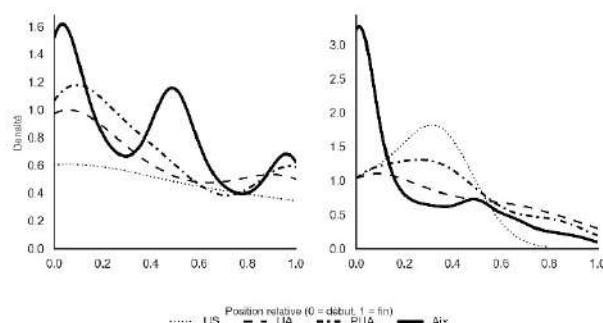


Fig. 3 : Distribution positionnelle EDN de *voilà* (gauche) et *ben* (droite).

(4) Pôle PD : interpersonnel : sous-emploi non convergent fortement marqué à faible diffusion (*tu vois*, *quoi*, *hein*)

Tu vois, *quoi* et *hein* relèvent principalement du domaine interpersonnel, fréquemment associé à la périphérie droite (PD), où se concentrent les fonctions intersubjectives, modales et interactionnelles (Degand, 2014) : marquage épistémique, régulation interactionnelle, recherche d'accord (Bolly, 2009 ; Lefeuve, 2011 ; Reali, 2011 ; Crible & Degand, 2019). Ils se caractérisent par un sous-emploi massif et une faible diffusion interindividuelle, même en PUA.

Sur le plan fréquentiel, l'écart avec Aix est particulièrement net. *Tu vois* est presque absent des données apprenantes (US : 0,0 ; UA : 56,3 ; PUA : 28,3), alors qu'il est bien attesté dans Aix (pm = 1821,9). *Quoi* reste également très inférieur à Aix en PUA (pm = 510,0 contre 4129,1). *Hein* demeure lui aussi peu fréquent en PUA (pm = 594,5 contre 3572,5 dans Aix). Les taux de locuteurs utilisateurs confirment cette faible diffusion : *tu vois* est utilisé par 3 % des locuteurs en PUA contre 81 % dans Aix, *quoi* par 13 % contre 96 %, et *hein* par 18 % contre 92 %.

Sur le plan positionnel, les distributions doivent être interprétées avec prudence en raison du faible nombre d'occurrences chez les apprenants. Dans Aix, *hein* et *quoi* présentent une concentration finale compatible avec leurs fonctions interactionnelles, tandis que *tu vois* est trop rare dans les groupes apprenants pour être interprété comme un marqueur pragmatique intersubjectif conventionnalisé au sens de Bolly (2009). Ce regroupement illustre ainsi la difficulté particulière des marqueurs PD–interpersonnels prototypiques.

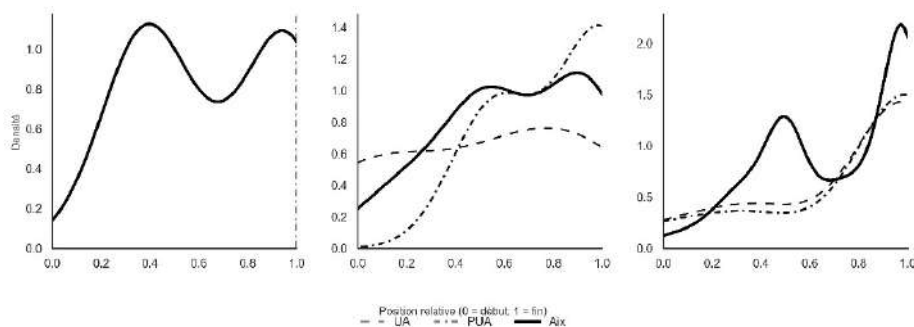


Fig. 4 : Distribution positionnelle EDN de *tu vois* (gauche), *quoi* (centre) et *hein* (droite).

5. Conclusion et perspectives

La CIA a identifié 11 MD présentant une différence significative dans au moins une comparaison adjacente parmi 286 mots-clés. Pour QR1, ces résultats montrent que les MD concernés relèvent de profils contrastés selon les groupes de séjour.

Pour QR2, les 11 MD ne progressent pas uniformément vers l'usage d'Aix avec l'allongement du séjour. Ils se répartissent le long d'une gradation entre un pôle de balisage explicite, principalement idéationnel/séquentiel, caractérisé par un suremploi non convergent (*donc*, *d'abord*), une zone de polyfonctionnalité sans spécialisation positionnelle, associée à un rapprochement apparent ou ponctuel (*enfin*, *c'est-à-dire*, *puis*, *en fait*), une zone de diffusion partielle limitée à la composante PG–interactive (*voilà*, *ben*), et un pôle PD–interpersonnel fortement sous-employé à faible diffusion (*tu vois*, *quoi*, *hein*). L'analyse leave-one-out confirme par ailleurs que certains rapprochements fréquentiels sont influencés par la concentration chez un petit nombre de locuteurs.

Pour QR3, cette gradation invite à nuancer l'hypothèse d'une progression linéaire liée à la durée de séjour. Elle s'articule entre domaine fonctionnel, position préférentielle, fréquence et diffusion interindividuelle : les marqueurs de balisage explicite, relativement accessibles,

donnent lieu à un suremploi durable ou à un rapprochement seulement apparent, tandis que les marqueurs interpersonnels prototypiques restent peu diffusés. Ce résultat est compatible avec l'hypothèse d'une acquisition des MD fortement dépendante du contact interactionnel (Sankoff et al., 1997 ; Deng, 2022).

Le dispositif transversal ne permet pas d'observer des trajectoires individuelles, et la durée de séjour fonctionne ici comme une approximation des occasions de contact, non comme une mesure directe de leur qualité. De plus, IPFC-J ne permet pas d'isoler les effets de cohorte, liés par exemple au programme d'études, à l'enseignant ou à la tâche, susceptibles d'expliquer certains pics observés à un stade donné, notamment en UA. Certaines propriétés générales de la production orale, notamment la longueur des énoncés, pourraient également être prises en compte dans de futures analyses, sans être interprétées comme des biais propres au dispositif. Ces limites ouvrent plusieurs perspectives : une étude longitudinale permettant de distinguer les effets de la durée de séjour, du niveau de compétence et de la qualité du contact, ainsi que l'intégration du contexte syntaxique et prosodique aux courbes EDN (Lee et al., 2019 ; Ferré, 2011).

Bibliographie

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.
- Bolly, C. (2009). Constructionnalisation et structure informationnelle : Quand la grammaticalisation ne suffit pas pour expliquer *tu vois*. *Linx*, 61, 103–130.
- Crible, L., & Degand, L. (2019). Domains and functions: A two-dimensional account of discourse markers. *Discours*, 24.
- Degand, L. (2014). “So very fast then”: Discourse markers at left and right periphery in spoken French. In K. Beeching & U. Detges (Eds.), *Discourse functions at the left and right periphery: Crosslinguistic investigations of language use and language change* (pp. 151–178). Brill.
- Deng, D. (2022). *Bon ben en fin fin* in non-native speech: The case of Chinese L1 speakers in Paris. *SHS Web of Conferences*, 138, Article 12002.
- Detey, S., & Kawaguchi, Y. (2008, December 11–13). Interphonologie du Français Contemporain (IPFC) : Récolte automatisée des données et apprenants japonais [Conference presentation]. *Journées PFC : Phonologie du français contemporain : variation, interfaces, cognition*, Paris, France.
- Dostie, G. (2012). *Ben en tant que* collocatif discursif. *Travaux de linguistique*, 65(2), 105–122.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.
- Eshkol-Taravella, I., & Grabar, N. (2014). Paraphrastic reformulations in spoken corpora. In A. Przepiórkowski & M. Ogrodniczuk (Eds.), *Advances in natural language processing: PolTAL 2014* (Lecture Notes in Computer Science, Vol. 8686, pp. 425–437). Springer.
- Ferré, G. (2011). Multimodal analysis of discourse markers *donc*, *alors*, and *en fait* in conversational French. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 671–674).
- Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, 31(7), 931–952.

- Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies* (pp. 37–51). Lund University Press.
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24.
- Hansen, M.-B. M. (1997). *Alors* and *donc* in spoken French: A reanalysis. *Journal of Pragmatics*, 28(2), 153–187.
- Hansen, M.-B. M. (1998). The semantic status of discourse markers. *Lingua*, 104(3–4), 235–260.
- Hardie, A. (2014). *Log ratio: An informal introduction*. ESRC Centre for Corpus Approaches to Social Science. <http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/>
- Koga, K., Akihiro, H., & Kawaguchi, Y. (2011). Aix hanashi kotoba kōpasu purojekuto [Projet de corpus de français parlé d'Aix]. *Flambeau*, 37, 37–54.
- Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, 127–165.
- Lee, L., Bartkova, K., Jouvett, D., Dargnat, M., & Keromnes, Y. (2019). Can prosody meet pragmatics? Case of discourse particles in French. In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1525–1529). Australasian Speech Science and Technology Association.
- Lefevre, F. (2011). *Bon et quoi* à l'oral : Marqueurs d'ouverture et de fermeture d'unités syntaxiques à l'oral. *Linx*, 64–65, 223–240.
- Magliacane, A., & Howard, M. (2019). Discourse markers in L2 Irish English. In M. Howard (Ed.), *Study abroad, second language acquisition and interculturality* (pp. 277–308). Multilingual Matters.
- Pellet, S. (2009). The pragmatics of the French discourse markers *donc* and *alors*. In R. P. Leow, H. Campos, & D. Lardiere (Eds.), *Little words: Their history, phonology, syntax, semantics, pragmatics, and acquisition* (pp. 159–169). Georgetown University Press.
- Real, V. (2011). *Forms and functions of the French discourse particle hein in French mundane conversation: A conversation analysis perspective* [Doctoral dissertation, University of Illinois at Urbana-Champaign]. IDEALS.
- Sankoff, G., Thibault, P., Nagy, N., Blondeau, H., Fonollosa, M.-O., & Gagnon, L. (1997). Variation in the use of discourse markers in a language contact situation. *Language Variation and Change*, 9(2), 191–217.
- Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. Wiley.
- Seimiya, T., & Kawaguchi, Y. (2025). Nihonjin Furansugo gakushūsha no jiyū kaiwa ni okeru danwa hyōshiki *bon* no ichi kōsatsu [Une étude sur le marqueur discursif *bon* dans les conversations libres d'apprenants japonais du français]. *Studia romanica*, 58, 68–82.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall.
- Sugiyama, K. (2017). Différences dans l'utilisation des marqueurs discursifs : Analyse comparative entre apprenants et Français natifs. In H. Tyne (Ed.), *Le français en contextes : Approches didactiques, linguistiques et acquisitionnelles*. Presses universitaires de Perpignan.
- Sugiyama, K. (2023). Analyse lexicale des verbes utilisés à l'oral par les apprenants japonais du français selon leur durée de séjour en pays francophone. In *Actes du quatrième colloque international de l'ATPF : Les multiples facettes du français et son enseignement au 21e siècle : Plurilinguisme, pluriculturalisme et innovations* (pp. 332–357).

Le potentiel lexicographique des néologismes des TIC dans une perspective franco-slovaque : analyse par le biais de *Sketch Engine*

Lucia Ráčková

Université Matej Bel à Banská Bystrica, Slovaquie – lucia.rackova@umb.sk

Abstract

In this study, we analyze suffixed neologisms in the field of ICT in French and their Slovak equivalents. The analysis is based on contemporary web corpora (frTenTen23 and skTenTen23) processed using Sketch Engine. We combine quantitative and qualitative criteria to evaluate their usage. The goal is to determine their degree of lexicographic potential. This approach is relevant for evaluating neologisms that are likely to be included in dictionaries.

Keywords: CQL query, Concordance, Word Sketch, Gallicagram, suffixed neologisms, web language

Résumé

Dans cette étude, nous analysons les néologismes suffixés du domaine des TIC en français et leurs équivalents slovaques. L'analyse repose sur des corpus web contemporains (frTenTen23 et skTenTen23) exploités via *Sketch Engine*. Nous combinons des critères quantitatifs et qualitatifs pour évaluer leur usage. L'objectif est de déterminer leur degré de potentiel lexicographique. Cette approche est pertinente pour prendre en compte des néologismes susceptibles d'être intégrés dans les dictionnaires.

Mots clés : CQL query, Concordance, Word Sketch, Gallicagram, néologismes suffixés, langue web

1. Introduction

Ce travail s'inscrit dans le cadre d'une recherche en néologie contrastive centrée sur les substantifs français dérivés et leurs équivalents slovaques (Ráčková, 2026, à paraître). L'objectif est d'évaluer, à partir de critères observables, dans quelle mesure certaines unités peuvent prétendre à un traitement lexicographique. Le potentiel lexicographique désigne la capacité d'un néologisme à être intégré dans les dictionnaires, en fonction de son degré de stabilisation dans l'usage, de sa fréquence, de sa diffusion et de sa conformité aux normes linguistiques. Le point de départ est empirique : un néologisme à faible diffusion n'a pas le même statut qu'une unité largement attestée et stabilisée. Dans cette perspective, nous postulons que la fréquence constitue un critère majeur pour l'intégration des unités dans les dictionnaires.

2. Cadre théorique

Dans le contexte de la lexicographie française, on sait que *Le Petit Robert* et *Le Petit Larousse* intègrent chaque année environ 150 nouvelles entrées, qu'il s'agisse de néologismes formels

ou sémantiques (Campese, 2025). Pour qu'un néologisme puisse être intégré au dictionnaire, il doit, selon le *Larousse*, « passer l'épreuve du temps », estimée à environ trois ans (Moraru, 2023, p. 184). Nous étudions également les néologismes français d'un point de vue temporel à l'aide de l'outil *Gallicagram*, qui retrace les unités lexicales du XIXe siècle à 2025, offrant ainsi une base solide pour la recherche lexicographique synchrone et diachronique. Dans une perspective français-slovaque, il est à noter que selon le cadre conceptuel de Catford (1965) sur l'intraduisibilité linguistique, ces formes peuvent être considérées comme des expressions qui résistent à une traduction exacte.

La définition la plus courante, bien qu'insuffisante, du néologisme est liée à son absence dans le dictionnaire (Mudrochová, 2021 : 296). Comme nous considérons ce critère, contrairement au « sentiment de nouveauté », comme un critère objectif, nous l'appliquons, dans le prolongement de Sablayrolles (2019), également dans la présente étude. D'autres termes importants utilisés par Sablayrolles (2000) sont le « cycle de vie des unités lexicales » et les « candidats néologiques ». Le premier adopte une perspective diachronique en tenant compte de la stabilisation dans le système linguistique et de l'intégration possible en tant qu'entrée de dictionnaire. Quant aux candidats néologiques, il s'agit de mots innovants susceptibles de se diffuser dans l'usage et de devenir des néologismes stabilisés. Ils oscillent entre la norme et le registre. Les locuteurs, en particulier sur internet, ne respectent pas souvent la norme linguistique et leurs productions oscillent entre la norme, les règles du système, le registre et la codification éventuelle du terme.

3. Méthodologie

L'étude repose sur les corpus web contemporains *French Web 2023* (*frTenTen23*) et *Slovak Web 2023* (*skTenTen23*) comparables par leur mode de constitution et accessibles dans *Sketch Engine*. L'identification des néologismes repose sur deux critères fondamentaux : leur absence dans les sources lexicographiques de référence (pour le français *Trésor de la langue française informatisé – TLFi*, *Dictionnaire de l'Académie française – DAF*, *Le Robert Dico en ligne – LR* ; pour le slovaque le portail lexical *JÚLŠ – SP JÚLŠ*, qui regroupe plusieurs dictionnaires sur un seul site) et, parallèlement, de leur présence dans les vastes corpus web contemporains de la famille *TenTen* : *frTenTen23* pour le français et *skTenTen23* pour le slovaque et traités à l'aide des outils de *Sketch Engine*, notamment les requêtes CQL, les concordances et les profils de type Word Sketch. En ce qui concerne le français, la dimension diachronique est également étudiée à l'aide de l'outil *Gallicagram*.

3.1. Matériel linguistique traité

Dans les corpus *TenTen*, les données ont été obtenues par collecte automatisée de données web (web crawling), puis soumises à un nettoyage, une déduplication, un filtrage linguistique, une tokenisation, une lemmatisation et une annotation morphosyntaxique (POS-tagging). Le potentiel néologique de ces textes est lié à leurs types : blogs, forums de discussion, médias en ligne.

Les corpus contiennent une large représentation de textes internet se caractérisant par leur actualité et leur lexique familier, et ce pas uniquement pour l'année 2023. Le corpus internet français *frTenTen23* contient 27 111 801 657 tokens, 23 191 789 469 mots et 1 196 093 735 phrases. La plupart des données ont été collectées en décembre 2022 et entre janvier et février 2023, tandis que les textes issus de Wikipédia en français proviennent d'une collecte réalisée entre septembre et novembre 2020. Le corpus slovaque *skTenTen23* a été collecté entre septembre et octobre 2023 et contient 1 086 909 655 tokens, 898 031 101 mots et 61 280 743 phrases ; la source principale est le domaine *wikipedia.org*. Bien que les deux corpus aient été créés selon les mêmes critères méthodologiques et puissent être considérés comme

comparables, ils diffèrent considérablement par leur taille, ce qui est lié au volume différent de contenu internet disponible dans chaque langue.

L'échantillon retenu comprend, côté français, les unités suivantes : *blogage*, *citage*, *fureteur*, *fureteuse*, *googlage*, *networkeur*, *networkeuse*, *plateformisation*, *plate-formisation*, *streameur*, *streameuse*, *supprimage*. Leurs équivalents slovaques observés dans le corpus sont notamment : *blogovanie*, *citovanie*, *surfér*, *surferka*, *gúglenie*, *gúglovanie*, *networker*, *networkerka*, *platformizácia*, *streamer*, *streamerka*, *stopnutie*.

3.2. Grille d'évaluation

Nous évaluons le potentiel lexicographique à l'aide d'une grille de critères combinant des paramètres quantitatifs et qualitatifs. Il s'agit de sept critères (1-7) pour le français et de cinq critères pour le slovaque (1-4, 7), en raison de l'absence d'outil diachronique strictement comparable et de l'absence des lexèmes slovaques dans les dictionnaires terminologiques :

- 1) fréquence dans le corpus ;
- 2) usage dans les différents domaines ;
- 3) transparence sémantique ;
- 4) neutralité stylistique relative ;
- 5) présence dans un dictionnaire terminologique ;
- 6) validation diachronique ;
- 7) adaptation morphologique et graphique.

Chaque critère est évalué de manière binaire (+ / -), puis synthétisé pour estimer le degré de potentiel lexicographique : fort, moyen ou faible.

4. Analyse et résultats quantitatifs

Dans la partie française de l'échantillon (cf. Tableau 1), les fréquences absolues les plus élevées concernent *fureteur* (5077 occurrences), *plateformisation* (1294), *streameuse* (1224), *streameur* (915) et *blogage* (892). À l'inverse, *citage* (10), *networkeur* (23), *networkeuse* (26) et *supprimage* (21) restent très marginaux. La fréquence relative confirme la saillance des mêmes unités, avec toutefois une réserve importante : l'intensité d'attestation ne signifie pas automatiquement stabilisation lexicographique. Par exemple, *fureteur* est très fréquent, mais son histoire lexicale est ancienne et sa valeur actuelle est partiellement reconfigurée dans le domaine numérique. À l'opposé, *plateformisation* a une fréquence déjà élevée tout en relevant clairement d'une dynamique récente. Un cas notable est la coexistence des variantes *plateformisation* et *plate-formisation*. La première forme est nettement dominante (1294 contre 32 occurrences), ce qui suggère un mouvement de stabilisation graphique au profit de la forme soudée. Cette dissymétrie est importante du point de vue lexicographique.

Du côté slovaque, les unités les plus fréquentes sont *blogovanie* (2005 occurrences) et *platformizácia* (1294). Les formes *streamer* (477), *citovanie* (338) et *stopnutie* (192) présentent elles aussi une implantation notable. En revanche, *networkerka* (2), *gúglovanie* (3) ou *surfér* (13) restent quantitativement faibles. Les différences entre les deux langues sont révélatrices. Certaines unités françaises à fréquence relativement faible ont des équivalents slovaques nettement plus attestés malgré la taille beaucoup plus réduite du corpus slovaque. C'est le cas de *citage* face à *citovanie*, ou de *supprimage* face à *stopnutie*. À l'inverse, des dérivés français en *-eur* / *-euse* comme *fureteur* / *fureteuse* ou *streameur* / *streameuse* conservent une forte visibilité en français, tandis que le slovaque privilégie d'autres schémas de désignation. Ces écarts indiquent que l'intégration d'une innovation lexicale n'est pas seulement liée au domaine

référentiel, mais aussi à la productivité des modèles dérivationnels disponibles dans chaque langue.

Tableau 1 : Les fréquences des néologismes TIC

NÉOLOGISME FRANÇAIS	FRÉQUENCE ABSOLUE	FRÉQUENCE RELATIVE	NÉOLOGISME SLOVAQUE	FRÉQUENCE ABSOLUE	FRÉQUENCE RELATIVE
<i>blogage</i>	892	$33 \times 10^{-7} \%$	<i>blogovanie</i>	2005	$18 \times 10^{-7} \%$
<i>citage</i>	10	$0,37 \times 10^{-7} \%$	<i>citovanie</i>	338	$3100 \times 10^{-7} \%$
<i>fureteur</i>	5077	$190 \times 10^{-7} \%$	<i>surfér</i>	13	$12 \times 10^{-7} \%$
<i>fureteuse</i>	278	$10 \times 10^{-7} \%$	<i>surferka</i>	26	$24 \times 10^{-7} \%$
<i>googlage</i>	119	$4,389 \times 10^{-7} \%$	<i>gúglenie</i>	23	$21 \times 10^{-7} \%$
			<i>gúglovanie</i>	3	$2,8 \times 10^{-7} \%$
<i>networkeur</i>	23	$0,85 \times 10^{-7} \%$	<i>networke r</i>	16	$15 \times 10^{-7} \%$
<i>networkeuse</i>	26	$0,96 \times 10^{-7} \%$	<i>networke rka</i>	2	$1,8 \times 10^{-7} \%$
<i>plateformisation</i>	1294	$48 \times 10^{-7} \%$	<i>platformizácia</i>	1294	$2,8 \times 10^{-7} \%$
<i>plateformisation</i>	32	$1,2 \times 10^{-7} \%$			
<i>streameur</i>	915	$34 \times 10^{-7} \%$	<i>streamer</i>	477	$440 \times 10^{-7} \%$
<i>streameuse</i>	1224	$45 \times 10^{-7} \%$	<i>streamerka</i>	9	$8,3 \times 10^{-7} \%$
<i>supprimege</i>	21	$0,77 \times 10^{-7} \%$	<i>stopnutie</i>	192	$180 \times 10^{-7} \%$

4.1. Synthèse des potentiels lexicographiques

L'évaluation croisée des sept critères (cf. Tableau 2) permet de distinguer trois groupes. Les unités à fort potentiel lexicographique sont *plateformisation*, *fureteur* et *fureteuse*. Elles cumulent plusieurs indicateurs favorables : fréquence significative, usage attesté, transparence sémantique, validation externe partielle et, pour certaines, appui diachronique. Les unités à potentiel moyen sont *blogage* et *streameur*. Elles sont visibles dans l'usage et sémantiquement claires, mais restent limitées par des facteurs de registre, de diffusion ou de stabilisation incomplète. Les unités à faible potentiel sont *citage*, *googlage*, *networkeur*, *networkeuse*, *plateformisation*, *streameuse* et *supprimage*. Les raisons varient : fréquence trop basse, forte coloration discursive, absence de validation externe, instabilité graphique ou adaptation incomplète. Un point mérite d'être souligné : *streameuse* est plus fréquente que *streameur*, mais cela ne suffit pas à lui conférer un potentiel supérieur. La fréquence doit être interprétée avec les autres variables, notamment la légitimation terminologique et la stabilité de l'usage.

Tableau 2 : Vue d'ensemble – section française

Néologismes FR	1	2	3	4	5	6	7	TOTAL
<i>blogage</i>	+	-	+	-	+	-	+	4
<i>citage</i>	-	+	+	-	-	-	+	3
<i>fureteur</i>	+	+	+	-	+	+	+	6
<i>fureteuse</i>	+	+	+	-	+	+	+	6
<i>googlage</i>	+	+	+	-	-	-	-	3
<i>networkeur</i>	-	+	+	-	-	-	-	2
<i>networkeuse</i>	-	+	+	-	-	-	-	2
<i>plateformisation</i>	+	+	+	+	+	+	+	7
<i>plate- formisation</i>	-	+	+	+	-	-	-	3
<i>streameur</i>	+	+	+	-	-	+	-	4
<i>streameuse</i>	+	+	+	-	-	-	-	3
<i>supprimage</i>	-	+	+	-	-	-	+	3

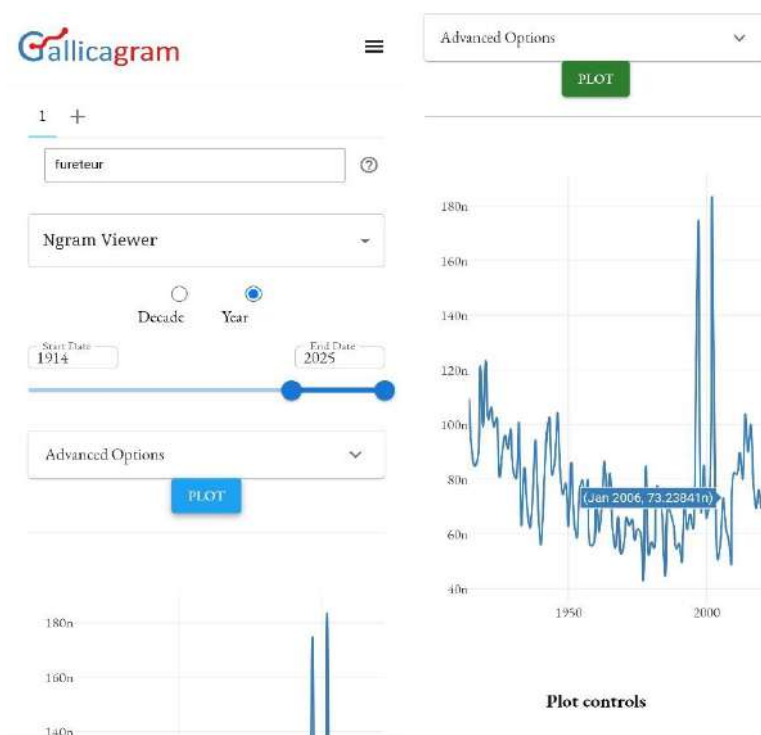
En slovaque, les unités à fort potentiel sont *blogovanie*, *citovanie*, *platformizácia*, *gúglovanie* et *stopnutie* (cf. Tableau 3). Elles sont bien adaptées morphologiquement, transparentes, bien distribuées ou fortement installées dans l'usage. Les unités à potentiel moyen sont *surfer*, *surferka*, *gúglenie* et *streamer*. Elles sont compréhensibles et attestées, mais demeurent plus marquées, moins fréquentes ou moins stabilisées. Les unités à faible potentiel sont *networker*, *networkerka* et *streamerka*. Leur hybridité formelle et leur faible fréquence limitent leur candidature à un traitement lexicographique prioritaire.

Tableau 3 : Vue d'ensemble – section slovaque

Néologismes SK	1	2	3	4	7	TOTAL
<i>blogovanie</i>	+	+	+	+	+	5
<i>citovanie</i>	+	+	+	+	+	5
<i>surfér</i>	-	+	+	-	+	3
<i>surferka</i>	-	+	+	-	+	3
<i>gúglenie</i>	-	+	+	-	+	3
<i>gúglovanie</i>	+	+	+	-	+	4
<i>networker</i>	-	+	+	-	-	2
<i>networkerka</i>	-	+	+	-	-	2
<i>platformizácia</i>	+	+	+	+	+	5
<i>streamer</i>	+	+	+	-	-	3
<i>streamerka</i>	-	+	+	-	-	2
<i>stopnutie</i>	+	+	+	-	+	4

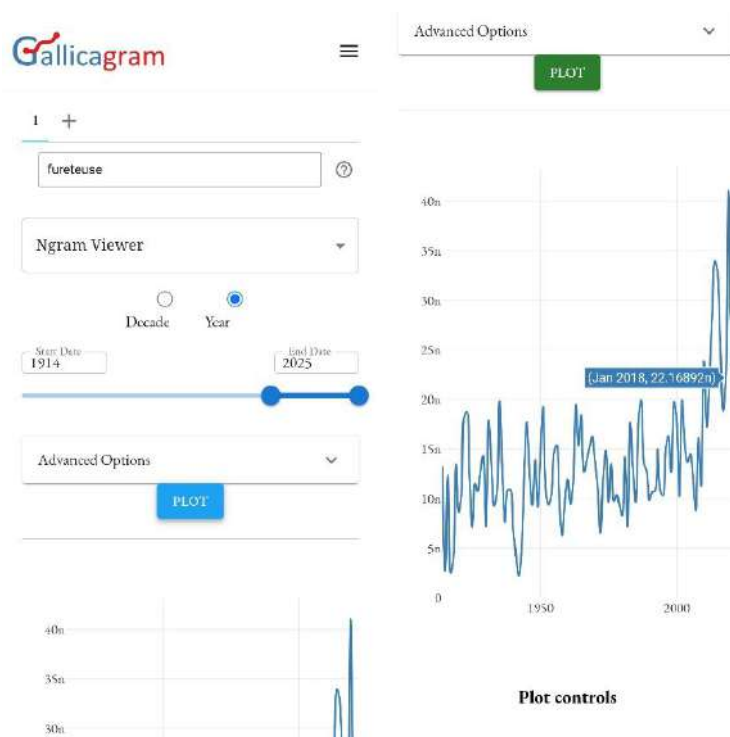
4.1. Dimension diachronique

Pour le français, la confrontation avec *Gallicagram* affine l'analyse. Plusieurs unités de l'échantillon ne laissent pas de trace significative dans la presse historique : *blogage*, *googlage*, *networkeur*, *networkeuse*, *streameur*, *supprimage*. Cette absence confirme leur caractère récent ou marginal hors web. D'autres, en revanche, présentent un profil diachronique exploitable. *Fureteur* (cf. Figure 1) montre une présence stable tout au long du XXe siècle, suivie d'un déclin progressif, puis de pics marqués à partir des années 2000. Ces variations ne traduisent pas l'émergence d'un néologisme, mais plutôt une réactivation liée à de nouveaux usages, notamment dans le domaine numérique.

Figure 1 : *Fureteur*

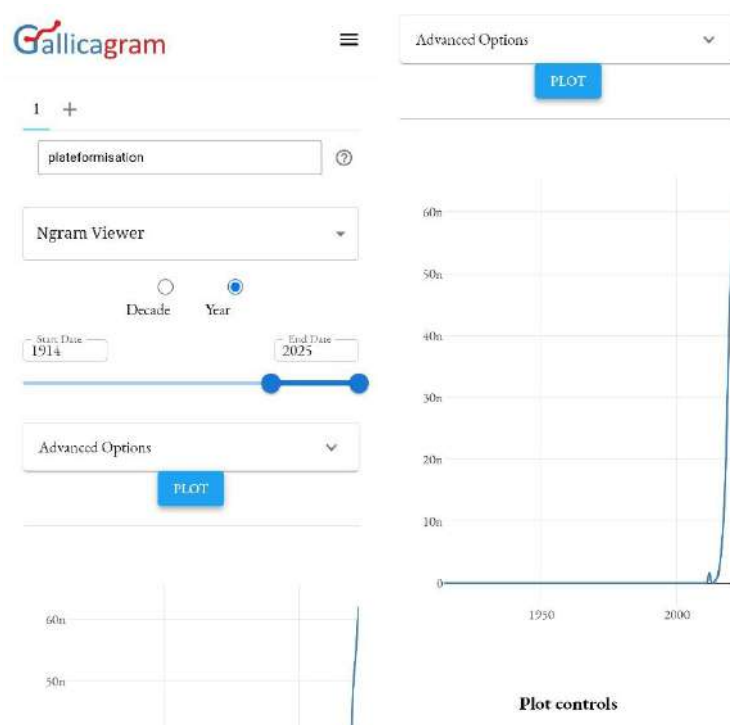
La courbe diachronique de *fureteuse* (cf. Figure 2) montre que, bien que marginale pendant une grande partie du XXe siècle, ce néologisme connaît une nette augmentation de fréquence à partir des années 2000, avec un pic autour de 2018. *Fureteuse* connaît donc une diffusion tardive, compatible avec la montée de la féminisation lexicale.

Figure 2 : Fureteuse



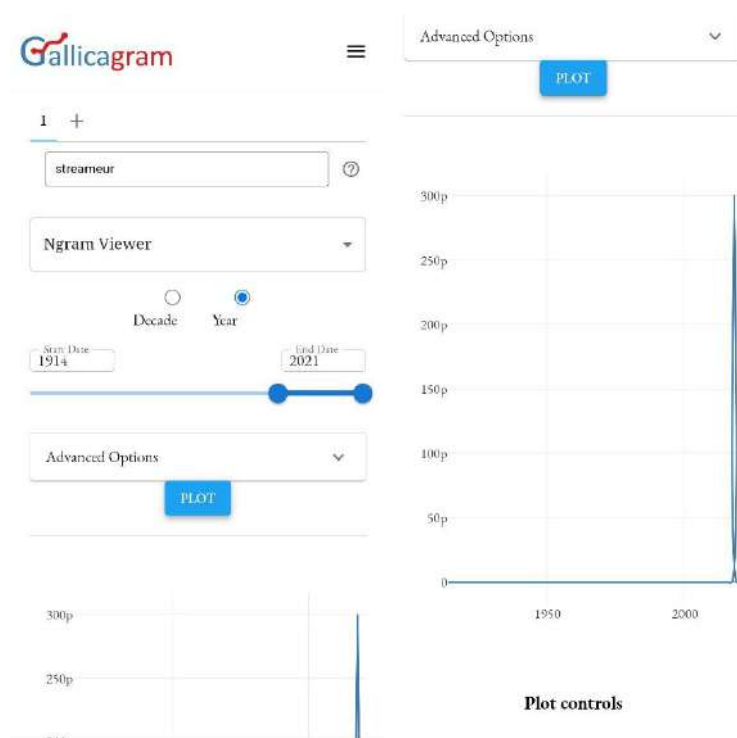
Plateformisation (cf. Figure 3), quant à elle, ne se développe qu'après 2010, avec une hausse sensible après 2020. Ce comportement est typique d'un néologisme en cours de stabilisation.

Graphique 3 : Plateformisation



Streameur (cf. Figure 4) apparaît également comme une innovation récente à croissance rapide.

Figure 4 : *Streameur*



5. Conclusion

Les résultats montrent que les néologismes issus du domaine des TIC présentent des degrés de stabilisation variables et que la fréquence seule ne constitue pas un critère suffisant pour justifier leur intégration dans les dictionnaires. De ce fait, le nombre d'occurrences dans les corpus web *frTenTen23* et *skTenTen23* a été mis en relation avec d'autres paramètres : usage contextuel, neutralité stylistique, stabilité graphique et morphologique, transparence sémantique et ancrage diachronique.

L'analyse quantitative menée à partir de *Sketch Engine* montre que les néologismes suffixés des TIC n'occupent pas tous le même degré de stabilisation dans l'usage. En français, *plateformisation*, *fureteur* et *fureteuse* apparaissent comme les meilleurs candidats à un traitement lexicographique renforcé. En slovaque, *blogovanie*, *citovanie*, *platformizácia*, *gúglovanie* et *stopnutie* présentent le profil le plus favorable. L'apport principal de l'étude est méthodologique. Elle montre qu'une évaluation crédible du potentiel lexicographique doit reposer sur une grille multicritère. C'est à cette condition que l'on peut passer du simple constat néologique à une décision de traitement lexicographique fondée empiriquement, la fréquence constituant un point d'articulation entre la norme et l'usage.

Références

- Campese, S. (2025). *Nouveaux mots du Petit Larousse et du Petit Robert 2025 : que disent-ils de nous ?* Projet Voltaire. <https://www.projet-voltaire.fr/ressources/nouveaux-mots-dictionnaire-2025/>
- Catford, J. C. (1965). *A linguistic theory of translation*. Oxford University Press.
- Dictionnaire de l'Académie française (DAF) (s. d.). *Académie française*. <https://www.dictionnaire-academie.fr/>
- frTenTen23 : *French Web 2023 Corpus* (s. d.). Sketch Engine. <https://www.sketchengine.eu/frtenten-french-corpus/>
- Le Petit Larousse (éd. annuelle). *Dictionnaire de la langue française*. Paris : Larousse.
- Le Petit Robert (éd. annuelle). *Dictionnaire alphabétique et analogique de la langue française*. Paris : Le Robert.
- Le Robert Dico en ligne (s. d.). *Le Robert*. <https://dictionnaire.lerobert.com/>
- Moraru, L. (2023). Considérations sur l'enseignement/apprentissage des néologismes dans la classe de FLE. *De la monem la text: reconfigurări lingvistice și practici didactice*, 182-186.
- Mudrochová, R. (2021). *Jean-François Sablayrolles : Comprendre la néologie. Conceptions, analyses, emplois*. *Časopis pro moderní filologii*, 103(2): 295-302.
- Ráčková, L. (2026, à paraître). *Les néologismes en français et leurs équivalents slovaques. L'exemple de la suffixation nominale*. L'Harmattan.
- Sablayrolles, J.-F. (2000). *La néologie en français contemporain*. Honoré Champion.
- Sablayrolles, J.-F. (2019). *Comprendre la néologie. Conceptions, analyses, emplois*. Lambert-Lucas.
- skTenTen23 : *Slovak Web 2023 Corpus* (s. d.). Sketch Engine. <https://www.sketchengine.eu/sktenten-slovak-corpus/>
- Slovníky slovenského jazyka (JÚEŠ SAV) (s. d.). *Jazykovedný ústav Ľ. Štúra*. <https://slovník.juls.savba.sk/>
- Trésor de la langue française informatisé (TLFi) (s. d.). *ATILF*. <https://www.atilf.fr/tlfi>

Vers un chatbot basé sur la recherche sémantique et le raisonnement à base de règles pour l'optimisation des opérations de la Sécurité Sociale française

Wenxiang Zhou^{1,2}, Touria Aït El Mekki¹, Frédéric Lardeux¹, Marc Legeay¹,
Jean-Olivier Boesmier²

¹Univ Angers, LERIA, F-49000 Angers, France — {firstname}.{lastname}@univ-angers.fr

²CPAM, La Roche-sur-Yon, France — {firstname}.{lastname}@assurance-maladie.fr

Abstract

Numerous LLM-based chatbots leveraging Retrieval-Augmented Generation (RAG) are being deployed across various industrial domains. However, in highly regulated knowledge-intensive sectors such as social security services who depend on exceptionally precise terminology and up-to-date legal and procedural information, their performance remains challenging concerning accuracy, reliability and traceability, as highlighted in recent work showing the limitations of RAG-only method in legal and regulatory context (Beyneix and Rovinski, 2025). These limitations are particularly challenging for experts in CPAM (French National Health Insurance Office) who handle complex cases while navigating large volumes of constantly evolving data. To address these issues, this study aims to develop a hybrid chatbot capable of providing accurate, consistent and traceable answers to resolve allowance claims submitted to the CPAM. The proposed system relies on authoritative regulatory data in French, including historical question-answer, legal texts from the French Social Security Code and internal technical guidance documents. First, a hybrid semantic similarity module retrieves relevant answers from historical database by weighting vector-based semantic similarity with lexical and terminological matching, thereby ensuring preservation of the importance of domain-specific key terms. Second, when sufficiently similar questions are not found, a rule-based reasoning engine is activated to generate a reliable answer from logical rules. The logical rules, extracted with an LLM from validated internal technical documents, are reviewed by the experts and triggered via key terms detection. Finally, a generative LLM produces the final answer by integrating the retrieved information and providing explicit source citations. Although the project is ongoing, several important steps have been completed, including the preprocessing and cleaning of heterogeneous textual datasets, the thematic classification of historical questions, the clustering of technical documents, whose thematic grouping corresponds to expert classifications, the terminology extraction and semantic similarity analysis, and the extraction of logical rules. Based on this, we obtained a first functional prototype. These results are crucial for structuring authoritative regulatory datasets and validating the feasibility of retrieval components. Beyond the final system implementation, this research contributes a reusable methodology for structuring regulatory french corpora and developing hybrid reasoning chatbot applicable to other administrative or legal contexts.

Keywords : Retrieval-Augmented Generation (RAG), Rule-based Reasoning, Legal context

Résumé

De nombreux chatbots basés sur le LLM et la technique RAG (Retrieval-Augmented Generation) sont actuellement déployés dans divers domaines industriels. Cependant, dans les secteurs hautement réglementés et à forte intensité de connaissances, tels que les services de sécurité sociale, qui dépendent d'une terminologie précise et d'informations juridiques et procédurales à jour, leurs performances restent difficiles en termes d'exactitude, de fiabilité et de traçabilité, comme le soulignent des travaux récents montrant les limites de la méthode RAG seule dans le contexte juridique et réglementaire (Beyneix and Rovinski, 2025). Ces contraintes sont particulièrement difficiles à gérer pour les experts de la Caisse primaire d'assurance maladie (CPAM) qui traitent des cas complexes

tout en naviguant dans de grands volumes de données évolutives. Cette étude vise à développer un chatbot hybride capable de fournir des réponses précises, cohérentes et traçables afin de résoudre les demandes d’allocations soumises à la CPAM. Le système proposé s’appuie sur des données réglementaires en français, y compris des données historiques Question-Réponse, des textes juridiques issus du Code de la Sécurité Sociale français et des documents techniques internes. D’abord, un module hybride de similarité sémantique récupère les réponses pertinentes dans la base historique en pondérant la similarité sémantique, lexicale et terminologique, garantissant ainsi la préservation de l’importance des termes clés spécifiques au domaine. Ensuite, lorsqu’aucune question suffisamment similaire n’est trouvée, un moteur de raisonnement basé sur des règles est activé afin de générer une réponse fiable à partir de règles logiques. Les règles logiques, extraites à l’aide d’un LLM à partir de documents techniques internes validés, sont examinées par les experts et déclenchées via la détection de termes clés. Enfin, un LLM génératif produit la réponse finale en intégrant les informations récupérées et en fournissant des citations explicites des sources. Bien que le projet soit toujours en cours, plusieurs étapes importantes ont déjà été franchies, notamment la constitution et le pré-traitement des données textuelles hétérogènes, la classification thématique des questions historiques, le clustering de documents techniques dont la classification thématique correspond à celle des experts, l’extraction terminologique et l’analyse de la similarité sémantique, ainsi que l’extraction des règles logiques. Sur cette base, nous avons obtenu un premier prototype fonctionnel. Ces résultats constituent une étape clé pour structurer des ensembles de données réglementaires et valider la faisabilité des composants du moteur de recherche. Au-delà de la mise en œuvre finale du système, cette recherche apporte une méthodologie réutilisable pour structurer des corpus réglementaires français et développer un chatbot à raisonnement hybride applicable à d’autres contextes administratifs ou juridiques.

Mots clés : Retrieval-Augmented Generation (RAG), Raisonnement basé sur des règles logiques, Contexte juridique

1. Introduction

La Caisse Primaire d’Assurance Maladie (CPAM) est l’institution française qui gère entre autres l’assurance santé en France. Son champ d’action est régi par plusieurs règles complexes majoritairement issues du Code de la Sécurité Sociale et d’instructions nationales. Les agents de la CPAM doivent ainsi gérer des dossiers très complexes et demandent une grande expertise. Pour les aider dans cette tâche, un agent de la CPAM peut, en cas de doute, solliciter un expert pour clarifier les consignes ambiguës ou interpréter une situation atypique. Ce processus de travail, bien qu’efficace, crée une charge importante pour les experts au détriment d’activités à plus forte valeur ajoutée. Dans ce contexte opérationnel, un chatbot d’assistance facilite l’interprétation des règles, propose une réponse pertinente et aide les agents dans leur prise de décision arbitraire. Parallèlement, les chatbots d’assistance basés sur les modèles de grande taille (Large Language Models - LLM) sont utilisés dans divers domaines pour assister les utilisateurs dans leur travail quotidien, notamment en santé (Nayinzira, 2025), en éducation (Nouali et al., 2025) ou en assurance automobile (Beauchemin et al., 2024). Toutefois, les approches existantes basées sur les LLM restent limitées lorsqu’elles sont appliquées dans un chatbot d’assistance dans un domaine spécifique. Les LLM généralistes manquent de connaissances spécifiques et risquent d’avoir des hallucinations (Metheniti et al., 2025), tandis que la méthode *fine-tuning*, bien que performante, est coûteuse et difficile à adapter aux données évolutives. Les méthodes basées sur la génération augmentée de recherche (Retrieval-Augmented Generation - RAG) présentent des risques d’hallucinations et des difficultés à raisonner de manière fiable et traçable dans un contexte réglementaire (Beyneix and Rovinski, 2025).

Malgré la généralisation des chatbots basés sur des LLM, très peu de recherches se sont concentrées sur le domaine juridique, notamment la sécurité sociale française. Les méthodes existantes, basées sur des LLM ou utilisant exclusivement un RAG, risquent de produire des réponses hallucinées ou obsolètes et manquent de traçabilité. Leur performance dépend énormément de la qualité des informations récupérées, surtout dans le contexte réglementaire où les données sont expertisées, complexes, hétérogènes et réparties sur plusieurs documents, ce qui complexifie la récupération d’informations pertinentes. Pour y répondre, dans cet arti-

cle, nous proposons une approche hybride combinant une recherche documentaire optimisée dans une base historique de questions/réponses et l'exploitation structurée de documents juridiques/réglementaires. Le module de récupération fonctionne en deux étapes. D'abord, il identifie les questions existantes les plus proches de celle posée grâce à un score composite. Ce score composite est calculé sur la similarité sémantique, lexicale et terminologique. Si aucune réponse pertinente n'est trouvée, un module de recherche active un moteur de règles logiques. Ces règles logiques sont extraites par un LLM et validées par des experts. Dans les deux cas, un modèle de langage produit une réponse finale adaptée à la question en garantissant la précision, la fiabilité et la traçabilité de la réponse.

Cet article est structuré de la manière suivante. La section 2 présente l'état de l'art. Les sections 3 et 4 décrivent respectivement la constitution, la préparation du corpus et la méthode envisagée. Pour mieux comprendre et structurer des données juridiques et réglementaires, dans la section 5, nous présentons les résultats préliminaires, notamment la classification thématique des questions historiques, le clustering des documents techniques, l'extraction terminologique et l'analyse de la similarité sémantique, ainsi que l'extraction des règles logiques. Enfin, l'article se conclut par une discussion du premier prototype et la continuité de notre projet.

2. État de l'art

Les systèmes conversationnels ont connu une évolution significative avec l'émergence des LLM ces dernières années. Grâce à leurs capacités de compréhension et de génération, ils dépassent des approches traditionnelles basées souvent sur des règles prédéfinies ou de workflow sous forme de graphes ou d'arborescence (Rémy et al., 2024). Cela permet d'avoir une amélioration des expériences utilisateurs en termes de flexibilité et de qualité conversationnelle. Toutefois, les travaux relatifs (Beyneix and Rovinski, 2025; Metheniti et al., 2025) relèvent les limites des LLM dans les domaines spécialisés, notamment le phénomène des hallucinations, la difficulté d'accès à des données spécifiques à jour et l'absence de traçabilité.

Pour pallier ces limites, les approches de type RAG, combinant un module de recherche documentaire avec un modèle génératif, ont été proposées. Les RAG ont été largement adoptés dans divers domaines, tels que la médecine (Liu, 2024), le business (Miyaji et al., 2025) et le juridique (Mohamed et al., 2025) pour apporter des réponses fiables en minimisant les hallucinations. Plusieurs produits industriels appliqués au juridique ont également émergé, comme LexisNexis, Lexis + AI et Predictice. Les RAG justifient des performances puissantes pour générer des réponses en langage naturel, mais restent dépendants largement de la pertinence des données récupérées (Jourdain et al., 2024) et ne permettent pas d'assurer un raisonnement logique explicite (Feillet, 2024). De ce fait, leur usage en contexte juridique nécessite souvent de la prudence du fait qu'ils peuvent possiblement produire des informations hallucinées, erronées ou obsolètes. Les outils conçus pour traiter spécifiquement des tâches juridiques à l'aide de RAG ne parviennent pas à réduire totalement ces risques (Beyneix and Rovinski, 2025).

Parallèlement, plusieurs approches méthodologiques pour interpréter des documents réglementaires avec des algorithmes de traitement automatique du langage naturel ont été proposées ces dernières années. (Aussenac-Gilles et al., 2023) ont montré une synthèse des méthodes d'extraction et de modélisation de règles formelles à partir de réglementations légales en mettant l'accent sur les aspects linguistiques. Leur travail est essentiel pour comprendre l'identification et l'extraction des entités clés des textes réglementaires, comme les obligations, les permissions, etc. De plus, (Mimouni et al., 2015) ont proposé une ontologie documentaire pour faciliter l'accès aux contenus juridiques en articulant le contenu au niveau local et global. Face aux contraintes du système conversationnel traditionnel basé uniquement sur RAG et aux avancées réalisées de l'interprétation fine des données juridiques, de nouvelles perspectives s'ouvrent en

combinant des approches symboliques et LLM dans le système conversationnel. Dans cette continuité, (Billi et al., 2024) ont renforcé un système expert juridique existant avec des règles logiques Prolog extraites par LLM pour surmonter la difficulté de la représentation des connaissances juridiques. (Feillet, 2024) a proposé plusieurs combinaisons entre LLM et moteurs de raisonnement basés sur des règles logiques pour développer un chatbot hybride. L'intégration du moteur d'inférence de règles dans le chatbot permet de compenser les limites des LLM en structurant et contextualisant les réponses générées à partir de la base de connaissances.

3. Corpus : présentation et préparation

Notre corpus d'analyse est constitué de différentes sources spécifiques à la CPAM: la base historique des Question-Réponse, les fiches techniques et les directives nationales internes comme les lettres-réseaux ou les circulaires. Les textes juridiques se caractérisent par leur nature normative marquée par l'utilisation d'un style et un vocabulaire particulier (Gémar, 2002). Par exemple, l'article R313-3-1 du Code de la Sécurité Sociale disponible sur Légifrance¹ : “ *En cas de redressement d'assiette sur la base d'une rémunération forfaitaire mis en recouvrement conformément aux dispositions de l'article L. 242-1-2, les conditions d'ouverture des droits (...)* ” utilise les terminologies comme *redressement d'assiette*, *rémunération forfaitaire* et la référence *l'article L. 242-1-2* pour définir les conditions d'ouverture des droits.

Dans notre recherche, les données réglementaires présentées sur des sites officiels, tels que Améli-réseau, ont été récupérées avec du *web scraping* et les données de la base historique ont été récupérées via une API. Notre corpus est constitué de trois sous-corpus:

- 315 fiches techniques : Documents internes rédigés par des experts métier qui traduisent les textes réglementaires en consignes opérationnelles, à la disposition des agents sur Améli-réseau.
- 1183 directives nationales internes : Instructions nationales disponibles sur Améli-réseau au format PDF définissant l'interprétation et l'application des règles. Elles constituent le cadre réglementaire et opérationnel sur lequel les experts s'appuient pour rédiger les fiches techniques.
- Base historique Question-Réponse (composée de 1315 paires question-réponse) : Recueil des échanges textuels entre agents et experts. La CPAM organise son activité autour de plusieurs thématiques, comme *maladie*, *accident du travail et de trajet*, *invalidité*, *maladie professionnelle*, *maternité* et *rente*. Chaque paire de question-réponse est annotée avec un ou plusieurs labels thématiques. La répartition des labels annotés est déséquilibrée: le thème *maladie* représentent 34% des annotations, *accident du travail et de trajet* 25%, *maladie professionnelle* 23%, *invalidité* 8%, tandis que *rentes* 5%, *maternité* 4% et *décès* 1% sont très minoritaires. L'extraction textuelle des PDF des directives nationales internes a été réalisée avec la bibliothèque Unstructured². Les tableaux contenus dans les directives ont été reformulés en langage naturel par un LLM (Mistral 7B³), puis validés par des experts.

Un pré-traitement spécifique a été appliqué à chaque ensemble de données selon ses exigences. Pour la base historique Question-Réponse, les acronymes métier ont été remplacés grâce à une liste prédéfinie des acronymes métier; les données personnelles ont été anonymisées grâce à SpaCy⁴ et à des expressions régulières, puis nettoyées en supprimant les caractères spéciaux, les ponctuations et les mots vides, avant la lemmatisation avec SpaCy. Les données structurées, comme les fiches techniques et directives nationales, ont été légèrement pré-traitées: suppression des caractères spéciaux, des ponctuations, des mots vides, puis lemmatisation avec SpaCy.

¹ Article consulté sur le site Légifrance (<https://www.legifrance.gouv.fr/>) le 3 mars 2026

² <https://pypi.org/project/unstructured/>

³ <https://mistral.ai/news/announcing-mistral-7b>

⁴ spaCy · Industrial-strength Natural Language Processing in Python

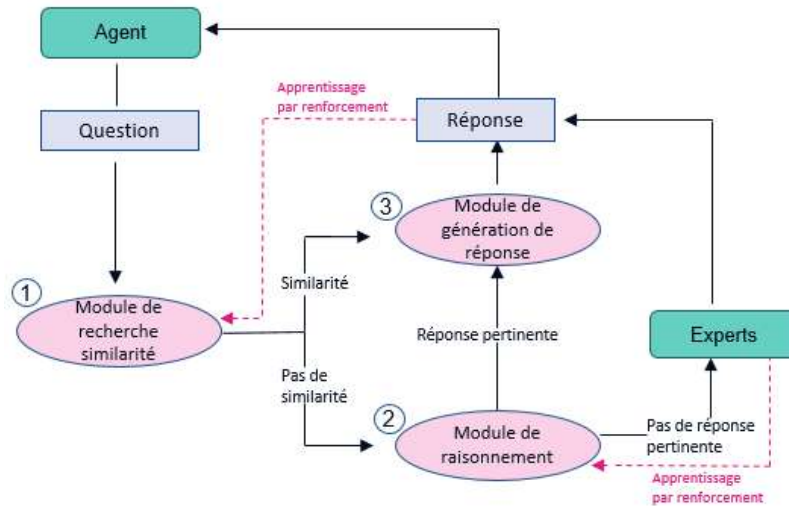


Figure 1: Chaîne de traitements du chatbot.

4. Méthode proposée

Nous proposons une architecture hybride pour développer un chatbot, illustrée en Figure 1 : un module de recherche de similarité entre la question posée et les questions qui ont déjà été posées; un module de raisonnement, basé sur des règles, visant à récupérer des éléments de réponse; un module de génération de réponse par un LLM. Lorsqu'une nouvelle question est soumise au système, elle passe d'abord par une étape de pré-traitement similaire à celle du corpus. Cette question pré-traitée est ensuite transmise au module de recherche de similarités.

4.1. Module de recherche de similarités

Pour récupérer des informations pertinentes, le module de recherche de similarités (Module 1 sur la Figure 1) repose sur un score composite de similarité basé sur trois dimensions complémentaires de similarité textuelle à partir de la base historique Question-Réponse. (1) la composante lexicale réalisée avec l'algorithme BM25 (Shetty, 2025), permet une mesure lexicale entre la question de l'utilisateur et les questions historiques. L'algorithme BM25 est une variante de TF-IDF qui intègre la fréquence des termes, la rareté des termes et la longueur du document. (2) la composante sémantique basée sur des vecteurs permet de capturer la proximité sémantique réalisée avec l'algorithme *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2*⁵. (3) la composante terminologique, mesurée par le taux de termes communs grâce à une liste des termes clés spécifiques entre la question de l'utilisateur et les questions existantes. Cette mesure intègre le recouvrement des termes clés qui sont souvent révélateurs de similarités thématiques.

La vectorisation étant coûteuse en termes de calcul, les 100 premières questions candidates basées sur la similarité lexicale grâce à l'algorithme BM25 sont d'abord filtrées. Ensuite, la question de l'utilisateur et les 100 questions candidates sont converties en vecteur avec l'algorithme. Puis la similarité terminologique entre la question de l'utilisateur et les questions candidates est calculée. Ces trois scores de similarité sont ensuite pondérés, comme défini dans l'équation (1), selon leur importance relative dans l'évaluation de la similarité.

$$\text{score}_{\text{final}} = \text{score}_{\text{embedding}} \cdot c_1 + \text{score}_{\text{BM25}} \cdot c_2 + \text{score}_{\text{terminologie}} \cdot c_3 \quad (1)$$

⁵ <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

Les coefficients c_1 , c_2 et c_3 seront fixés après une étude des différentes combinaisons afin d'équilibrer entre les similitudes des termes métier et des termes génériques. Cette agrégation pondérée permet d'identifier, à l'aide d'un seuil, s'il existe une ou plusieurs questions candidates similaires à la question posée. Si tel est le cas, ces éléments sont transmis au module de génération de réponse, sinon, le module de raisonnement est enclenché.

4.2. Module de raisonnement

Le module de raisonnement (Module 2 sur la Figure 1) repose sur un moteur d'inférence de règles. Ces règles logiques sont générées à partir des documents légaux, des directives nationales et des fiches techniques grâce à un modèle LLM (Mistral 7B). Les règles logiques sont liées à chaque section des documents d'origine, afin de pouvoir sourcer les règles et de pouvoir les désactiver si un nouveau texte vient rendre la section obsolète. Ces règles logiques sont représentées sous forme de structure conditionnelle en intégrant différents types de normes (comme obligations, permissions, interdictions) et sont ensuite validées par des experts. Chaque règle générée a été sauvegardée dans des fichiers JSON structurés, avec le nom de la fiche, la date de mise à jour et les références du document source sur lesquelles cette fiche s'appuie. Cette étape vise à transformer les règles textuelles du langage naturel en langage logique, en y intégrant des métadonnées complémentaires pour rendre les règles plus lisibles et traçables.

L'identification des règles pertinentes repose sur une correspondance entre les termes clés de la question et les éléments de l'intitulé des règles. Une fois les règles candidates sélectionnées, le moteur évalue leurs conditions d'application, examine la présence éventuelle d'exceptions et détermine la conclusion à la question de l'utilisateur. Si aucune règle pertinente n'est trouvée, la question sera envoyée à un expert. Sa réponse sera ensuite introduite dans la base Question-Réponse. Si une ou plusieurs règles pertinentes sont activées, les éléments de réponse ainsi générés sont transmis au module de génération de réponse.

4.3. Module de génération de réponse

Le module de génération (Module 3 sur la Figure 1) constitue l'étape finale du système et consiste à produire une réponse contextualisée, cohérente et sourcée à partir des éléments textuels fournis par le module de recherche de similarités ou le module de raisonnement. Le mécanisme du fonctionnement repose sur un prompt structuré combinant la requête d'utilisateur, les passages textuels jugés pertinents identifiés par les différentes méthodes hybrides de récupération en amont (réponse associée à une question existante dans la base historique lorsque celle-ci est jugée proche de celle de l'utilisateur, sinon réponse obtenue en déclenchant les règles logiques). Ce prompt intègre aussi une instruction définissant le comportement attendu, pour guider la génération vers une réponse cohérente et fidèle aux données fournies.

5. Résultats préliminaires

Les travaux sont toujours en cours sur les différentes étapes de la méthodologie. Nous présentons ici les résultats du traitement de la base historique et des fiches techniques. Les autres directives nationales, comme les textes de loi ou les circulaires, n'ont pas encore été exploitées.

5.1. Classification thématique des questions par apprentissage automatique

Chaque paire question-réponse de la base historique est annotée avec un ou plusieurs labels thématiques, ce qui reflète la complexité des questions posées. Cette observation justifie la conception d'un modèle prédictif des multi-labels thématiques correspondants à une nouvelle question, facilitant l'orientation de la recherche vers les fiches techniques pertinentes. Les questions

Table 1: Évaluation comparative pour la classification multi-label

vectorisation	classifieur	precision macro	recall macro	F1 macro	F1 weighted
TF-IDF	SVM	0.94	0.79	0.85	0.90
TF-IDF	RL	0.85	0.79	0.81	0.87
TF-IDF	Ridge	0.93	0.80	0.85	0.89
TF-IDF	RF	0.67	0.42	0.47	0.74
SBERT	SVM	0.57	0.73	0.63	0.73
SBERT	RL	0.46	0.70	0.53	0.68
SBERT	Ridge	0.52	0.73	0.59	0.72
SBERT	RF	0.46	0.27	0.33	0.55

historiques ont été divisées en un ensemble d’entraînement (80%) et d’évaluation (20%). Nous avons vectorisé les questions pré-traitées grâce à deux méthodes: TF-IDF avec un n-gramme de taille 1 à 3 et une représentation sémantique basée sur Sentence-BERT (SBERT), à savoir *sentence-transformers/paraphrase-multilingual-mpnet-base-v2*⁶. La classification multi-label a été réalisée avec *OneVsRestClassifier*. Nous avons utilisé *MultilabelStratifiedShuffleSplit* pour garantir une répartition proportionnelle de chaque thème. Quatre classifieurs ont été évalués: Régression logistique (RL), Random Forest (RF), SVM, Ridge Classifier (Ridge), tous configurés avec l’option *class_weight=“balanced”* pour compenser certaines classes rares, comme *décès*, *maternité*. Les résultats (illustré dans le Tableau 1) mettent en évidence deux points: d’une part, les performances obtenues avec TF-IDF sont nettement supérieures à celles de SBERT pour l’ensemble des métriques. Cette remarque suggère que les questions montrent une forte distinctivité lexicale, probablement liée aux termes métiers propres à chaque thème. D’autre part, les modèles linéaires, notamment SVM, sont plus performants, tandis que le modèle à base d’arbres comme RF est moins efficace. L’ensemble des résultats indique que la structure de la tâche de classification thématique est globalement linéaire et que les thématiques définies par des experts sont exploitables de manière automatisée. Le modèle SVM configuré avec TF-IDF est suffisamment performant pour la prédiction thématique d’une nouvelle question.

5.2. Clustering thématique des fiches techniques

Nous avons utilisé BERTopic⁷, une méthode non supervisée de modélisation de sujets, réglé avec Sentence-Camembert-base⁸ pour regrouper les fiches techniques en clusters thématiques. Les 3 clusters générés montrent une structuration thématique cohérente des fiches techniques. Le cluster 0, le plus grand, assemble majoritairement les fiches liées aux sujets *maladie* et *indemnité journalière accident du travail*. Le cluster 1 regroupe des fiches abordant *invalidité* et *rente*, aux liens sémantiques validés par des experts, tandis que le 2 est dominé par les fiches sur la *maternité*. La cohérence des clusters, validés par les experts, confirme la capacité du modèle à refléter fidèlement la proximité sémantique entre fiches, ce qui facilite l’orientation vers la fiche pertinente pour une nouvelle question, une fois son sujet identifié par le classifieur.

5.3. Extraction terminologique et analyse de similarité sémantique

Les termes désignent des concepts métiers, simples ou composés, et essentiels à la compréhension des textes. Des travaux antérieurs existent sur l’extraction de termes clés, notamment

⁶ <https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2>

⁷ <https://github.com/MaartenGr/BERTopic>

⁸ <https://huggingface.co/dangvantuan/sentence-camembert-base>

(Alrahabi et al., 2025). Avec LLM (Mistral 7B), nous avons créé un ensemble terminologique complet et adapté, utilisé par le module de recherche de similarités. Cet ensemble contient des termes comme *arrêt de travail*, *accident de travail*, validés par des experts, ensuite complétés par ceux qui avaient déjà été identifiés par experts. Contrairement aux termes métiers, les entités nommées (EN) désignent les noms de personnes, de lieux, d'organisations, les dates et les unités monétaires. La reconnaissance des entités nommées (NER) identifie les éléments essentiels pour interpréter des textes. Les approches NER reposent soit sur des règles linguistiques, soit sur des modèles d'apprentissage profond, comme BERT, dont la variante CamemBERT Judiciaire (Mahmoudi et al., 2022). Les outils tels que SpaCy ou Stanza permettent aussi l'extraction des EN pour les textes français. Toutefois, leur performance reste limitée lorsqu'ils traitent des corpus spécialisés, notamment le recueil d'échange textuel riche en abréviations (par exemple *HDF* pour *Haut de France*). La technicité lexicale réduit la précision des EN extraites. Par exemple, les termes *Caisse des français de l'étranger* et *Agence régionale de santé* rarement présentés dans les corpus d'entraînement généralistes. Pour y répondre, nous avons extrait les EN avec *EntityRuler*, un composant de SpaCy intégré des règles régulières personnalisées. Cette méthode est légère et déterministe, sans entraînement d'un modèle supervisé. Les experts ont défini un ensemble de catégories d'entités essentielles (références réglementaires comme les articles du code, lettres-réseau, circulaires, noms des fiches techniques, types de régime, durée, date, organismes). Nous avons extrait avec succès les EN de notre base historique. Une fois les termes métiers extraits, nous avons analysé la similarité sémantique sur l'ensemble de la base historique. Les méthodes classiques TF-IDF permettent de repérer des termes spécifiques, tandis que les modèles Word2Vec (Mikolov et al., 2013) et FastText (Bojanowski et al., 2017) représentent distributionnellement des termes sensibles aux contextes locaux. Nous avons utilisé plusieurs méthodes de vectorisation pour explorer des lexiques proches dans la base historique, à savoir Word2Vec de type Skip-gram, FastText, SBERT. L'évaluation comparative des méthodes mobilisées indique que SBERT propose les résultats les plus pertinents jugés par les experts. À titre indicatif, nous avons noté que les termes comme *congé paternité*, *congé légal de naissance* se rapprochent de *congé maternité*; tout comme *accident mortel*, *accident de trajet* se rapprochent d'*accident de travail*; *décéder*, *mort*, *suicide* se rapprochent de *décès*. Les résultats de similarité sémantique sont ensuite validés par des experts.

5.4. Extraction des règles logiques

Pour interroger la base de règles logiques, nous envisagerons d'extraire des connaissances à partir des fiches techniques définissant les consignes opérationnelles. À titre exploratoire, nous avons commencé par les fiches sur la maternité, plus réduites en taille. Les règles en langage naturel sont transformées en règles conditionnelles en mobilisant plusieurs modèles LLM (Mistral Nemo 12B⁹, Mistral 7B, Llama3.2 3B¹⁰). Nous avons optimisé les prompts pour améliorer la qualité des règles logiques, en ajustant précisément les instructions données au modèle pour encadrer la génération et minimiser l'hallucination. Chaque section de la fiche est traitée indépendamment pour extraire les règles logiques selon la structure hiérarchique (titres et sous-titres numérotés). Les règles générées ont été stockées dans des fichiers JSON enrichis de métadonnées (nom de la fiche, date de la mise à jour et références) pour la traçabilité, comme illustré dans la Figure 2. Mistral Nemo 12B a montré la meilleure performance. À ce stade du projet, une phase exploratoire a permis d'extraire des règles à partir des fiches maternité avec Mistral Nemo 12B. L'extraction à partir de l'ensemble des fiches et leur évaluation sont en cours.

⁹ <https://mistral.ai/news/mistral-nemo>

¹⁰ <https://huggingface.co/meta-llama/Llama-3.2-3B>


```

{"meta":{"date_mise_jour":"10/02/2002",
"nom_fiche":"MALADIE-001 - maternité",
"doc_source":"article R123-3"},
"paragraphe": "L'indemnité du congé maternité sont accordées aux assurées ayant au moins 10 mois
d'affiliation et au moins 150 heures de travail, sauf celles en arrêt maladie longue durée (LR01-01)",
"regles":{"intitulé":"condition accès prestation congé maternité",
"condition":"si l'assurée ayant au moins 10 mois d'affiliation et 150 heures de travail",
"effet":"avoir l'indemnité du congé maternité",
"formule_logique":"(mois_affiliation≥10^heure_travail≥150)⇒indemnité_congé_maternité",
"exception":"arrêt maladie longue durée",
"reference":"LR01-101"}}

```

Figure 2: Exemple des règles logiques générées

6. Conclusion et perspectives

Nous avons présenté une démarche structurée en exploitant des données juridiques et réglementaires de la sécurité sociale française pour le développement d'un chatbot d'assistance. Une particularité majeure de ce corpus est son hétérogénéité, puisqu'il réunit des PDF, des textes tabulaires et bruts, ce qui exige un pré-traitement robuste pour la segmentation, l'extraction et la représentation textuelle. Malgré ces difficultés, plusieurs analyses ont été menées avec succès, notamment la classification thématique des questions par apprentissage automatique, le regroupement thématique des fiches techniques, l'extraction terminologique et l'analyse de similarité sémantique, ainsi que l'extraction des règles des fiches techniques. Sur cette base, nous avons développé un premier prototype fonctionnel basé sur la recherche de questions proches dans la base historique Question-Réponse. Lorsqu'une question proche est identifiée grâce à un score composite pondéré, la réponse associée est transmise à Mistral 7B pour reformuler une réponse adaptée à la question posée. En cas de faible similarité, le système active le module de recherche basé sur des règles (en développement). Les résultats montrent que la méthode proposée offre une base solide pour un chatbot hybride alliant recherche sémantique, réutilisation des connaissances historiques et raisonnement déterministe. Toutefois, il montre plusieurs limites. La précision des réponses dépend de la qualité des passages retrouvés. Le prototype repère parfois les questions contextuellement proches, liées au profil de l'assuré, alors que leurs intentions diffèrent. Ainsi, deux questions similaires en apparence peuvent relever de problématiques distinctes, surtout lorsqu'elles sont longues, ce qui peut fausser les réponses.

Plusieurs perspectives se dégagent de ce travail: (1) compléter le module de raisonnement pour activer les règles logiques lorsqu'aucune question similaire n'est trouvée dans la base historique; (2) affiner le module de recherche dans la base historique grâce à une meilleure identification des intentions; (3) réaliser l'évaluation quantitative des différents modules de la chaîne et valider les performances du système, y compris son impact opérationnel à grande échelle.

References

- Alrahabi M., Mansour N. B., and Rahimi H. (2025). Extraction de mots-clés à partir d'articles scientifiques: comparaison entre modèles traditionnels et modèles de langue. In *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2: traductions d'articles publiés*, pp. 15–27.
- Aussenac-Gilles N., Bouché-Pillon J., Zaraté P., Chevalier Y., and Gicquel P.-Y. (2023). Éléments d'état de l'art sur l'extraction et la modélisation de règles formelles à partir de textes légaux. In *34es Journées francophones d'Ingénierie des Connaissances (IC 2023)@ PFIA 2023*, pp. 79–84. AFIA.

- Beauchemin D., Khoury R., and Gagnon Z. (2024). Quebec automobile insurance question-answering with retrieval-augmented generation. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pp. 48–60.
- Beyneix I. and Rovinski A. (2025). L'intégration de l'IA générative: quels impacts pour les juristes? *Entreprendre & Innover*, 64(1):24–33.
- Billi M., Pisano G., and Sanchi M. (2024). Fighting the knowledge representation bottleneck with large language models. In *Legal Knowledge and Information Systems*, pp. 14–24. IOS Press.
- Bojanowski P., Grave E., Joulin A., and Mikolov T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.
- Feillet P. (2024). Grands modèles de langage (large language models) et règles logiques pour une automatisation décisionnelle avancée. In *10^{ème} Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle*, volume 1, pp. 115–120.
- Gémar J.-C. (2002). Le plus et le moins-disant culturel du texte juridique. langue, culture et équivalence. *Meta*, 47(2):163–176.
- Jourdain L., Hellal S., and Marini T. (2024). Why just search when you can expand ? Enhancing RAG with Query Expansion Strategies and Document Aggregation.
- Liu Y. (2024). The application of rag technology in traditional chinese medicine. In *2024 international conference on artificial intelligence and communication (ICAIC 2024)*, pp. 402–408. Atlantis press.
- Mahmoudi S. A., Condevaux C., Mathis B., Zambrano G., and Mussard S. (2022). Ner sur décisions judiciaires françaises: Camembert. *Extraction et Gestion des Connaissances: EGC'2022*, 38:281.
- Metheniti E., Bhar S., and Asher N. (2025). Une revue sur les hallucinations des llm. In *Actes des 32^{ème} Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1: articles scientifiques originaux*, pp. 756–779.
- Mikolov T., Chen K., Corrado G., and Dean J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mimouni N., Nazarenko A., and Salotti S. (2015). Une ontologie documentaire pour l'accès aux contenus juridiques. In *26^{es} Journées Francophones d'Ingénierie des Connaissances*.
- Miyaji R., Moulin R., Monção S., and Machado L. (2025). Empowering business decisions and knowledge management through advanced rag-driven qa systems. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pp. 55–60. IEEE.
- Mohamed T. N., Pudaruth S., and Coste-Manière I. (2025). A chatbot for the legal sector of mauritius using the retrieval-augmented generation ai framework. *International Journal of Advanced Computer Science & Applications*, 16(2).
- Nayinzira J. P. (2025). *SentimentCareBot: Chatbot à génération augmentée par recherche pour le soutien en santé mentale avec intégration des sentiments*. PhD thesis, Université du Québec à Rimouski.
- Nouali S., Badache I., and Bellot P. (2025). Exploration du rag pour la génération de réponses à des questions en contexte éducatif: étude sur les données sciq. In *Actes de l'atelier Intelligence Artificielle générative et ÉDUCation: Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU)*, pp. 29–41.
- Rémy N., Deschamps F., and Kreckelbergh S. (2024). Construire la confiance des chatbots à base de llm. In *Congrès Lambda Mu 24 «Les métiers du risque: clés de la réindustrialisation et de la transition écologique»*.
- Shetty R. (2025). Enhancing context-aware search with retrieval-augmented generation. *Authorea Preprints*.

L'extraction semi-automatique de termes métaphoriques grâce à la spécificité sémantique : étude appliquée à un corpus français-italien de presse spécialisée dans le domaine de la bijouterie-joaillerie

Marius François¹, Maria Grazia Massimo²

¹Université Lumière Lyon 2 – marius.francois@univ-lyon2.fr

²Université de Naples L'Orientale/Université Lumière Lyon 2 – m.massimo1@unior.it

Abstract

Many studies have demonstrated the importance of metaphorical phenomena for the efficient communication of specialised knowledge (Resche, 2002; Rossi, 2014, 2017, 2021; Temmerman, 2000). For terminologists, the extraction of metaphorical terms, forged through the encounter of “two incompatible notional domains” (Kleiber 1994: 42, our translation), is therefore of paramount a significant issue. Our approach is complementary of those based on metaphorical markers (Ureña Gómez-Moreno and Faber, 2010; Volanschi and Kübler, 2011) or the measurement of the overlap of cooccurents for a given term (Oliveira and Ploux, 2011). The approach relies on a specificity calculation (Lafon, 1980), traditionally applied at the level of lemma frequency (LeS) for terminological extraction purposes (Drouin, 2003). We introduce the consideration of the semantic tag as a complementary unit of analysis. The tags are automatically assigned by PyMUSAS (Moore et al., 2026) according to the semantic field (SF), namely the combination of a notional and a lexical domain. Semantic specificity (LaS) allows to distinguish between exogenous SFs ($LaS < 3.09$, $p < 0.001$) and those endogenous to the domain ($LaS > 3.09$). The study is applied to two specialised jewelry press corpora, one in Italian (562,226 tokens) and one in French (365,686 tokens). Noun and adjective cooccurents of various pivot-lemmas (*bague*, *anello*, etc.) were extracted using three methods: an unsupervised approach considering exogenous SFs, one targeting exogenous SFs with manual partial semantic reattribution of tags (PSR), and one combining PSR with the inclusion of both exogenous and endogenous SFs ($LeS/LaS > 3.09$). All three approaches were implemented with and without statistical association measures (SAMs) and subsequently evaluated against a manual goal standard. The third approach, despite relatively low precision in both languages ($\bar{P}_{fr} = 0.24$; $\bar{P}_{it} = 0.2$), achieves the highest recall scores ($\bar{R}_{fr} = 0.52$; $\bar{R}_{it} = 0.47$), especially without SAMs. This provides terminologists with “corpus grounding” (L’Homme, 2005, our translation), despite significant noise. Notwithstanding OCR or semantic/grammatical tagging, it allows for the retrieval of both exogenous metaphors and those integrated into the domain.

Keywords: semantic specificity – semi-automatic extraction – jewels – metaphorical term.

Résumé

De nombreuses études démontrent l'importance de la métaphore pour la communication efficace de connaissances spécialisées (Resche, 2002 ; Rossi, 2014, 2017, 2021 ; Temmerman, 2000). Pour les terminologues, l'extraction de termes métaphoriques issus de la rencontre entre « deux domaines notionnels incompatibles » (Kleiber, 1994 : 42) constitue un enjeu majeur. Notre approche s'inscrit en complément des travaux basés sur des marqueurs lexicaux (Ureña Gómez-Moreno et Faber, 2010, Volanschi et Kübler, 2011) ou sur la mesure du recouvrement des cooccurents d'un terme donné (Oliveira et Ploux, 2011).

La méthode repose sur un calcul de spécificité (Lafon, 1980), traditionnellement mobilisé au niveau de la fréquence des lemmes (*LeS*) à des fins d'extraction terminologique (Drouin, 2003). Nous considérons ici l'étiquette sémantique comme unité d'analyse complémentaire. Ces étiquettes sont attribuées automatiquement par

PymUSAS (Moore et al., 2026) en fonction du champ sémantique (CS), soit la combinaison d'un domaine notionnel et lexical. La spécificité sémantique (*LaS*) permet de distinguer les CS exogènes ($LaS < 3,09$, $p < 0,001$) des CS endogènes au domaine ($LaS > 3,09$).

L'étude est appliquée à deux corpus de presse spécialisée en bijouterie-joaillerie, en italien (562 226 tokens) et en français (365 686 tokens). Les cooccurents nominaux ou adjectivaux de différents lemmes-pivots (*bague*, *anello*, etc.) ont été extraits selon trois méthodes : une approche non supervisée considérant les CS exogènes, une ciblant les CS exogènes mais avec réétiquetage sémantique partiel (RSP) manuel, et une combinant RSP et considération des CS exogènes comme endogènes ($LeS/LaS > 3,09$). Les trois approches considèrent, ou non, des mesures d'association statistique (MAS) et sont évaluées en regard d'un standard manuel. La troisième approche obtient, en dépit d'une précision relativement basse dans les deux langues ($\bar{P}_{fr} = 0,24$; $\bar{P}_{it} = 0,2$), les scores de rappel les plus élevés ($\bar{R}_{fr} = 0,52$; $\bar{R}_{it} = 0,47$), particulièrement sans MAS. Elle offre ainsi au terminologue un « ancrage dans le corpus » (L'Homme, 2005), en dépit d'un bruit non négligeable. Malgré les limites liées à l'océrisation ou à l'étiquetage sémantico-grammatical, elle inclut tant les métaphores exogènes que celles intégrées au domaine.

Mots clés : spécificité sémantique – extraction semi-automatique – bijoux – terme métaphorique.

1. Introduction

Le phénomène de la métaphore dans les terminologies des domaines spécialisés a été largement étudié, notamment par Assal (1994), ainsi qu'au cours des vingt dernières années (Resche, 2002 ; Rossi, 2014, 2017, 2021 ; Temmerman, 2000). En particulier, nous citons le recueil de communications de Dury et al. (2009), qui vise l'analyse des aspects métaphoriques dans divers domaines, selon des approches à la fois théoriques et pratiques. De plus, la diversité des typologies métaphoriques soulève des enjeux épistémologiques concernant la motivation des termes, leur traduisibilité et leur production au sein d'un contexte synchronique ou diachronique (Rossi, 2021 : 170).

Dans le cadre de notre étude, nous nous intéressons à la métaphore descriptive à fonction dénominateur qui, d'après de nombreux chercheurs, parmi lesquels Humbley (2018), Rossi (2021), Prandi (2025) et Zanella (2024), correspond à une typologie de métaphore précise et couramment mobilisée, à savoir la catachrèse isolée sur la base d'une analogie formelle. De son côté, Oliveira (2005), à la lumière de ses études sur l'identification de la métaphore terminologique, définit cette typologie comme une métaphore d'assimilation, « car elle rallie de manière explicite deux univers différents tout en supposant un certain parallélisme entre eux » (Oliveira, 2005). À titre d'exemple, citons *vanne papillon* et *tige solaire* (Prandi et Rossi, 2012 : 15) ou *anneau mitral* (Oliveira, 2005), dont la relation métaphorique est issue de la rencontre entre « deux domaines notionnels incompatibles » (Kleiber, 1994 : 42).

Pour cette recherche, nous analyserons la métaphore dans la terminologie de la bijouterie-joaillerie, s'inscrivant dans le domaine de l'orfèvrerie, un art à la structure complexe pour le travail des métaux précieux destinés à plusieurs services, tels que les arts de la table et de la décoration, et qui mobilise de nombreux sous-domaines (la minéralogie, la chimie et la technologie). Toujours sur le plan des savoir-faire, nous considérons la bijouterie et la joaillerie comme des domaines à part entière. Par ailleurs, nous entendons le domaine comme « un ensemble de procédés [...] destinés à produire certains résultats » (de Bessé, 2000 : 184), et, en suivant les réflexions de Dury (2024), comme un champ de connaissance d'activités spécialisées.

L'objectif de notre recherche est de proposer une méthode semi-automatique d'extraction statistique afin d'offrir au terminologue un ancrage dans ses données capable de générer un gain de temps relativement conséquent pour le relevé des termes métaphoriques en corpus.

2. La détection des termes métaphoriques : état de l'art

La question de l'extraction de métaphores spécialisées a fait l'objet d'approches multiples. Nous citerons ici les travaux de Volanschi et Kübler (2011) qui, dans le domaine de la biologie, mobilisent un ensemble de marqueurs textuels de métaphoricité tels que des guillemets ou des occurrences d'expressions comme *so-called*, *which we call*, etc. De plus, les chercheuses ont extrait 600 termes avec au moins 10 occurrences, dont ont ensuite été conservés ceux pour lesquels existait une tension sémantique, soit un usage inattendu d'une expression dans un contexte précis. Ureña Gómez-Moreno et Faber (2010) mobilisent également des marqueurs textuels et se penchent sur les cooccurrences de termes fréquents du domaine de la biologie marine, jugés plus susceptibles de posséder des cooccurrents métaphoriques (CM), comme *fish*, à partir duquel on pourra observer *lanternfish* ou *surgeonfish*. Nous citerons également les travaux d'Oliveira et Ploux (2011), lesquelles déterminent, via le modèle géométrique ACOM, une mesure de distance sémantique établie via le taux de recouvrement des cooccurrents de termes. Ainsi, un terme comme *fruit civil*, étant donné que *fruit* et *civil* n'apparaissent presque jamais dans des contextes similaires, constitue un bon candidat au statut métaphorique.

Dans notre cas, nous aurons recours à la spécificité (Lafon, 1980), un indice statistique mobilisé, notamment, à des fins d'extraction terminologique par Drouin (2003). En comparant les fréquences d'un lemme donné dans deux ensembles de données, en l'occurrence un ensemble spécialisé et un ensemble de langue générale, elle permet de distinguer les lemmes sur-/sous-représentés. Une valeur élevée, généralement au-dessus d'un seuil fixé à 3,09 ($p < 0,001$), constitue dès lors un « indice fort du statut terminologique » (L'Homme, 2005 : 1124). La spécificité sémantique, *LaS*, que nous considérons, diffère de cette première spécificité « lemmatique », *LeS*, par l'unité d'analyse qu'elle mobilise, à savoir l'étiquette sémantique. Cette dernière situe le lemme auquel elle est rattachée dans un champ sémantique, soit un groupement de mots liés au même « concept mental » (Archer et al., 2002 : 1, notre traduction) et par des relations paradigmatiques telles que l'hyponymie, l'hyponymie, la synonymie, etc. (*ibid*). Ainsi, des lemmes comme *piano* ou *clarinette* sont susceptibles d'être regroupés sous le champ sémantique « Musique », représenté par une étiquette précise. La *LaS* mesure la sur-/sous-représentation d'un champ sémantique entre un corpus d'analyse et un corpus de comparaison. De fait, notre approche s'inspire des travaux de Rayson et Garside (2000) ou Rayson (2008), lesquels mesurent statistiquement l'écart entre la fréquence attendue d'une étiquette sémantique et sa fréquence réelle, via la comparaison d'un corpus d'analyse et d'un corpus de langue générale, afin de déterminer les champs sémantiques saillants des corpus d'analyse considérés.

3. Méthodologie

Grâce à l'outil *PyMUSAS*¹ (Moore et al., 2026), chaque lemme d'un corpus donné est attribué à une étiquette sémantique. Toutes les étiquettes sont constituées d'une combinaison lettre + chiffre. L'outil propose 232 classifications différentes établies à partir du *Longman Lexicon of Contemporary English*. Ces dernières sont appliquées à un lexique anglais, lequel a été traduit dans d'autres langues, dont le français et l'italien. Chaque élément du lexique est classifié à partir de la désambiguïsation de ses acceptions possibles, réalisée à partir de plusieurs filtres, parmi lesquels l'étiquetage grammatical, qui permet par exemple de distinguer *ferme* (Adj.), relatif à la consistance, de *ferme* (N.), relatif à l'exploitation agricole, ou encore la

¹ <https://ucrel.github.io/pymusas/>

fréquence d'utilisation avec un sens précis : *bleu* sera vraisemblablement utilisé plus souvent pour faire référence à la couleur qu'au néophyte (Rayson et al., 2004).

Dans le cadre de ce travail, nous appliquons le calcul de spécificité à partir du décompte des étiquettes sémantiques précédemment évoquées dans les différents corpus présentés *infra*. Nous extrairons ensuite, avec ou sans mesure d'association statistique (MAS) – *Pointwise Mutual Information* (PMI; Church et Hanks, 1990) et *log-likelihood* (LL; Dunning, 1993) – les cooccurents directs des deux lemmes-pivots (LP) français, *bracelet* et *bague*, et deux LP italiens, *bracciale* et *anello*, correspondant à des adjectifs ou à des noms. Ces différents LP font partie des éléments principaux de la parure féminine, dont l'usage remonte à l'Antiquité. À l'inverse de la plupart des approches, qui considèrent exclusivement la sur représentation, nous proposons ici de considérer les étiquettes sémantiques sous-représentées ($LaS < 3,09$), afin de circonscrire les différents cooccurents en principe exogènes au domaine de la bijouterie-joaillerie (approche n°1).

Nous réaliserons, en complément de cette première extraction, un réétiquetage sémantique partiel (RSP), afin de parer à d'éventuelles erreurs d'étiquetage sémantique par l'outil (approche n°2). Par partiel, nous entendons limité aux lemmes identifiés comme CM dans le standard, ainsi qu' à ceux ayant au moins 50 occurrences dans leur corpus d'analyse respectifs. Il aurait été plus intéressant d'envisager un réétiquetage complet. Cela constitue toutefois un investissement trop chronophage à l'échelle de nos moyens actuels, notamment au vu de la quantité de lemmes différents dans nos deux corpus (16 369 en français, 30 557 en italien).

Pour notre étude, nous disposons de deux corpus d'analyse, un en français (421 023 tokens)² et un en italien (638 248 tokens)³, issus de la presse spécialisée « technique et professionnelle » (Balle, 2006) du domaine, ou magazines professionnels (Laffont et Petit, 2007 ; Millot, 2023), ainsi que de deux corpus de comparaison de langue générale pour les deux langues (2 031 526 tokens⁴ pour le français et 1 980 791 tokens⁵ pour l'italien). L'ensemble des corpus a été lemmatisé et étiqueté grammaticalement. La construction des corpus d'analyse a été motivée par le fait que les magazines qui les composent constituent un observatoire privilégié de la terminologie des bijoux féminins.

Bien que le corpus soit construit en diachronie, nous procédons dans cette étude à une analyse synchronique des cooccurrences précédemment évoquées, suivies d'un nom et d'un adjectif, isolées à l'aide du logiciel de textométrie TXM⁶ (Heiden et al., 2010). Afin d'aboutir à un standard pour évaluation, nous avons filtré manuellement ces premiers résultats en ne conservant que les résultats métaphoriques et, dans une deuxième phase, nous avons retiré du standard les lemmes qui n'avaient pas été étiquetés correctement du point de vue grammatical par notre outil (voir Tableau n°2).

De plus, comme nos résultats et notre relevé manuel le montreront, cette approche ne permet que d'extraire la partie non-saillante des cooccurents métaphoriques. Le recours à certains champs sémantiques (CS) *a priori* exogènes s'est en réalité pérennisé ($LaS > 3,09$). Nous

² Composé de 77 numéros de *L'Officiel – Horlogerie et bijouterie*, publiés entre 2008 et 2023.

³ Composé de 76 numéros de *Preziosa Magazine*, publiés entre 2008 et 2023.

⁴ <https://wortschatz.uni-leipzig.de/en/download/fra>

⁵ <https://wortschatz.uni-leipzig.de/en/download/ita>

⁶ <https://txm.gitpages.huma-num.fr/textometrie/>

insistons ici sur le *a priori*, la sur-représentation d'un CS pouvant s'expliquer par une utilisation métaphorique « massive » d'un CS donné, par exemple, le CS des végétaux (L3), dans *bague trèfle*, *bague fleur* ou *anello tulipano*, etc., comme par le recours à un CS à des fins non métaphoriques, comme c'est le cas du CS lié à l'anatomie (B1). Ici, le domaine de la bijouterie-joaillerie porte majoritairement sur des objets portés à différents endroits du corps : *poignet*, *cou*, *main*, etc. La *LaS* de B1 pourrait donc, en principe, être supérieure à notre seuil, ce qui aurait pour conséquence d'invisibiliser certains termes métaphoriques comme *bracelet œil*, *bracelet cœur* ou encore *anello cuore*.

Il convient donc de considérer également les lemmes attribués à ces CS sur-représentés et nous proposons, afin de réduire le bruit généré par cette extraction plus large, d'introduire une double condition de sur-représentation au niveau lemmatique ($LeS > 3,09$) comme sémantique ($LaS > 3,09$). Cela constituera notre approche n°3.

4. Résultats

La section qui suit est consacrée aux résultats obtenus via l'application de nos trois différentes approches. Chacune a été évaluée via les métriques d'évaluation classique de rappel (*R*), précision (*P*) et F-mesure (*F1*), calculées à partir de la comparaison avec le standard.

Tableau n°1 : Résultats des différentes extractions terminologiques.

LP	MAS	R_1	P_1	$F1_1$	R_2	P_2	$F1_2$	R_3	P_3	$F1_3$
<i>bracelet</i>	PMI	0,23	0,58	0,33	0,35	0,21	0,26	0,61	0,28	0,38
<i>bracelet</i>	LL	0,1	0,6	0,17	0,19	0,19	0,19	0,29	0,23	0,26
<i>bracelet</i>	Ø	0,32	0,46	0,38	0,39	0,17	0,24	0,81	0,22	0,35
<i>bague</i>	PMI	0,12	0,5	0,19	0,28	0,17	0,21	0,44	0,21	0,28
<i>bague</i>	LL	0,09	0,57	0,16	0,23	0,25	0,24	0,3	0,23	0,26
<i>bague</i>	Ø	0,26	0,48	0,34	0,33	0,17	0,22	0,67	0,22	0,33
<i>bracciale</i>	PMI	0,23	0,19	0,21	0,08	0,08	0,08	0,69	0,24	0,36
<i>bracciale</i>	LL	0,08	0,14	0,1	0,08	0,14	0,1	0,46	0,27	0,34
<i>bracciale</i>	Ø	0,23	0,11	0,15	0,08	0,06	0,06	0,77	0,14	0,24
<i>anello</i>	PMI	0,15	0,29	0,2	0,15	0,33	0,21	0,27	0,19	0,22
<i>anello</i>	LL	0,08	0,22	0,11	0,12	0,38	0,18	0,15	0,18	0,17
<i>anello</i>	Ø	0,15	0,17	0,16	0,19	0,21	0,2	0,5	0,18	0,26

5. Discussion

À partir de la comparaison avec les résultats de la méthode n°1, un premier constat peut être dressé que la réattribution partielle des étiquettes sémantiques, en français comme en italien, produit des extractions moins bonnes (les F-mesures sont en moyenne inférieures de 21 % pour le français, 4,7 % pour l'italien). Le rappel, en français, est toutefois plus élevé (+80 % en moyenne), et légèrement moins bon pour l'italien (-8,9 % en moyenne). Le réétiquetage partiel a pour effet de faire drastiquement baisser la précision en français (-64 % en moyenne), et ne change rien pour l'italien (+1,1 % en moyenne).

Outre les F-mesures globalement meilleures, en français (+27,8 % en moyenne) comme en italien (+83 % en moyenne), ce sont davantage les mesures de rappel, dans les deux langues (en moyenne +194 % pour le français, +218 % pour l'italien), qui soulignent l'apport de la considération des lemmes avec $LeS/LaS > 3,09$, en regard de l'approche initiale. Le postulat

relatif au fait que la métaphore terminologique se constitue l'apanage de l'exogénéité sémantique est donc mis à mal. Les cooccurents métaphoriques semblent être répartis en deux catégories : les CS exogènes, qui constituent une minorité du genre, et les CS endogènes, majoritaires.

Enfin, sur l'utilisation d'une métrique d'association, les résultats indiquent, comme on pouvait s'y attendre, que le fait d'inclure une exigence statistique d'association restreint la quantité de CM correctement extraits au profit de la précision : $\bar{P}_{MAS_{fr}} = 0,34$ contre $\bar{P}_{\emptyset_{fr}} = 0,29$ et $\bar{P}_{MAS_{it}} = 0,22$ contre $\bar{P}_{\emptyset_{it}} = 0,15$. Du reste, les valeurs de rappel sont en moyenne plus élevées : $\bar{R}_{MAS_{fr}} = 0,41$ contre $\bar{R}_{\emptyset_{fr}} = 0,74$ et $\bar{R}_{MAS_{it}} = 0,39$ contre $\bar{R}_{\emptyset_{it}} = 0,64$. Nous nous inscrivons dans une démarche qui vise un ancrage dans le corpus à des fins subséquentes d'analyse terminologique. Notre approche suit donc les recommandations de L'Homme (2020) à ce sujet :

Du point de vue de l'utilisateur, il est préférable de réduire le silence au maximum, quitte à augmenter le bruit de façon importante : il est plus facile d'épurer une liste contenant des candidats non souhaités que de retrouver un terme omis dans un corpus parfois volumineux (*ibid* : 217)

Sans exclure de notre considération l'ensemble des métriques d'évaluation, nous accorderons une attention toute particulière au rappel et utiliserons cet argument afin de trancher contre l'utilisation d'une MAS.

Du reste, afin d'illustrer le gain de temps pour l'utilisateur, notons que, pour les tests en français, environ 43 % des cooccurents de nos pivots sont exclus pour, *in fine*, obtenir trois quarts des CM recherchés. Pour l'italien, le tri est plus important, 60 % en moyenne pour environ 65 % des CM identifiés, comme récapitulé dans le tableau suivant :

Tableau n°2 : Bilan de l'extraction avec RSP et considération des lemmes avec LeS/LaS sans MAS du point de vue du terminologue.

LP	CM N./Adj. (total)	CM (standard)	Cooccurents (extraction)	Proportion évacuée	CM obtenus
<i>bracelet</i>	163	31	92	44 %	21 (68 %)
<i>bague</i>	183	43	107	42 %	29 (81 %)
<i>bracciale</i>	177	13	70	60 %	10 (77 %)
<i>anello</i>	190	26	74	61 %	13 (50 %)

Si l'objectif est d'épargner au terminologue un nombre trop conséquent d'allers-retours entre les listes de résultats et les données, nous poserons ici que notre approche ne prétend pas à la perfection et à l'exhaustivité, mais à une quantité qui « peut suffire pour répondre aux besoins de la recherche considérée » (Pincemin, 2025 : §24).

Nous noterons en premier lieu des erreurs d'océrisation du corpus d'analyse. Par exemple, le lemme *trèfle*, cooccurent métaphorique de *bague*, a été océrisé dans le corpus en trois graphies distinctes : *trefle*, *tréfle* et *trèfle*. L'unique occurrence de *bague* avec un de ces cooccurents a considéré la graphie *tréfle*. Ce dernier possède une LaS égale à 50,49 pour une LeS égale à 0,82 et n'a donc pas pu être détecté par notre méthode. Si ces différentes formes avaient été mutualisées, la LeS aurait alors atteint 23,49 et le TM aurait été extrait.

En second lieu, nous constatons, à plusieurs reprises, un étiquetage grammatical erroné. Ainsi, des lemmes comme *araignée* ou *schiaiva* ont été reconnus avec une forme nominale et une forme verbale. Cela ne semble toutefois pas avoir eu d'incidence trop importante sur les valeurs de LeS déterminées et donc sur l'extraction.

Notons également, comme nous l'évoquions *supra*, certains problèmes d'étiquetage sémantique. Une majorité peut être imputée à des modèles d'étiquetage peu performants pour l'italien (0,5 MB) comme pour le français (0,08 MB). Toutefois, nous ne saurions reprocher à l'outil un étiquetage erroné de lemmes empruntés à des langues différentes, plus ou moins lexicalisés. À cet effet, nous remarquons par exemple *snake*, cooccurrent métaphorique de *bague*, ou encore *cuff*, pour *bracciale*. Par ailleurs, notons que les CM identifiés dans le standard manuel pour l'italien incluent davantage d'emprunts à l'anglais (*cuff*, *tennis*) ou au français (*animalier*, *chevalière*) que les CM des pivots français. Cela pourrait partiellement expliquer les résultats statistiquement moins bons pour l'italien. Les dysfonctionnements de *PyMUSAS* peuvent toutefois être partiellement corrigés par le linguiste. Ce dernier devra veiller, via un recours au CA, à un traitement rigoureux de l'homographie, voire de l'ambiguïté sémantique. Dans le premier cas, prenons l'exemple du lemme *morse* (N.), CM de *bracelet*. Ici, la désambiguïsation par la catégorie grammaticale ne suffit pas pour trancher entre analogie au mammifère ou au code/alphabet. S'il s'agit du premier, il devra être rattaché à l'étiquette L2 avec une *LaS* de 24,24 qui ne sera pas compensée par une *LeS* suffisamment élevée (1,63). S'il s'agit du second, cela correspond à l'étiquette Q3, dont la *LaS* a été déterminée à 2,16 et le lemme sera correctement extrait. C'est le recours au CA et, étant donné la nature de notre corpus, à l'image, qui permet de trancher en faveur du second. En ce qui concerne l'ambiguïté sémantique, considérons la classification de *croce*. Le lemme peut tout à fait être classé dans deux CS différents : O2 (Objets général, *LaS* = 8,9) ou S9 (Religion et surnaturel, *LaS* = 9,95). Nous recommandons ici de le classer dans O2 pour s'en tenir à un usage littéral, non connoté, à partir duquel nous pourrions définir un usage métaphorique, comme illustré dans les travaux du groupe de Pragglejaz (2007). Dans ce cas précis, l'incidence ne sera que relative, le lemme étant quoi qu'il arrive extrait du fait de sa valeur *LeS* égale à 5,77.

6. Conclusion

La méthode présentée s'appuie sur l'hypothèse que la métaphore s'établit, en contexte spécialisé, par le rapprochement de deux domaines *a priori* incompatibles. Elle comprend une comparaison d'un corpus spécialisé et d'un corpus de langue générale, en français et en italien. Celle-ci s'établit à partir de deux unités d'analyse : la fréquence d'un lemme donné, classique, et celle plus originale de l'étiquette sémantique attribuée par l'outil *PyMUSAS* et permet, dans les deux cas, de quantifier une sur-/sous-représentation. Quatre lemmes-pivots ont été considérés, deux dans chaque langue, et un relevé manuel de leurs cooccurrents métaphoriques correspondant à des adjectifs ou des noms a été réalisé et érigé en standard. Nous avons ensuite extrait l'ensemble des cooccurrents correspondant à des noms ou à des adjectifs dont la *LaS* est inférieure à 3,09 et dont les valeurs *LeS/LaS* sont supérieures à 3,09. Ainsi, nous postulons l'extraction de TM construits à partir d'un domaine alors exogène à celui de la bijouterie-joaillerie (Nourriture avec *bague cake*, Boissons avec *bague cocktail*, etc.) comme ceux construits à partir de domaines *a priori* exogènes mais qui ont, en raison d'un usage métaphorique conséquent (Végétaux avec *bague fleur*, Êtres vivants avec *bracelet papillon*, etc.) ou d'un usage littéral tout aussi conséquent (Anatomie lié à l'utilisation de termes relatifs à l'emplacement où sont portés les bijoux), été intégrés aux CS constitutifs du domaine. La méthode s'inscrit à la suite d'autres et, en cela, rencontre des problèmes similaires relatifs à l'outillage (lemmatisation et étiquetage grammatical erronés [Drouin, 2002]), ou relatifs aux données (extraction à base statistique moins efficace pour les lemmes à faible fréquence [L'Homme, 2020]). Sur la question des problèmes liés aux lemmes à faible fréquence, notons que notre méthode, puisqu'elle extrait en fonction de deux conditions de spécificité distinctes,

a permis d'isoler des lemmes à faible fréquence candidats au statut terminologique et, en l'occurrence, à la métaphoricité, tels que *cake*, qui ne totalise que deux occurrences. Nous contournons donc partiellement le problème lié à l'extraction d'éléments linguistiques employés sporadiquement, mais mobilisés avec un sens spécialisé (Drouin, 2004). Quoi qu'il en soit, la méthode permet un « ancrage dans le corpus » (L'Homme, 2005 : 1124) pour quiconque s'intéresse aux questions métaphoriques en langue de spécialité.

7. Quelques pistes de réflexion

Un travail subséquent à celui-ci consistera en une réflexion sur les possibilités d'évacuation du bruit. Une première étape pourrait consister à retirer les étiquettes Z99, qui correspondent à la catégorie « Non reconnu », qui abondent dans nos résultats. En français et sans MAS, la précision est presque doublée pour *bracelet* et *bague* (0,41 pour les deux). Nous ne notons toutefois pas d'incidence sur la précision pour l'italien.

Nous posons l'hypothèse que la considération de la *LaS* peut également s'appliquer en diachronie : l'évolution de la *LaS* d'un CS se constituerait alors « indice » (Dury et Picton, 2009 : 36) supplémentaire au moyen duquel il serait possible de postuler des évolutions terminologiques au sein d'un domaine. La progression de la *LaS* d'un CS entre un moment T-1 et un moment T0 pourrait indiquer l'incorporation d'un CS au domaine, représentée potentiellement par un foisonnement néologique spécialisé. Puisque nous avons travaillé ici à partir d'un lemme pivot, il sera également intéressant de tester notre approche à l'échelle d'un corpus sans considérer de pivot.

De plus, nous gageons que la *LaS* peut, quoi qu'il en soit, permettre de cartographier les différents CS du domaine. Si nous avons initialement considéré les domaines les moins saillants, nous ne pouvons que constater que ce qui est saillant est tout aussi pertinent. À ce titre, Prandi et Rossi (2012) constatent l'existence, dans plusieurs domaines, d'isotopies, des projections collectives du type « LE VIN EST UNE PERSONNE », à l'origine, du fait de l'anthropomorphisation induite, d'un foisonnement terminologique important ; pourront alors être évoqués le *corps* ou encore le *nez* du vin. Posons ici l'hypothèse que le CS associé, dans ce cas B1 (Anatomie et physiologie) est susceptible d'avoir, dans un corpus du domaine de la dégustation de vin, une *LaS* élevée. Le cas échéant, il serait rendu possible d'identifier les isotopies d'autres domaines. Cela inclurait celles dont les faibles valeurs de *LaS*, partagées par un nombre restreint de locuteurs, peuvent être à l'origine d'une dynamique de néologie sémantique spécialisée. De plus, notre approche permet d'identifier des CS à l'origine de CM, dont nous posons qu'ils peuvent engendrer un foisonnement métaphorique plus ample : la prévalence de CM relatifs à l'étiquette L3 (Végétaux) avec *bague* + *fleur/poire* ou encore *anello* + *fiore*, incite à considérer les résultats de LP + lemme (L3, N./Adj.). Cela permettra de constater l'existence de TM, tels que *bague trèfle* ou encore *anello quadrifoglio*, que l'extraction initiale n'avait pas inclus.

Du reste, nous travaillons d'ores et déjà sur la mise à disposition d'un outil permettant d'appliquer nos différentes approches.

Références

- Altmanova J. (2019). *Dictionnaire de l'orfèvrerie français/italien. Termes, cultures, traditions*. Loiralab.
- Archer D., Wilson A. and Rayson P. (2002). Introduction to the USAS category system. https://ucrel.lancs.ac.uk/usas/usas_guide.pdf

- Assal A. (1994). La métaphorisation terminologique. *Terminologie et traduction*, 2: 235-242.
- Balle F. (2006). *Lexique d'information communication*. Dalloz.
- Church K. and Hanks P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1): 22-29.
- De Bessé B. (2000). Le domaine. In Béjoint H. and Thoiron P., editors, *Le sens en terminologie*, Presses Universitaires de Lyon, pp. 182-197.
- Dressen A., Heuzé M. and Lignel B. (2017). *Médusa, bijoux et tabous*. Musée d'Art Moderne de la ville de Paris, Éditions Paris-Musées.
- Drouin P. (2002). Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés. *Thèse de doctorat*, Université de Montréal.
- Drouin P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1): 99-116.
- Drouin P. (2004). Spécificités lexicales et acquisition de la terminologie. In Purnelle G., Fairon C and Dister A., editors, *Le poids des mots - Actes des 7èmes journées internationales d'analyse statistique - JADT*, pp. 345-352.
- Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74.
- Dury P., Maniez F., Arlin N. and Rougemont C. (2009). *La métaphore en langue de spécialité*. Travaux du CRTT, Presses Universitaires de Grenoble.
- Dury P. and Picton A. (2009). Terminologie et diachronie : vers une réconciliation théorique et méthodologique ?. *Revue française de linguistique appliquée*, 13(2): 31-41.
- Dury P. (2024). Variations diachroniques dans le domaine de l'environnement, en français et en anglais, une étude basée sur corpus. *Repères DoRiF*, 30.
- Heiden S., Magué J.-P. and Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In Bolasco S., Chiari I. and Giuliano L., editors, *10th International Conference on the Statistical Analysis of Textual Data - JADT*, pp. 1021-1032.
- Humbley J. (2018). *La néologie terminologique*. La Lexicothèque, Éditions Lambert-Lucas.
- Kleiber G. (1994). Métaphore : le problème de la déviance, *Langue Française, les figures de rhétorique et leur actualité en linguistique*, 101: 35-56.
- Laffont H. and Petit M. (2007). Presse et anglais de spécialité : quelques éléments de caractérisation de la presse professionnelle. *Les cahiers de l'APLIUT*, 26(3): 27-41.
- Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1: 127-165.
- L'Homme M.-C. (2005). Sur la notion de « terme ». *Meta*, 50(4): 1112-1132.
- L'Homme M.-C. (2020). *La terminologie - Principes et techniques, 2e éd. revue et mise à jour*. Les Presses de l'Université de Montréal.
- Millot P. (2023). Le genre de l'article d'expert professionnel : Une étude sur corpus dans le domaine de la documentation en entreprise. In Le Cor G. and Couterut M., editors, *Les genres en anglais de spécialité : définitions, méthodologies d'analyse et retombées pédagogiques*, pp. 89-116.

- Moore A., Rayson P., Archer D., Czerniak T., Knight D., Lal D., Ó Donnchadha G., Ó Meachair M., Piao S., Uí Dhonnchadha E., Vuorinen J., Yabo Y. and Yang X. (2026). Creating a Hybrid Rule and Neural Network Based Semantic Tagger using Silver Standard Data: the PyMUSAS framework for Multilingual Semantic Annotation. [arXiv:2601.09648](https://arxiv.org/abs/2601.09648)
- Oliveira I. (2005). La métaphore terminologique sous un angle cognitif. *Meta*, 50(4), s.p.
- Oliveira I. and Ploux S. (2011). Vers une méthode de détection et de traitement automatique de la métaphore. In Van Campenhoudt M., Lino T. and Costa R., editors, *Actes des 8e Journées scientifique du Réseau de chercheurs Lexicologie, terminologie, traduction - Passeurs de mots, passeurs d'espoir*, pp. 429-440.
- Pincemin B. (2025). Apprivoiser le « bruit » en linguistique de corpus : expérience d'une analyse factorielle et propositions. *Corpus*, 26.
- Pragglejaz Group. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1): 1-39.
- Prandi M. and Rossi M. (2012). Les métaphores dans la création de terminologie : quelques perspectives ouvertes. In GLAT, *Terminologie : textes, discours et accès aux savoirs spécialisés*, pp. 7-18.
- Prandi M. (2025). Termini e tecnicismi nei lessici di specialità. *Italiano LinguaDue*, 17(2): 191-200.
- Rayson P. and Garside R. (2000). Comparing Corpora using Frequency Profiling. In Kilgarriff A. and Sardinha T. B., editors, *Proceedings of the workshop on comparing corpora, held in conjunction With the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pp. 1-6.
- Rayson P., Archer D., Piao S. and McEnery T. (2004). The UCREL Semantic Analysis System. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, in association with LREC-04*, European Language Resources Association, pp. 7-12.
- Rayson P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4): 519-549.
- Resche C. (2002). La métaphore en langue spécialisée, entre médiation et contradiction : étude d'une mutation métaphorique en anglais économique. *ASp*, 35-36: 103-119.
- Rossi M. (2014). Métaphores terminologiques : fonctions et statut dans les langues de spécialité. In *SHS Web of Conferences 8*, pp. 713-724.
- Rossi M. (2017). De l'usage des métaphores terminologiques dans les sciences : définition de nouveaux paradigmes et vulgarisation auprès du grand public. In Frassi P. and Tallarico G., editors, *Autrement dit : définir, reformuler, gloser : Hommage à Pierluigi Ligas*, Hermann, pp. 163-184.
- Rossi M. (2021). Termes et métaphores, entre diffusion et orientation des savoirs. *La Linguistique*, 57(1): 153-173.
- Temmerman R. (2000). *Towards New Ways of Terminology Description. The sociocognitive approach. (Terminology and Lexicography Research and Practice)*. John Benjamins Publishing Company.
- Ureña Gómez-Moreno J. M. and Faber P. (2010). Strategies for the Semi-Automatic Retrieval of Metaphorical Terms. *Metaphor and Symbol*, 26(1): 23-52.
- Volanschi A. and Kübler N. (2011). The impact of metaphorical framing on term creation in biology. *Terminology*, 17(2): 198-223.
- Zanella G. (2024). La variation de la métaphore dans la vulgarisation scientifique. Une étude en diachronie courte en français et en italien liée aux trous noirs. *Testi e linguaggi*, 18(1): 170-183.

Utilisation des emprunts à l'anglais de type V-*ing* en espagnol, français et italien : une étude en diachronie courte des corpus du web du Wortschatz de Leipzig

François Maniez¹, Maria-Belén Villar-Díaz², Sidonie Larato³,

¹Université de Lyon – francois.maniez@univ-lyon2.fr

²Université de Lyon – maria-belen.villar-diaz@univ-lyon2.fr

³Université de Lyon – s.joannes-larato@univ-lyon2.fr

Abstract 1

Words beginning with a verb root and ending with the *-ing* morpheme are often borrowed from English in many Indo-European languages, and the growing popularity of the *-ing* morpheme has often been attributed to its nominalizing syntactic function (Picone 1996). Following such borrowings, equivalents formed by using native words are frequently created in the recipient language and sometimes included in standard dictionaries (cf. the case of *brainstorming* and *remue-méninges* in French, Humbley 2008). These neologisms sometimes take the form of hybrid borrowings (*relooking* or *surbooking* in French), or pseudo-Anglicisms such as *mailing*. In their report on the Néoveille project, Cartier et al. (2018) identified 22,475 new word forms (types) in a corpus of 92 million words of the French contemporary press (2015-2017). Among the borrowings (which accounted for 18.19% of all word forms), there were no fewer than 303 lexical items ending in *-ing*, representing 21.3% of all borrowings in the corpus.

Numerous borrowings from English, whether used in the same sense (*shopping*), in a different sense (*footing*), or modified by suffixation (*surbooking*), coexist with officially recommended terms, whose implementation has met with varying degrees of success. For example, the French equivalents proposed for “opinion shopping” (*chalandage de firmes d'audit*) and “treaty shopping” (*chalandage fiscal*) in the Journal Officiel of September 22, 2000, have few occurrences when searched on Google, more than twenty years after being suggested by the CGTN (General Terminology and Neology Committee).

Using the News2020 and News2024 web news corpora from the University of Leipzig (more than 700,000 words for all three languages), we study the use of one hundred of the most frequent *-ing* borrowings in Spanish, French, and Italian in the two corpora.

In quantitative terms, Italian has the highest number of *-ing* loanwords from English (309 per million words), whilst French is relatively less inclined to adopt such loanwords (218), and Spanish even less so (107). These loanwords are often linked to the fields of economics and finance (*banking*, *dumping*, *holding*, *rating*, *trading*), sport (*diving*, *jogging*, *racing*, *rafting*, *trekking*) as well as technology and communications (*e-learning*, *mailing*, *phishing*, *streaming*).

Keywords: anglicism, corpus, *-ing* morpheme, lexicology, loanwords, neology, Romance languages.

Abstract 2

Les mots commençant par une racine verbale et se terminant par le morphème *-ing* occupent une place importante parmi les mots empruntés à l'anglais dans de nombreuses langues indo-européennes, et la popularité croissante du morphème *-ing* a souvent été attribuée (Picone 1996) à sa fonction syntaxique nominalisante. Ces emprunts sont fréquemment suivis de la création d'équivalents à partir de mots natifs de la langue réceptrice, et parfois de leur inclusion dans les dictionnaires standard (cf. le cas de *brainstorming* et *remue-méninges* en français, Humbley 2008). Ces néologismes se présentent parfois sous la forme d'emprunts hybrides (*relooking* ou *surbooking* en français), ou de pseudo-anglicismes tels que *mailing*. Dans leur rapport sur le projet Néoveille, Cartier et al. (2018) ont identifié 22 475 nouvelles formes de mots (types) dans un corpus de 92 millions de mots de français

journalistique contemporain (2015-2017). Parmi les emprunts (qui représentaient 18,19% de tous les mots-formes), il n'y avait pas moins de 303 items lexicaux se terminant par *-ing*, soit 21,3% de tous les emprunts du corpus.

De nombreux emprunts à l'anglais, qu'ils soient utilisés dans le même sens (*shopping*), dans un autre sens (*footing*) ou modifiés par suffixation (*surbooking*) coexistent avec les termes officiellement recommandés, dont l'implantation a connu des succès divers. Par exemple, les équivalents français proposés pour *opinion shopping* (chalandage de firmes d'audit) et *treaty shopping* (chalandage fiscal) dans le Journal Officiel du 22/09/2000 ne comptent que peu d'occurrences lorsque l'on interroge le moteur de recherche Google, plus de vingt ans après avoir été suggérés par la Commission générale de terminologie et de néologie.

À l'aide des corpus de nouvelles du web News2020 et News2024 de l'université de Leipzig (plus de 700 000 mots pour l'ensemble des trois langues), nous étudions l'utilisation de cent de ces emprunts les plus fréquents en espagnol, en français et en italien dans les deux corpus.

D'un point de vue quantitatif, l'italien compte le plus d'emprunts en *-ing* à l'anglais (309 par million de mots), le français étant relativement plus réticent à ce type d'emprunts (218), et l'espagnol encore plus (107). Ces emprunts sont souvent liés au domaine de l'économie et de la finance (*banking*, *dumping*, *holding*, *rating*, *trading*), au sport (*diving*, *jogging*, *racing*, *rafting*, *trekking*) ainsi qu'à la technologie et aux communications (*e-learning*, *mailing*, *phishing*, *streaming*).

Mots clés : anglicisme, corpus, emprunt, langues romanes, lexicologie, morphème *-ing*, néologie.

1. Introduction

De nombreux néologismes créés par l'emprunt de mots anglais dans les langues romanes contiennent un vocable commençant par une base verbale et se terminant par le morphème *-ing*. Très peu d'articles de recherche se sont jusqu'à présent concentrés exclusivement sur les emprunts en *-ing* ; cependant, ce morphème figure toujours parmi les sources les plus productives d'emprunt à l'anglais en français et en italien. Dans leur rapport sur le projet Néoveille, Cartier et al. (2018) ont identifié 22 475 nouvelles formes de mots (types) dans un corpus de 92 millions de mots de français journalistique contemporain (2015-2017). Parmi les emprunts (qui représentaient 18,19% de tous les mots-formes), il n'y avait pas moins de 303 items lexicaux se terminant par *-ing*, soit 21,3% de tous les emprunts du corpus. Les auteurs notent que le morphème *-ing* s'est progressivement imposé dans la langue française depuis plus d'un siècle, à l'origine par le biais de transformations métonymiques dans lesquelles des lieux (*parking*, *camping*, *dancing*) étaient désignés par la troncation d'un mot anglais (*parking lot*, *camping site*, *dancing hall*) alors que les formes anglaises font référence à un processus. Ils remarquent également que le morphème concurrent *-age* limite l'utilisation de la forme *V-ing* à des sports spécifiques (*running*, *beatboxing*, *snorkeling*, *cardiotraining*), à un vocabulaire lié au commerce (*networking*, *packaging*, *branding*, *fact-checking*, *coworking*, *crowdfunding*) ou à des activités "socioculturelles" (*bashing*, *ghosting*, *pet-sitting*) qui n'ont pas d'équivalents synthétiques en français. Les magazines sportifs et féminins sont les deux supports dans lesquels les auteurs enregistrent le plus grand nombre de nouveaux items lexicaux, ce que reflètent deux exemples mentionnés par Gazzardi et Vásquez (2022) dans leur approche taxonomique de l'usage de l'anglais dans les médias italiens, *linea per il grooming* ('ligne de toilettage') et *running* (dans la question *Fai running ?*, utilisée dans le titre d'un article décrivant du matériel de sport à la mode).

Dans son étude de l'influence de l'anglais sur l'italien dans le domaine de l'économie, Musacchio (2005) note que l'utilisation de composés hybrides tels que *agenzie di rating*, *società di outsourcing* ou *trading di derivati* montre qu'en dépit de l'incorporation de ces emprunts en *-ing* en italien, les mécanismes de formation des mots typiques des langues romanes (tels que l'utilisation d'un syntagme prépositionnel) continuent de prévaloir sur le modèle anglais de prémodification qui se fonde sur la juxtaposition des mots. Renner et Fernández-Domínguez

(2015) se concentrent sur les faux anglicismes en français, espagnol et italien et enregistrent un nombre élevé de formes de mots en *-ing* (*dribbling, footing, forcing, franchising, outing*), notant qu'ils pourraient alternativement être classés comme des exemples d'affixation ou d'emprunts directs de formes anglaises qui ont acquis un nouveau sens dans la langue réceptrice.

Ces emprunts stimulent très souvent le processus néologique, puisque de nouveaux mots utilisant le stock lexical de la langue d'accueil sont créés. Toutefois, dans des pays comme la France, certaines institutions nationales ont tenté d'endiguer le flot de ces emprunts en proposant des alternatives par l'intermédiaire de commissions dont les travaux sont financés par le ministère français de la Culture (Commission générale de terminologie et de néologie, ci-après CGTN)¹. Il arrive que l'usage des expressions qui résultent du travail de ces commissions s'impose au fil du temps, auquel cas les dictionnaires finiront par refléter leur adoption par le grand public quelque temps après le début de leur implantation. Par exemple, Humbley (2008) rapporte que seul *brainstorming* était inclus dans la version 1986 du dictionnaire *Petit Robert*, alors que *brainstorming* et *remue-méninges* (le terme recommandé par la CGTN) ont été inclus dans la version 2007 du même dictionnaire avec des renvois mutuels. Les entrées des nouveaux mots proposés par les commissions fournissent généralement une définition ainsi que le mot étranger qu'ils sont censés remplacer. Voici l'entrée du mot *remue-méninges* tirée du *Vocabulaire de l'économie et des finances* publié en 2006 par la CGTN :

remue-méninges, n.m. ♦ **Définition** : Technique de groupe destinée à stimuler l'imagination des participants en vue de leur faire produire le maximum d'idées dans le minimum de temps. ♦ **Équivalent étranger** : brainstorming.

Source : *Journal officiel* du 22 septembre 2000.

De nombreux emprunts à l'anglais, qu'ils soient utilisés dans le même sens (*shopping*), dans un autre sens (*footing*) ou modifiés par suffixation (*surbooking*) coexistent avec les termes officiellement recommandés, dont l'implantation a connu des succès divers. Par exemple, les équivalents français proposés pour *opinion shopping* (chalandage de firmes d'audit) et *treaty shopping* (chalandage fiscal) dans le *Journal Officiel* du 22/09/2000 ne comptent respectivement que 30 et 8 950 occurrences lorsque l'on interroge le moteur de recherche Google, plus de vingt ans après avoir été suggérés par la CGTN.

L'utilisation fréquente du morphème *-ing* dans la néologie des langues romanes a été attribuée à sa fonction nominalisante (Picone, 1996). Il convient toutefois de noter que de nombreuses utilisations de ces emprunts en français sont des formes singulières ou tronquées qui ne sont pas considérées comme acceptables en anglais standard (*jogging* pour *jogging suit*, *parking* pour *parking lot*, *passing* pour *passing shot*). De même, l'adjectif composé anglais *meat-sounding* (présent dans des syntagmes tels que *meat-sounding denominations* ou *meat-sounding labels*) est utilisé comme nom en italien depuis quelques années pour désigner l'utilisation de noms de marque à connotation carnée.

¹ Notons à ce sujet la création en 2015 du groupe « Incipit » à l'Accademia della Crusca, qui vise à proposer des substituts italiens à certains anglicismes dont le remplacement est jugé souhaitable (Rati, 2025). Bachis (2023) évoque le positionnement « libéral », parfois contesté, de la Crusca face à la pénétration des anglicismes. Selon certains linguistes italiens, cette retenue s'explique par le souvenir de la politique fasciste, qui contraignait à l'italianisation des mots et noms propres empruntés (Gobbo, 2017).

2. Objectifs et méthode

L'étude des emprunts en *-ing* dans le corpus Europarl (Maniez, 2014) souffrant de certains biais liés au processus de traduction et de la relative obsolescence du corpus (plus d'une vingtaine d'années), nous avons choisi d'avoir recours à des corpus comparables plus récents issus de la *Leipzig Corpora Collection* (News 2020 et News 2024), corpus du web dont la taille varie entre 150 et 446 millions de mots-formes en fonction de la langue concernée et qui sont composés de nouvelles (au sens de l'anglais *news*) publiées sur le Web pendant les années 2020 et 2024.

Notre objectif était de déterminer laquelle des trois langues romanes choisies compte le plus d'emprunts en *-ing*, et dans quels domaines de spécialité ils sont le plus utilisés. Par l'utilisation de deux corpus de taille comparable pour les années 2020 et 2024, nous souhaitons également observer l'émergence éventuelle de certaines tendances d'usage en diachronie courte.

Dans le corpus Europarl, les mots se terminant par le morphème *-ing* de fréquence supérieure à 5 avaient été extraits du corpus dans les trois langues grâce à l'expression régulière [word= ».*ing »]. Cependant, les corpus unilingues de la *Leipzig Corpora Collection* ne permettent pas ce type de requête sous la forme d'une expression régulière, leur interface n'étant conçue que pour la recherche de mots-formes individuels.

Nous avons donc dans un premier temps eu recours à un autre corpus disponible sur la plateforme OPUS, le corpus Subtitles 2018 (le processus d'extraction de ce type de corpus à partir de sous-titres filmiques et télévisuels est expliqué en détail par Lison et Tiedemann, 2016). Ce corpus nous a permis d'extraire une liste des 200 formes en *-ing* les plus fréquemment employées dans chacune des langues concernées grâce à l'expression régulière mentionnée plus haut (seules les formes utilisées 18 fois dans au moins une des langues ont été conservées). Chacune de ces formes a ensuite fait l'objet d'une requête distincte dans chacun des corpus News 2020 et News 2024 des trois langues de notre étude.

3. Résultats quantitatifs

Le tableau 1 donne les chiffres obtenus dans chacun des corpus News 2020 et News 2024 pour les formes les plus fréquemment employées dans le corpus Subtitles 2018.

Subtitles 2018 (frq absolue)	Forme en <i>-ing</i>	News 2020 (frq. PMW)			News 2024 (frq. PMW)		
		ES	FR	IT	ES	FR	IT
14277	parking	2,2	29,6	0,4	1,46	1,19	0,82
6301	shopping	1,1	4,7	14,5	1,16	0,62	0,80
5313	bowling	0,1	1,0	0,4	1,03	1,01	1,50
4441	smoking	0,1	0,5	1,3	1,48	1,12	1,06
4352	timing	0,2	5,9	0,8	1,55	1,38	1,13
4143	planning	0,1	5,6	0,4	0,61	0,86	0,79
3374	camping	2,1	8,6	0,7	1,47	0,95	1,24
3247	marketing	16,6	20,3	24,8	1,23	1,05	1,00
3033	briefing	0,1	2,0	3,0	0,95	0,87	0,62
2837	streaming	19,2	22,4	73,3	1,06	1,17	0,81
2704	catering	1,1	0,3	2,0	1,21	0,88	0,77
2369	jogging	0,2	3,5	3,3	1,23	0,65	0,25
2065	living	1,1	0,1	0,6	1,58	1,55	1,17
Mots-formes en <i>-ing</i> (PMW)		83,9	204,2	309,1	107,4	217,8	359,8
% augmentation/espagnol :		1	2,43	3,68	1	2,03	3,35

Tableau 1 : Nombre d'occurrences dans chacun des corpus News 2020 et News 2024 des formes en *-ing* les plus fréquentes dans le corpus Subtitles 2018

Afin d'extraire de nos corpus les emprunts les plus spécifiques de chacune des langues de notre étude, nous avons calculé pour chacune des formes un indice que nous avons dénommé Ratio de Fréquence Comparative (RFC). Pour chaque langue, le RFC est le double de la fréquence obtenue dans cette langue, divisé par la somme des fréquences obtenues dans les deux autres langues. Ainsi, une forme deux fois plus utilisée en moyenne dans une langue que dans chacune des deux autres aura un RFC de valeur 2, celui-ci étant de valeur 1 pour une forme qui serait en moyenne autant utilisée dans une langue que dans chacune des deux autres.

Les tableaux 2, 3 et 4 donnent les Ratios de Fréquence Comparative (RFC) dans le corpus News 2024 pour les dix formes les plus fréquemment employées dans chaque langue.

Forme en -ing	RFC ES	RFC FR	RFC IT
scrapbooking	12,63	0,09	0,20
fracking	4,81	0,03	0,77
living	3,69	0,18	0,73
trending	3,40	0,13	0,89
skateboarding	3,28	0,04	1,12
birdwatching	2,57	0,00	1,56
sexting	2,51	0,19	1,10
ranking	2,21	0,11	1,46
sleeping	2,20	0,00	1,82
surfing	2,04	0,15	1,48

Tableau 2 : Ratios de Fréquence Comparative (RFC) des 10 formes en -ing les plus fréquemment employées en espagnol dans le corpus News 2024

Forme en -ing	RFC ES	RFC FR	RFC IT
monitoring	0,00	50,58	0,08
planning	0,02	26,57	0,13
parking	0,18	20,09	0,02
curling	0,01	19,20	0,19
travelling	0,08	16,45	0,15
footing	0,11	15,37	0,13
timing	0,08	13,58	0,20
lobbying	0,02	12,48	0,30
clubbing	0,14	7,82	0,32
brushing	0,34	7,67	0,13

Tableau 3 : Ratios de Fréquence Comparative (RFC) des 10 formes en -ing les plus fréquemment employées en français dans le corpus News 2024

Forme en -ing	RFC ES	RFC FR	RFC IT
stalking	0,00	0,02	178,59
dribbling	0,04	0,00	112,19
styling	0,03	0,01	84,31
multitasking	0,05	0,00	83,34
debriefing	0,09	0,00	46,31
screening	0,02	0,09	36,17
outing	0,07	0,10	22,24
closing	0,01	0,18	22,04
wrestling	0,20	0,00	19,87
stretching	0,04	0,17	18,44

Tableau 4 : Ratios de Fréquence Comparative (RFC) des 10 formes en -ing les plus fréquemment employées en italien dans le corpus News 2024

La comparaison des tableaux 2 à 4 montre clairement une tendance à l'emprunt des formes en *-ing* plus nette en français qu'en espagnol, et beaucoup plus nette en italien que dans les deux autres langues. Ainsi, le RFC de la dixième forme la plus employée en espagnol (*surfing*) n'est que de 2,04, en comparaison des valeurs obtenues pour les formes de même rang en français (*brushing*, avec un RFC de 7,67) et en italien (*stretching*, avec un RFC de 18,44)

Afin de délimiter plus clairement les différences entre l'italien et les deux autres langues, nous avons comptabilisé deux groupes de lexies : celles qui sont employées en moyenne au moins 10 fois plus dans l'une des trois langues ($\text{RFC} > 20$) et celles qui sont employées entre 4 et 9 fois plus dans l'une des trois langues.

En italien, le premier groupe ne dénombre pas moins de 18 formes (*stalking, dribbling, styling, multitasking, debriefing, screening, outing, closing, wrestling, stretching, sharing, killing, publishing, painting, dating, snorkeling, drifting, meeting*) et le deuxième en compte 27 (*watching, contouring, peeling, racing, melting, shocking, trekking, dancing, engineering, banking, rating, feeling, landing, mobbing, mining, tracking, learning, shopping, walking, coming, spoofing, trading, punching, gaming, mailing, smoking, warning*).

Pour le français, le premier groupe contient environ deux fois moins de lexèmes qu'en italien (huit formes en *-ing* au total : *monitoring, planning, parking, curling, travelling, footing, timing, lobbying*), et le deuxième groupe en contient trois fois moins (9 formes au total : *clubbing, brushing, casting, coaching, consulting, passing, jogging, popping, camping*). En espagnol, les deux groupes ne compte qu'un seul lexème (*scrapbooking / fracking*).

Ces résultats font apparaître que l'espagnol est bien moins perméable aux emprunts en *-ing* que l'italien, le français résistant également davantage à ce type d'emprunt, quoique dans une moindre mesure.

4. Comparaison des résultats des corpus News 2020 et News 2024

L'analyse des fréquences relevées pour les cent formes de notre étude révèle une augmentation moyenne de 14,7% de leur fréquence par million de mots sur l'ensemble des trois langues. Cette augmentation est toutefois plus importante en espagnol (28%) qu'elle ne l'est en italien ou en français (respectivement 16,4 et 6,7%)². Mais les données comparatives restent très similaires : on relève toujours plus de deux fois plus de formes en *-ing* en français qu'en espagnol, le ratio dépassant 3 quand on compare l'italien à ce dernier. Les fréquences observées pour les trois langues dans les deux corpus sont reproduites dans le Tableau 5.

	News 2020			News 2024		
	SP	FR	IT	SP	FR	IT
Fréquence / million de mots	83,9	204,2	309,1	107,4	217,8	359,8
% croissance 2020-2024				28,0	6,7	16,4
Ratio / Espagnol	1	2,43	3,68	1	2,03	3,35

Tableau 5 : Fréquence comparée des formes en *-ing* dans les corpus News 2020 et News 2024

² Cette augmentation pourrait être liée au fait que la RAE (Académie Royale Espagnole de la Langue), instance normative, donne de plus en plus de place, notamment au sein de son dictionnaire (DLE), à de nouveaux termes intégrés sans l'adaptation graphique habituellement recommandée (*ránking* pour *ranking*).

Afin d'examiner les formes dont l'usage avait le plus augmenté pendant les quatre ans séparant les deux corpus, nous les avons triées pour chaque langue par ordre décroissant de leur facteur d'augmentation, soit le quotient du nombre de formes relevées dans les corpus News 2024 et News 2020. Les tableaux 6 à 8 donnent les résultats obtenus dans chacune des trois langues pour les 10 formes dont l'usage a le plus augmenté. Pour chacune des formes, nous avons également fait figurer dans le tableau le nombre total d'occurrences dans le corpus News 2024 toutes langues confondues et celui obtenu pour la langue considérée.

Forme en -ing	Total News 2024	Total ES	Facteur Augmentation ES
dressing	17	12	12,07
scrapbooking	80	76	10,92
breaking	442	221	8,55
engineering	170	23	4,63
rebranding	253	99	4,53
drinking	37	9	4,53
shocking	65	21	4,22
skateboarding	144	118	3,83
clubbing	58	12	3,02
trekking	1 407	510	2,98

Tableau 6 : Augmentation de la fréquence des formes en -ing entre les corpus News 2020 et News 2024 en espagnol

Dans certains cas, le facteur d'augmentation peut être élevé sans que l'usage de la forme soit forcément très répandu. Ainsi, *dressing*, qui était un hapax en 2020, compte 12 occurrences en 2024, mais elles concernent surtout des noms composés (*power dressing*, *latte dressing*). Le fait que 2024 ait été une année olympique peut également expliquer l'augmentation des occurrences des formes désignant certains sports (*breaking*, *skateboarding*, *kickboxing*, *surfing*). Les termes quantitativement les plus employés concernent souvent le domaine du commerce et de l'entreprise (*engineering*, *rebranding*) et celui des nouvelles technologies (*phishing*, *networking*). Notons également que certaines formes (par exemple *trekking*), sont presque exclusivement employées dans la partie sud-américaine du corpus.

Forme en -ing	Total News 2024	Total FR	Facteur Augmentation FR
contouring	97	8	10,43
rebranding	253	31	6,74
breaking	442	144	5,87
spoofing	154	20	5,22
curling	281	241	4,84
stalking	1 343	10	4,35
spinning	140	21	3,91
passing	83	32	3,79
mining	239	43	3,30
screening	2 102	60	3,13

Tableau 7 : Augmentation de la fréquence des formes en -ing entre les corpus News 2020 et News 2024 en français

On observe la même augmentation qu'en espagnol des occurrences des formes désignant certains sports ou actions liées au sport (*breaking*, *curling*, *passing*, *pressing*, *jumping*). Notons qu'un tiers des occurrences de *passing* ne concerne pas le tennis, mais la précision des passes dans les sports de ballon, toutes les occurrences concernant ce sens provenant de la presse belge. Par ailleurs, certaines formes en -ing sont quasiment absentes des organes de presse de la France métropolitaine : les termes *branding* et *rebranding* apparaissent majoritairement dans la presse francophone d'Afrique, et tous les emplois de

sharing concernent le *car sharing* (exclusivement employé dans la presse suisse) ou les bigrammes *food sharing/sharing food* (exclusivement employés dans la presse belge).

Forme en -ing	Total News 2024	Total IT	Facteur Augmentation IT
listing	106	62	21,08
skateboarding	144	25	12,75
dancing	207	169	7,50
spoofing	154	101	4,68
surfing	74	17	4,34
travelling	46	4	4,08
bodybuilding	79	50	3,92
ranking	12 233	2 735	3,55
swinging	10	10	3,40
mining	239	191	2,91

Tableau 8 : Augmentation de la fréquence des formes en -ing entre les corpus News 2020 et News 2024 en italien

À l'exception des termes *listing* et *mining* (relevant du monde de la finance) et du terme *spoofing*, qui concerne celui des télécommunications, la majorité des items dont la fréquence augmente fortement sont également liés au sport. Certaines formes comme *stalking* ne sont presque pas employées dans les deux autres langues du corpus, et d'autres sont exclusivement employées en tant que constituants d'une lexie de taille supérieure (*binge drinking*). Notons que l'entrée dans la langue d'une lexie de ce type (*nordic walking*, très fréquemment employé dès 2020) semble faciliter l'emprunt ultérieur d'autres lexies contenant la même forme en -ing (*power walking*, *silent walking*, *retro walking*). Quant à la forme *opening*, si elle figure en tant qu'adjectif dans de nombreux bigrammes empruntés à l'anglais (*opening act/party/night/credits/ceremony*), elle est majoritairement employée en tant que nom en italien.

Forme en -ing	Total News 2024	Facteur Augmentation
listing	106	7,52
skateboarding	144	5,74
breaking	442	5,48
meeting	15 333	4,87
rebranding	253	4,21
dresssing	17	4,15
contouring	97	4,10
scrapbooking	80	4,03
spoofing	154	3,78
dancing	207	3,08
drinking	37	2,77
ranking	12 233	2,55
engineering	170	2,53
networking	1 660	2,50
surfing	74	2,33
stalking	1 343	2,27
mining	239	2,24
shocking	65	2,18
curling	281	2,11
travelling	46	2,07

Tableau 9 : Augmentation de la fréquence des formes en -ing entre les corpus News 2020 et News 2024 toutes langues confondues

Le tableau 9 montre les vingt formes pour lesquelles le facteur d'augmentation de la fréquence est le plus élevé sur l'ensemble des trois langues. Notons que les formes figurant dans le tableau sont très souvent fortement employées dans l'une des trois langues et nettement moins dans les deux autres. On compte toutefois 29 formes dont l'usage a augmenté dans les trois langues (*bowling, branding, breaking, building, casting, coaching, drinking, engineering, going, kickboxing, living, lobbying, making, marketing, merchandising, networking, opening, racing, ranking, rebranding, shocking*).

5. Conclusion

Qu'il s'agisse des résultats obtenus dans le corpus News 2020 ou le corpus News 2024, l'italien emprunte plus de formes en *-ing* que le français (et beaucoup plus que l'espagnol). Les résultats obtenus ici profiteraient d'une comparaison avec des données similaires portant sur des emprunts autres que ceux contenant le morphème *-ing* dans les trois langues romanes de notre étude. Dans le cas présent, cette comparaison a été facilitée par l'extraction automatique des formes se terminant par le morphème *-ing*, même si ce processus a dans un premier temps fourni dans les trois langues un nombre important de formes ne correspondant pas à des emprunts à l'anglais (notamment des noms propres). Mais comme l'ont montré les publications issues du projet Néoveille pour le français (Cartier E. et al. 2018, Cartier 2019), l'extraction automatique des néologismes reste un processus complexe, notamment dans le cas de la néologie sémantique (Sablayrolles, 2012). L'italien et l'espagnol ne semblant pas actuellement disposer de projets de l'ampleur de Néoveille pour la détection des néologismes³, une comparaison quantitative de l'ensemble des emprunts à l'anglais dans les trois langues se fondant sur l'utilisation de corpus comparables devra probablement se faire attendre encore quelque temps.

Comme nous l'avons fait remarquer plus haut, l'actualité sportive a probablement eu un impact sur l'augmentation de la fréquence des formes étudiées observée entre les années 2020 et 2024. Cependant, la forte augmentation des chiffres concernant l'espagnol est peut-être due à une diminution de la résistance à l'emprunt de formes anglaises dans cette langue. Il conviendra de vérifier la confirmation de cet accroissement de l'utilisation des formes en *-ing* en espagnol en ayant recours à d'autres corpus dans un futur proche.

Bibliographie

- Alvar Ezquerro M. (1995). *La formación de palabras en español*, Madrid, Arco Libros.
- Bachis, D. (2023). Tra misoneismo e “fedeltà linguistica”: la percezione dei neologismi in un campione di parlanti. *Circula*, (17), 237-259.
- Bistarelli, A. (2008). L'interferenza dell'inglese sull'italiano. *inTRAlinea*, 10, 1-11.
- Furiassi, C., Pulcini, V., et González, F. R. (ed.) (2012). *The anglicization of European lexis*. John Benjamins Publishing.
- Gobbo, F. (2017). Review of: Claudio Marazzini e Alessio Petrali (2015). *La lingua italiana e le lingue romanze di fronte agli anglicismi*. *Language Problems and Language Planning*, 41(2), 210-213.
- Grossmann M., Rainer F. (dirs.) (2004). *La formazione delle parole in italiano*, Tübingen, Niemeyer.
- Haensch G. (2005). « Anglicismos en el español de América », *ELUA*, 19, pp. 243-251
- Humbley J. (2008). Emprunts, vrais et faux, dans le *Petit Robert 2007*, In PRUVOST J. (dir.), *Les journées des dictionnaires de Cergy : Dictionnaires et mots voyageurs. Les 40 ans du Petit Robert, de Paul Robert à Alain Rey*, 221-238, Herblay, Éditions des Silves.

³ Notons cependant pour l'espagnol l'existence de la banque de néologismes de l'OBNEO (Observatoire de la néologie de l'Université Pompeu Fabra de Barcelone).

- Lison P., Tiedemann J. (2016), « OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles », in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 923–929.
- Maniez, F. (2014). “Implantation of English terms including the -ING morpheme in French, Spanish and Italian: A corpus-based study of the debates of the European Parliament” in Pascaline Dury et al. (dirs) *La néologie en langue de spécialité : détection, implantation et circulation des nouveaux termes*, pp. 189-201, Lyon, Travaux du CRTT.
- Marazzini, C ; et Petralli, A. (2015), *La lingua italiana e le lingue romanze di fronte agli anglicismi*, tiré des travaux des travaux du " Gruppo Incipit " de l'Accademia della Crusca.
- Picone M. D. (1996). *Anglicisms, Neologisms and Dynamic French*, Amsterdam, Benjamins.
- Rati, M. S. (2025). “Gli anglicismi nell’italiano accademico”. *Italiano LinguaDue*, 17(2), 201-212.
- Renner, V., & Fernandez-Dominguez, J. (2015). 6. False Anglicization in the Romance languages: A contrastive analysis of French, Spanish and Italian. In *Pseudo-English* (pp. 147-158). De Gruyter Mouton.
- Rodríguez González. F. (2022), *Anglicismos en el español contemporáneo*, Berlin/Bern/Bruxelles/New York/Oxford/Warszawa/Wien, Peter Lang.

Sitographie

BOBNEO, Banc de Neologismes, <https://bobneo.upf.edu/>
 Leipzig Corpora Collection, <https://corpora.wortschatz-leipzig.de/>
 OPUS, the open parallel corpus, <<http://opus.lingfil.uu.se/>>

C01

Textometry, corpus methods and visualization

09-lug
17:00 - 18:30
Aula 11

Articuler méthodes quantitatives et épistémologies critiques en linguistique : enjeux, tensions et perspectives méthodologiques

Tiago Joseph¹

¹Ghent University – tiago.joseph@ugent.be

Abstract

This article explores how quantitative methods and critical epistemologies— decolonial, feminist and queer —can be articulated within the language sciences. Although often perceived as incompatible, these approaches can mutually enrich one another. The first section reviews the critiques directed at quantitative methods, accused of essentializing and obscuring minority experiences, as well as reproducing positivist assumptions. It nevertheless highlights the emergence of research that reintegrates quantitative linguistics within a reflexive, political, and collaborative framework. The second section identifies key convergences between textometrics and critical perspectives, including the importance of context, low frequencies, exploratory analysis, and multimodality. Finally, a case study based on a feminist and LGBTQIA+ Instagram corpus focusing on (non-)consent illustrates the methodological challenges of such hybridization. Through issues related to importation, structuring, weighting, lemmatization, and metadata, the article reflects on methodological contingencies and the forms of bricolage they require, emphasizing the need to adapt tools and methods for critically analyzing multimodal digital corpora.

Keywords: corpus linguistics; textometrics; TXM; decoloniality; queer; feminism

Résumé

Cet article explore les possibilités d'articulation entre méthodes quantitatives et épistémologies critiques – décoloniales, féministes et queer – en sciences du langage. Souvent perçues comme opposées, ces approches peuvent pourtant s'enrichir mutuellement. La première partie retrace les critiques adressées aux méthodes quantitatives, accusées d'essentialiser et d'invisibiliser des expériences minoritaires, et de positivisme. Elle montre toutefois l'émergence de travaux qui réinscrivent la linguistique quantitative dans une démarche réflexive, politique et collaborative. La seconde partie met en évidence les convergences entre textométrie et perspectives critiques, notamment l'importance du contexte, des basses fréquences, de l'exploration et de la multimodalité. Enfin, un cas d'étude portant sur un corpus Instagram féministe et LGBTQIA+ consacré au (non-)consentement illustre les enjeux méthodologiques de cette hybridation. À travers des défis liés à l'import, la structuration, la pondération, la lemmatisation et les métadonnées, l'article revient sur des imprévus méthodologiques et les bricolages qu'ils nécessitent, et souligne l'importance d'adapter les outils et méthodes pour analyser des corpus numériques multimodaux dans une perspective critique.

Mots clés : linguistique de corpus ; textométrie ; TXM ; décolonialité; queer ; féminisme

1. Introduction

Cet article réfléchit sur les possibilités de croisement entre méthodes quantitatives et épistémologies critiques – féministes, décoloniales et queer – en sciences du langage. Alors que ces épistémologies insistent sur la réflexivité, la situation des savoirs, l'attention portée aux marges et la prise en compte des rapports de pouvoir, elles perçoivent parfois les approches quantitatives comme une pratique scientifique ambivalente, reproduisant le mythe positiviste

de l'objectivité, l'homogénéisation et la disqualification des savoirs et expériences marginalisées.

La première section interroge les tensions historiques entre ces épistémologies et la linguistique quantitative. J'y montre que, malgré de fortes critiques, plusieurs réflexions s'efforcent de réinscrire la méthode quantitative dans une démarche politique, collaborative et plurielle. Ces recherches démontrent que la linguistique de corpus peut devenir un espace d'exploration, de créativité et de décentrement, en phase avec les épistémologies critiques.

La seconde section met en lumière les convergences existantes entre textométrie et approches critiques, souvent sous-estimées. Elle rappelle que la textométrie se présente comme une méthode mixte, qui accorde une place centrale au contexte, au retour au texte, à la complexité des données ou encore à la prise en compte des basses fréquences – autant d'éléments cohérents avec une perspective anti-essentialiste et critique.

Enfin, la troisième section revient sur un cas d'étude issu d'un corpus multimodal texte-image issu d'Instagram. Ce corpus est composé de discours féministes et LGBTQIA+ sur la notion de (non-)consentement. À travers une série de défis méthodologiques (import, édition, pondération, lemmatisation, métadonnées), cette étude expose les ajustements nécessaires pour articuler corpus atypique, méthode textométrique et épistémologies critiques, en soulignant les potentialités ouvertes par ce *métissage* disciplinaire.

2. Articuler les études quantitatives et les épistémologies critiques

2.1. La méthode quantitative face aux épistémologies décoloniales, queer et féministes

Au-delà d'un mouvement social complexe et d'une grille d'analyse des rapports sociaux, le féminisme fonctionne comme un socle épistémologique : il constitue "une démarche critique et un système de construction de la connaissance" (Bruneel et al., 2023, p. 3). À l'instar de toute démarche scientifique (critique), une démarche féministe accorde de l'importance au doute et à l'exploration, à l'instabilité/contextualité du sens et des catégories, à la réflexivité et à la transparence de la démarche. En particulier, les épistémologies féministes se caractérisent par une réflexion sur la situation des savoirs : ils sont considérés comme nécessairement subjectifs et politiques, construits depuis une position sociale et de recherche singulière traversée par les idéologies (Bruneel et al., 2023 ; Caldeira, 2024 ; franzke, 2020 ; Millette, 2023).

Il importe de démultiplier les perspectives autour d'un "objet" de recherche dans le but d'entretenir un rapport critique au savoir, attentif aux rapports de pouvoir qui le traversent. Il s'agit alors d'interroger des phénomènes tels que l'universalisation d'un point de vue (par exemple la blancheur) ou encore l'homogénéisation d'une catégorie sociale (par exemple "les femmes"). Assumant la dimension politique de la recherche, ces épistémologies critiques sont orientées vers la justice sociale et cherchent "à dévoiler la colonialité comme matrice persistante d'organisation du monde" (Chbat et Enriquez, 2026, p. 11), ainsi que les autres rapports sociaux qui s'y imbriquent, tels que l'impérialisme raciste, le validisme, le genre, la sexualité, l'âgisme, etc. C'est dans cette optique qu'on parle d'approche intersectionnelle (Paveau, 2023).

Les épistémologies queer, féministes, décoloniales, intersectionnelles ont trouvé des applications au sein de la linguistique, notamment au travers de l'analyse du discours féministe (Lazar, 2017 ; Marignier, 2016 ; Abbou, 2022), queer (Motschenbacher et Stegu, 2013) et antiraciste et décoloniale (Salim et al., 2016 ; Telep, 2019 ; Paveau et Telep, 2022 ; Flores et Rosa, 2023 ; Paveau, 2023) – bien que cette dernière soit encore très marginalisée dans les sciences du langage.

Cependant, en ce qui concerne les approches textuelles quantitatives, les travaux demeurent rares. Certaines réflexions émergent en anglais comme en français, depuis les sciences des données (Leurs, 2017 ; D'Ignazio et Klein, 2020), la linguistique de corpus et la textométrie (entre autres Baker, 2018 ; Motschenbacher, 2018 ; Flesch, 2020 ; Prévost et al., 2021 ; Atzeni, 2022) et les sciences de la communication (Millette, 2023 ; Caldeira, 2024), mais elles sont loin de constituer un champ de recherche structuré et dynamique.

Cette rareté des travaux quantitatifs en études féministes, queer et décoloniales peut s'expliquer par une critique et une méfiance envers les revendications d'objectivité, de neutralité et de généralité/universalité des études quantitatives (Motschenbacher, 2018 ; Gowdie, 2022 ; Bruneel et al., 2023 ; Caldeira, 2024). Les études quantitatives ont également été critiquées pour leurs risques d'homogénéisation et d'essentialisation (catégorisation), de polarisation (focus sur le contraste) et d'in/sur-visibilisation (focus sur les hautes fréquences) de certaines catégories sociales, des discours, des normes, des rapports de pouvoir, là où les épistémologies critiques (poststructuralistes) portent une attention particulière aux marges, fluidités et résistances. Tandis que les démarches quantitatives favorisent une recherche distanciée, non-collaborative, où les savoirs non académiques sont anonymisés, appropriés et infériorisés, interprétés de l'extérieur, les épistémologies critiques proposent de renouveler le processus scientifique en l'inscrivant dans une logique de dialogue horizontal, de collaboration consentie, de redistribution matérielle et symbolique, de responsabilisation et d'implication politique collective (Gowdie, 2022 ; Abbou et Burnett, 2023 ; Bruneel et al., 2023 ; Caldeira, 2024 ; franzke, 2020 ; Millette, 2023).

Néanmoins, ces critiques émergent de travaux qui, justement, s'essaient à la méthode quantitative et lui reconnaissant une légitimité – notamment pour "montrer la dimension systémique et endémique des injustices sociales" (Millette, 2023, p. 12). Face aux limites respectives du qualitatif comme du quantitatif, ces travaux reconnaissent l'importance des méthodes mixtes et tentent d'inscrire la méthode quantitative dans une démarche (plus) critique et éthique (Bruneel et al., 2023 ; Caldeira, 2024 ; franzke, 2020 ; Millette, 2023 ; Motschenbacher, 2018). La linguistique de corpus devient alors l'occasion d'un exercice réflexif, où sont questionnés les effets discursifs et politiques des choix méthodologiques, où les contextes et les identités sociales deviennent indispensables au travail d'interprétation ; créatif, où l'attention aux contrediscours et aux polysémies/polyphonies permettent la désessentialisation et la subversion ; extensif, où les personnes, perspectives et occurrences habituellement absentes et alternatives sont davantage intégrées, pour aller au-delà des actrices et paroles socialement valorisées (Motschenbacher, 2018 ; cfr. Baker, 2018 ; Bruneel et al., 2023 ; Caldeira, 2024 ; franzke, 2020 ; Millette, 2023). Dans cette perspective, la démarche quantitative se transforme en espace de pluralité, d'hétérogénéité, qui permet de rendre les voix et de revaloriser des savoirs personnels et militants non hégémoniques, à l'instar d'une postlinguistique folk et décoloniale (Paveau, 2023).

2.2. Convergences entre épistémologies critiques et textométrie

Ainsi la méthode quantitative apparaît-elle comme un lieu traversé de tensions, auxquelles les épistémologies critiques cherchent des solutions. Il faut néanmoins rendre justice aux travaux de linguistique de corpus et de textométrie qui reconnaissent certaines limites de la méthode quantitative, et qui convergent en plusieurs points avec les critiques précédentes. Ainsi, les travaux axés sur la méthodologie textométrique insistent sur l'importance du contexte, la nécessité constante d'un retour au texte originel et une connaissance fine du corpus (Baker, 2018 ; Leblanc et al., 2017 ; Pincemin, 2018, 2020). La textométrie est à ce titre considérée

comme une méthode intrinsèquement mixte. Ces travaux soulignent également l'importance de réfléchir aux impacts qu'ont la composition du corpus et la structuration des données (Barats, 2024), et de sélectionner les outils selon leurs affordances, d'en comprendre les soubassements statistiques, d'en connaître les limites (Fleury et al., 2017 ; Pincemin, 2020).

À ce titre, certains travaux montrent la nécessité de prendre en compte la complexité et la multimodalité des données, au-delà de leur seule dimension textuelle (Baker, 2018 ; Née et Fleury, 2017). Face aux limites des méthodes automatisées pour les données visuelles (Barats et al., 2013 ; Bouko, 2024), il importe de développer des outils d'analyse capables de compléter les logiciels de textométrie (Baker, 2018 ; Joseph et al., ce volume). Enfin, des discussions ont lieu à propos des critères et des attentes de représentativité, d'interprétation et d'homogénéité des corpus, sur la richesse discursive des basses fréquences et sur les enjeux de la lemmatisation (Fleury et al., 2017 ; Leblanc et al., 2017 ; Pincemin, 2020 ; Sitri et Barats, 2017). La textométrie se pense parfois davantage comme une pratique exploratoire que confirmatoire (Pincemin, 2020).

Au-delà des idées reçues, cette réflexion tend à montrer la possibilité et la richesse de croiser les méthodes de linguistique quantitative et les épistémologies critiques, décoloniales, féministes et queer. La section suivante présente un cas d'étude pratique et réfléchit aux enjeux épistémo-méthodologiques d'une telle démarche.

3. Retour sur un cas d'étude : quelques enjeux méthodologiques d'une analyse du discours quantitative, féministe et queer

3.1. Présentation du corpus

Ma recherche doctorale s'intéresse aux idéologies féministes et LGBTQIA+ sur les violences et éthiques sexuelles en contexte numérique (Instagram). Mon corpus est constitué de 1447 profils francophones (Belgique, France), dont j'ai collecté plus de 304 000 posts (Joseph et Bouko, 2024, 2025). En m'intéressant aux représentations du consentement dans les comptes individuels (projet d'une seule personne), j'ai constitué un sous-corpus de 402 profils et de 3696 publications (Joseph, à paraître). Chaque publication est composée d'une ou plusieurs images, lesquelles peuvent contenir du texte qui a été OCRisé, et d'une légende, qui diffère la plupart du temps du texte des images et intègre souvent un ou plusieurs hashtags. Ce sous-corpus est composé de 1 872 003 mots (textes OCRisés et légendes). Ce corpus multimodal, numérique et complexe est relativement atypique pour une étude textométrique. Je souhaite revenir sur certains challenges méthodologiques que ce corpus a posés, autant lors de son import que lors des analyses dans TXM (Heiden et al., 2010). Ces retours ne sont pas des lacunes de l'outil, mais davantage la trace d'une inadéquation partielle entre mon corpus et les outils de textométrie disponibles à l'heure actuelle.

3.2. Enjeux d'import et d'édition

Le logiciel TXM a originellement été désigné pour des analyses de quelques longs textes. S'il permet l'édition de données visuelles, c'est sur base du modèle de la page scannée, dont le texte a été retranscrit (par exemples des manuscrits médiévaux).

Importer un corpus composé de milliers de textes courts a constitué un premier défi. En effet, la mémoire du logiciel ne supporte pas la structuration en milliers de textes, enrichis de multiples métadonnées. La solution consiste alors à importer l'ensemble des publications

comme un texte unique, où chaque publication constitue un chapitre¹. Si cette solution n'est pas en soi problématique, elle a néanmoins des incidences sur les calculs (voir section 3.3).

Par ailleurs, TXM permet l'édition des métadonnées et des images des publications, ce qui constitue un atout majeur pour le retour au texte. Néanmoins, trop de métadonnées alourdissent l'affichage et nécessitent de faire systématiquement défiler l'onglet d'édition avant d'arriver au texte, ce qui peut s'avérer contrariant. Afin de ne pas réduire l'ergonomie de l'outil, j'ai choisi de sélectionner uniquement les métadonnées les plus utiles.

De plus, TXM a été développé pour associer une portion de texte à une seule image correspondante. Dans mon cas, les publications contiennent jusqu'à 10 images. Pour résoudre cet écart, il a fallu générer un lien associant chaque image à son emplacement (sur le disque de l'ordinateur par exemple), et associer ces liens aux données textuelles lors de l'import. La flexibilité de TXM est une fois encore un atout, mais cela demande aux chercheuses davantage de temps et de connaissances informatiques. Si la navigation entre images à l'aide d'une barre de défilement est tout à fait pratique, la possibilité de faire défiler les images (comme sur Instagram) permettrait d'améliorer l'ergonomie de l'outil pour ce type de corpus. À titre anecdotique, une autre amélioration possible serait de rendre cliquable les différents liens URL présents dans les métadonnées, afin de faciliter l'accès au contexte numérique d'origine.

Face à ces difficultés de retour au texte et au contexte numérique, ainsi qu'à l'absence d'outil adapté pour l'analyse manuelle de corpus texte-image pourtant nécessaire à toute recherche mixte, nous avons développé avec Michaël Stappers l'outil SMIQQDA, *Social media interface for quantitative and qualitative data analysis* (Joseph et Stappers, 2025). Cet outil a été désigné de façon à pouvoir dialoguer avec TXM, dans une logique de complémentarité logicielle permettant d'améliorer la structuration, l'édition (visualisation) et l'annotation de gros corpus texte-image.

3.3. Enjeux statistiques liés aux calculs des mots, à la pondération et à la lemmatisation

TXM se base sur la sémantique différentielle et interprétative de François Rastier (Pincemin, 2020). Les textes sont découpés en unités graphiques, parfois lemmatisées, qui sont comptées et caractérisées en contexte, par contraste avec les unités graphiques qui les accompagnent et entre les différents textes.

Face à ce fonctionnement, mon corpus a représenté trois difficultés importantes. Premièrement, le comptage des mots est basé non pas sur l'unité textuelle (la publication) mais sur l'ensemble du corpus. Certes, il est possible de partitionner le corpus selon telle variable et de compter les mots de chaque modalité, mais cela ne donne qu'une vague idée des unités textuelles analysées. Dans mon cas, la solution a été de compter le nombre d'espaces (et d'ajouter 1)² au sein des cellules de texte (hors cellules vides) dans Excel, pour approximer le nombre de mots et calculer différentes statistiques descriptives (Tableau 1).

¹ Je remercie chaleureusement Serge Heiden et Mathieu Decorde, les deux ingénieurs en charge de TXM, pour leur aide précieuse et répétée.

² `"=NBCAR(A1)-NBCAR(SUBSTITUE(A1; " "; ""))+1"`. Cette formule est approximative ; elle compte par exemple les unités avec apostrophe comme un seul mot (par exemple "l'individu").

Statistiques descriptives (nombre de mots)	OCR	Légende
Minimum	0	0
1er décile	3	13
1er quartile	24	33
Médiane	128	74
Moyenne	291	110
3ème quartile	394	169
Dernier décile	788	271
Maximum	5597	404

Tableau 1. Statistiques descriptives des OCR et légendes des publications du corpus total (en nombre de mots)

Deuxièmement, le format des données textuelles complique l'interprétation des calculs textométriques. En effet, même si c'était techniquement possible (ce qui pour rappel n'est pas le cas), le nombre important de publications et d'autaires rend inenvisageable de structurer le corpus en autant de textes que de publications ou d'autaires : calculer des spécificités ou des analyses factorielles de correspondances (AFC) sur plus de 400 unités va à l'encontre de l'esprit de synthèse de la textométrie, et est donc peu intéressant. L'alternative est d'importer l'ensemble des publications comme un seul et même texte, puis de regrouper des ensembles de publications ou d'autaires en le partitionnant. L'import des publications comme un texte unique ne pose à priori pas de problème.

Néanmoins, lors de la partition, la pondération des occurrences n'est plus réalisée relativement à la taille du texte dont elles sont issues (la publication – comme ça serait le cas si chaque publication équivalait à un texte) mais selon la taille de la partie (l'ensemble des publications de telle modalité). Cette différence de logique est importante à prendre en compte dans l'interprétation des résultats. Par exemple, lors du calcul de spécificités, certains usages linguistiques issus d'une publication ou d'une autaire en particulier apparaissent comme spécifiques à l'ensemble de la partie, et non à la publication ou à l'autaire dont ces usages sont issus. À titre d'exemple, si je choisis de partitionner les publications sur le consentement entre les comptes qui revendiquent l'étiquette "féministe" dans leur profil (groupe A) et ceux qui ne le font pas (groupe B), et qu'un compte du groupe A fait un usage extensif du mot "phantasme", ce dernier apparaît comme statistiquement spécifique au groupe A, et non à la publication ou à l'autaire. Or, les spécificités de chaque publication n'étant pas disponibles (contrairement à quelques textes longs), il est difficile de prêter du crédit à ces calculs sans un examen systématique des fréquences et des concordances.

Une solution complémentaire est d'annoter via TXM certains marqueurs discursifs au niveau de la publication (par exemple, les publications contenant le mot "harcèlement" seront annotées "violence"), d'exporter le corpus et de l'importer dans SMIQQA, qui permet de convertir les annotations TXM en modalités d'une variable, d'exporter le corpus en format tableur et de réaliser des calculs sur base du nombre de posts concernés (et non plus sur base du nombre d'occurrences). C'est en croisant nombre d'occurrences et nombre de posts contenant tel mot qu'on obtient une idée plus représentative de son usage.

L'absence de pondération par auteur pose un autre challenge : les différents comptes étudiés produisent un nombre de posts très variables, d'autant plus lorsque ces posts concernent une thématique particulière (comme le consentement). Là où certains comptes publient moins de 10 posts sur la question, quelques autres en publient jusqu'à 300. En l'absence de pondération, les comptes produisant beaucoup invisibilisent les récurrences discursives des comptes produisant ponctuellement. Pour résoudre cette difficulté, j'ai constitué deux corpus alternatifs : le premier où le nombre de publications des gros producteurs est aléatoirement réduit à 30, soit le triple de la moyenne du nombre de publications par compte (corpus pondéré), afin de réduire l'influence de ces comptes tout en la reconnaissant ; le second se compose seulement des comptes ayant produit moins de 10 publications, afin de faire émerger d'autres discours. Ces choix s'inscrivent dans la continuité d'une linguistique de corpus queer, qui prête attention aux faibles fréquences et aux marges et tente de réduire les injustices épistémiques (Motschenbacher, 2018).

La troisième difficulté est typique du terrain numérique (Barats et al., 2013) et plus particulièrement des réseaux sociaux. Mes données textuelles se caractérisent par beaucoup de néologismes, des caractères typiquement numériques (comme les smileys, les hahstags), du multilinguisme, ainsi qu'à une grande variation formelle. Au-delà des variantes (ortho)graphiques, on y retrouve des variations de style (police, capitales) et des *algospeaks* (Tillewein et al., 2024), variations graphiques permettant d'échapper à la censure des algorithmes et des modérateurs. C'est par exemple le cas du mot "vi*1" pour "viol". À cela s'ajoute la variation et l'hétérogénéité du contenu textuel des images, obtenu par OCRisation : si la qualité est globalement satisfaisante, ce contenu contient des erreurs de reconnaissance, la transcription d'autres alphabets et de signes alphanumériques, et du bruit (par exemple la composition d'un shampoing, si le dos de la bouteille apparaît sur l'image).

Ces particularités textuelles appauvrissent considérablement les résultats de l'annotation morphosyntaxique et de la lemmatisation automatisées réalisées via TreeTagger, et donc des calculs qui ne sont pas basés sur les formes graphiques. Dans mon cas, le mot "consent" pouvait correspondre à la fois à l'indicatif présent du verbe *consentir* en français, ainsi qu'à différentes modalités verbales et au substantif en anglais. Autant les calculs basés sur les formes graphiques que sur les lemmes se sont avérés partiellement insatisfaisants. Ces résultats soulignent l'intérêt de penser les outils de lemmatisation dans une approche de linguistique décoloniale, qui reconnaît la prévalence du multilinguisme, et prend également en compte la multiplicité des langues (autres que coloniales), cessant de postuler le monolinguisme comme situation de fait. Si ces limites logicielles sont plus souvent techniques qu'idéologiques, il me semble que les dépasser nécessite toutefois une telle posture.

3.4. Enjeux liés aux métadonnées. Choisir, c'est renoncer

L'avantage des données issues des réseaux sociaux est qu'elles constituent des données riches, assorties de nombreuses métadonnées. C'est par exemple le cas des données métriques (nombre de likes et de commentaires) qui offrent la possibilité de s'intéresser aux dynamiques de réception de différents discours, et d'analyser la façon dont elles reproduisent et subvertissent l'hégémonie culturelle et les idéologies dominantes (franzke, 2020). C'est ainsi que j'ai pu faire l'hypothèse de biais libéraux et sensationnalistes dans la réception des publications sur le (non-)consentement sur Instagram (Joseph, à paraître).

Mais la richesse des métadonnées constitue également un risque. Elle favorise la démultiplication des analyses, qui s'avèrent particulièrement chronophages et peuvent in fine entraîner une confusion dans la masse des résultats. Il me semble alors intéressant de voir ces

métadonnées comme autant de fils à tirer, tisser, et parfois abandonner, compatibles avec la logique exploratoire de la textométrie.

4. Remerciements

Je remercie chaleureusement Serge Heiden et Mathieu Decorde, les deux ingénieurs en charge de TXM, pour leur aide précieuse et répétée, ainsi que Michaël Stappers, l'ingénieur du logiciel *SMIQQDA*.

References

- Abbou J. (2022). *Tenir sa langue. Le langage, lieu de lutte féministe*. Les Pérégrines.
- Abbou J. and Burnett H. (2023). Des corpus féministes face à l'institution éthique. Limites matérielles et administratives de l'interdisciplinarité en genre et langage. In Arab C. et al., comité scientifique et organisationnel, *3^{ème} Congrès International de l'Institut du Genre. Nos futurs. Genre : bouleversements, utopies, impatiences*, Toulouse, 04–07 July 2023.
- Atzeni C. (2022). Exploration textométrique du discours de l'Organisation des Nations Unies sur la violence envers les femmes et de son évolution entre 1996 et 2019 : une étude sur corpus de deux genres discursifs d'organisation internationale en anglais et en français. Thèse de doctorat, Université Paris Cité.
- Baker P. (2018). Language, sexuality and corpus linguistics: Concerns and future directions. *Journal of Language and Sexuality*, 7(2): 263–279.
- Barats C. (2024). Corpus. *Publictionnaire*. <https://publictionnaire.huma-num.fr/notice/corpus/>
- Barats C., Leblanc J.-M. and Fiala P. (2013). Chapitre 5. Approches textométriques du web : corpus et outils. In Barats C., éditrice, *Manuel d'analyse du web en Sciences Humaines et Sociales*. Armand Colin.
- Bouko C. (2024). *Visual Citizenship: Communicating Political Opinions and Emotions on Social Media*. Routledge.
- Bruneel E., Olivesi A. and Verquere L., éditrices (2023). Enquêter sur le genre en communication. *Communication*, 40(2).
- Caldeira S. (2024). Exploring feminisms on Instagram: Reflections on the challenges and possibilities of incorporating digital methods strategies in feminist social media research. *Journal of Digital Social Research*, 6(1): 74–89.
- Chbat M. and Enriquez C. (2026). Introduction. In Chbat, M., éditrice, *Décolonialité et résistances queers*. Les Éditions du Remue-Ménage.
- D'Ignazio C. and Klein L.F. (2020). *Data feminism*. The MIT Press.
- Flesch M. (2020). lol thats how reddit talks ;) : le site américain Reddit comme espace de variation de l'anglais. Étude de corpus intersectionnelle et quantitative d'usages non standard, au prisme du genre, de l'âge et de l'ethnicité. Thèse de doctorat, Université de Lorraine.
- Fleury S., Née E. and Barats C. (2017). Chapitre I. Compter les mots ? Pas si simple... In Née E, éditrice, *Méthodes et outils informatiques pour l'analyse des discours*. Presses universitaires de Rennes.
- Flores N. and Rosa J. (2023). Undoing raciolinguistics. *Journal of Sociolinguistics*, 27: 421–427.
- franzke a.s. (2020). Feminist Research Ethics. In Association of Internet Researchers, éditrice, *Internet Research: Ethical Guidelines 3.0*. <https://aoir.org/reports/ethics3.pdf>
- Gowdie P. (2022). Epistemologies Underlying Feminist and Anti-Racist Methodologies and Their Contributions to Critical Sociological Research. *Academia Letters*, Article 4779.

- Heiden S., Magué J.-P. and Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In Bolasco S., Chiari I. and Giuliano L., éditeurs, *Proc. of JADT 2010*, pp. 1021–1032.
- Joseph T. (à paraître). *Consentir sur Instagram. Perspectives féministes et queers*.
- Joseph T. and Bouko C. (2024). Espace, discours et identité féministes sur Instagram. Modes de construction d'un soi féministe dans les profils Instagram : propositions méthodologiques. *Essachess*, 17(1/33): 65–93.
- Joseph T. and Bouko C. (2025). Mapping feminist identities on Instagram: identity markers, profile types, and social logics of digital activism. *Feminist Media Studies*, 1-25.
- Joseph T., Stappers M., Gent University (2025). *SMIQQDA, Social Media Interface for Quantitative and Qualitative Data Analysis (Version 1.0)*. Zenodo, <https://doi.org/10.5281/zenodo.15642544>.
- Joseph T., Stappers M. and Bouko C. (ce volume). Bridging the tool gap in visual social media research: SMIQQDA, a free and open-source solution.
- Lazar M.M. (2017). Feminist critical discourse analysis. In Flowerdew J. and Richardson J., éditeurs, *The Routledge Handbook of Critical Discourse Studies*. Routledge.
- Leblanc J.-M., Fleury S. and Née E. (2017). Chapitre V. Quels outils logiciels et pour quoi faire ? In Née E, éditrice, *Méthodes et outils informatiques pour l'analyse des discours*. Presses universitaires de Rennes.
- Leurs K. (2017). Feminist Data Studies: Using Digital Methods for Ethical, Reflexive and Situated Socio-Cultural Research. *Feminist Review*, 115(1): 130–154.
- Marignier N. (2016). Les matérialités discursives du sexe. La construction et la déstabilisation des évidences du genre dans les discours sur les sexes atypiques. Thèse de doctorat, Université Sorbonne Paris Cité.
- Millette M. (2023). Théorie et méthodologie féministes pour la recherche en contexte numérique. *Communication*, 40(2): 1–20.
- Motschenbacher H. and Stegu M. (2013). Queer Linguistic approaches to discourse. *Discourse & Society*, 24(5): 519-535.
- Motschenbacher H. (2018). Corpus linguistics in language and sexuality studies. *Journal of Language and Sexuality*, 7(2): 145–174.
- Née E. and Fleury S. (2017). Chapitre III. Constituer un corpus en trois scénarios. In Née E, éditrice, *Méthodes et outils informatiques pour l'analyse des discours*. Presses universitaires de Rennes.
- Paveau M.-A. (2023). Une analyse du discours contre-hégémonique. *Langage et société*, 178(1): 161–190.
- Paveau M.-A. and Telep S., éditrices (2022). Penser la race dans les approches sociales du langage. *Langage & société*, 3(177).
- Pincemin B. (2018). Sept logiciels de textométrie. *HAL open science*. <https://shs.hal.science/halshs-01843695v1>
- Pincemin B. (2020). La textométrie, une approche quantitative et qualitative des données textuelles. Présentation, *Semaine Data SHS Traiter et analyser des données en sciences humaines et sociales*, MSH Lyon Saint-Étienne & Data infrastructure Progedo, Lyon, France.
- Prévost H., Figeac J., Salord T. and Cabanac G. (2021). Internet dans la reconfiguration des répertoires d'action du mouvement « féministe agroécologique » brésilien. *Politiques de communication*, 17(2): 129–156.
- Alim H.S., Rickford J.R. and Ball A.F. (2016). *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford Academic.

- Sitri F. and Barats C. (2017). Introduction. Pour comprendre ce qui va suivre... In Née E., éditrice, *Méthodes et outils informatiques pour l'analyse des discours*. Presses universitaires de Rennes.
- Telep L. (2019). « Whitiser, c'est parler comme un Blanc ». Langage, subjectivité et postcolonialité chez des militants afrodescendants d'origine camerounaise à Paris. Thèse de doctorat, Université Paris Cité.
- Tillewein H., Mohon-Doyle K. and Cox D. (2024). Critical Discourse Analysis of Sexual Violence Survivors and Censorship on the Social Media Platform TikTok. *Archives of Sexual Behavior*, 53(10): 3785–3794.

A la croisée des disciplines informatique et psychanalytique : une application sémiotique de la méthode Reinert

Claire Tissot¹

¹INRS – claire.tissot@inrs.fr

(Analyse réalisée dans un cadre personnel et indépendant des missions de mon institut)

Abstract

The Reinert method Alceste is a textual analysis method developed by the mathematician and sociologist Max Reinert in the 1980s. It is used by researchers in sociological and scientific fields to analyze open-ended survey questions or textual variables in databases. Reinert's work also intersects with semiotic research based on C. S. Peirce's theory of the sign and with psychoanalytic theory in exploring the possible connections between Lacanian registers and Peircean dimensions. The present study focuses on this expanded interpretations of classification results through Peircean and Lacanian concepts and the understanding of meaning in a text. According to Peirce, meaning emerges from the triadic dimensions of Icon–Index–Symbol. According to Lacan, the psychic structure is organized around the balance of the registers Imaginary–Real–Symbolic illustrated by his conception of the Borromean knot. The semiotician psychoanalyst Michel Balat has proposed a possible analogy between Peircean dimensions and Lacanian registers whose interweaving constitute the underlying framework of any discourse. Reinert's Hierarchical Descending Classification (HDC) becomes here relevant, allowing the separation and the identification of these three threads of discourse that are not immediately perceptible through enunciation. The present analysis deals with James Joyce's novel *Ulysses*. Lacan devoted his Seminar XXIII *The Sinthome* to this author in which he developed the hypothesis that Joyce lacked the Imaginary register, accounting for the loss of meaning readers often experience in his text. Is this hypothesis consistent with the HDC results? Do the resulting classes indeed confirm the absence of the Peircean iconic dimension/Lacanian Imaginary register ? The findings of this study highlight the relevance of the Reinert method as a tool for guiding or confirming hypotheses that go beyond a mere characterization of typologies.

Keywords: Natural Language Processing, Textmining, Reinert method, Peircean semiotics, Lacanian registers, Borromean knot, meaning, sign.

Résumé

La méthode Reinert Alceste est une méthode d'analyse textuelle élaborée par le mathématicien sociologue Max Reinert dans les années 1980. Elle est utilisée par les chercheurs de disciplines sociologiques ou scientifiques traitant les questions ouvertes de questionnaires ou les variables textuelles de bases de données. Les travaux de Max Reinert s'inscrivent également à la croisée des recherches sémiotiques sur les bases conceptuelles de la théorie du signe de Charles Sanders Peirce et psychanalytiques grâce aux rapprochements pouvant s'établir entre les registres lacaniens et les dimensions peirciennes. Le présent travail se situe dans cet élargissement de l'interprétation des résultats selon les concepts peirciens et lacaniens et le sens à donner à un texte. Selon Peirce, le sens émerge des dimensions icône-index-symbol. Selon Lacan, la structure psychique se construit autour d'un équilibre Imaginaire-Réel-Symbolique illustrée par sa conception du nœud borroméen. Le psychanalyste sémioticien Michel Balat a proposé une analogie possible entre les dimensions peirciennes et les registres lacaniens dont le nouage constitue la trame sous-jacente de tout énoncé. C'est ici qu'intervient la classification descendante hiérarchique (CDH) de Reinert qui permet la séparation et l'identification de ces trois brins du discours non repérables à la simple lecture d'un texte. La présente analyse traite du roman *Ulysses* de James Joyce. Lacan a

consacré le séminaire XXIII « Le sinthome » à cet auteur. Il y développe l'hypothèse de l'absence, chez Joyce, du registre Imaginaire pour expliquer la perte de sens ressentie à la lecture du texte. Cette hypothèse est-elle cohérente avec les résultats de la CDH ? Les classes obtenues confirment-elles l'absence de dimension iconique peircienne/registre Imaginaire lacanien ? Les résultats obtenus montrent l'intérêt de la méthode Reinert comme outil d'orientation ou de confirmation d'hypothèses dépassant la simple caractérisation de typologies.

Mots clés : traitement automatique du langage, fouille de texte, méthode Reinert, sémiotique peircienne, registres lacaniens, nœud borroméen, sens, signe.

1. Introduction

La méthode Reinert Alceste est une méthode d'analyse textuelle élaborée par le mathématicien sociologue Max Reinert (1944-2024) dans les années 1980. Une large bibliographie retrace les applications de cet algorithme à des disciplines sociologiques ou scientifiques traitant les questions ouvertes de questionnaires ou les variables textuelles de bases de données. Il procède à une classification automatique de textes en s'appuyant sur son algorithme particulier de classification descendante hiérarchique (CDH) et non ascendante comme la plupart des algorithmes de classification. Son objectif est de regrouper dans une même classe des fragments d'énoncés constitués des mêmes répétitions de co-occurrences et étant les plus différents de ceux des autres classes (Reinert, 1983). Ces répétitions correspondent aux tendances d'un auteur à revenir sans cesse, sans le choisir, sur une même position d'énonciation. Au cours de ses recherches, Max Reinert a été amené à établir une correspondance entre les classes fournies par la CDH et les trois dimensions sémiotiques de C. S. Peirce (1839-1914) : l'icône, l'indice, le symbole (Reinert, 2001, 2003, 2007 ; Peirce, 1978).

Dans ses dix dernières années, Jacques Lacan (1901-1981) a développé le concept de nœud borroméen notamment à partir de l'œuvre de James Joyce. Il connaissait le travail de Peirce et l'a nommé à plusieurs reprises dans ses séminaires. Ses réflexions l'ont fait passer d'une conception structuraliste du psychisme (névrose/psychose/perversion) à une approche borroméenne dynamique basée sur le nouage de trois brins : Imaginaire (I), Réel (R) et Symbolique (S). A partir d'un passage du roman de Joyce *A Portrait of the Artist as a Young Man*, Lacan fait l'hypothèse d'un rapport particulier de l'auteur à son corps. Ce passage relate une agression physique de Stephen par ses camarades, agression qui ne l'affecte pas, ayant la sensation que son corps se détache comme une pelure d'oignon. Lacan en déduit une déficience du registre Imaginaire. Ce registre Imaginaire représente pour Lacan le rapport au corps : idée de soi comme corps, le corps en tant qu'image, et constitue le fondement de l'ego et du narcissisme. Pour garder une stabilité mentale, Joyce a dû se forger un ego particulier reliant le brin Imaginaire aux deux autres brins Réel et Symbolique, restituant le nouage complet du nœud (Lacan, 2005).

Michel Balat (Balat, 2000) a étudié de manière très approfondie les rapprochements possibles entre les dimensions peirciennes et les trois registres lacaniens. Un pont peut ainsi être établi entre les résultats de la méthode Reinert et les hypothèses lacaniennes suggérées à propos d'*Ulysses* dans le séminaire XXIII : si le rapport au corps (Imaginaire) est dénoué de la réalité psychique de Joyce, existerait-il une dimension iconique dans les classes obtenues avec la méthode Reinert appliquée à *Ulysses* ? Outre la difficulté d'approche des textes joyciens, la réponse à cette question a été le moteur principal de la présente analyse.

2. Mise en œuvre de la méthode Reinert

Le texte *Ulysses* a été téléchargé du site Gutenberg en juin 2021 au format texte brut UTF-8 (eBook #4300). C'est le texte original en anglais qui a été analysé et toutes les sorties graphiques présentent les mots en anglais.

Alceste existe en open source dans l'environnement de programmation R avec le package rainette. A la différence de la version commerciale Alceste qui fournit tout un processus d'analyse, rainette ne procède qu'à la classification descendante hiérarchique. Tous les autres traitements sont à programmer dans R.

Le présent article est un extrait d'une étude complète disponible en ligne sur le site « Textmining et sémiotique » (Tissot, 2022).

2.1. Préparation du texte

Ulysses est composé de 18 chapitres répartis en trois parties. La méthode Reinert commence par découper le texte de façon aléatoire en segments d'environ 40 mots. Le corpus obtenu comprend 6 806 segments de texte. Les traitements habituels de *textmining* sont appliqués à ce corpus. Si les résultats sont satisfaisants sans lemmatisation, les regroupements féminin/masculin, singulier/pluriel, etc. ne sont pas appliqués, ce qui était le cas ici. Une tokenisation transforme les segments en suite de mots isolés. Une étape particulièrement importante pour la méthode Reinert est la définition de « mots pleins » et de « mots-outils ». Les mots pleins (noms, verbes, adjectifs, adverbes) sont utilisés pour la CDH alors que les mots-outils aident *a posteriori* à identifier les dimensions peirciennes des classes. Il s'agit d'articles, de pronoms, de conjonctions, des verbes « être » et « avoir », des verbes modaux « devoir », « vouloir », etc. Dans cette analyse les mots-outils ont été définis à partir de la liste standard des *stopwords* anglais et également celle des *stopwords* français car *Ulysses* contient un grand nombre de mots français. Certains traitements complémentaires peuvent être nécessaires pour obtenir un texte propre : identifier des mots soit à ajouter à la liste des mots-outils, soit à définir dans une vraie liste de *stopwords*. Ces repérages sont assez longs et jamais vraiment terminés. Sur 318 907 mots du lexique initial, 28 850 mots pleins sont gardés pour les analyses et 398 mots-outils mis de côté. Un filtrage supplémentaire est effectué sur les mots pleins encore trop nombreux : on garde ceux présents au moins cinq fois, ce qui réduit le lexique à 5 027 mots. Le tableau lexical final croise les 6 806 segments de texte en ligne et les 5 027 mots en colonne. La matrice est prête pour la classification descendante hiérarchique.

```
> dfmRainTrim
Document-feature matrix of: 6,806 documents, 5,027 features (99.71% sparse) and 13 docvars.
features
docs      plump buck mulligan came bearing bowl mirror razor lay crossed
text1_1    1    1      1      1      1    2      1    1    1    1
text1_2    0    0      0      1      0    0      0    0    0    0
text1_3    0    0      0      0      0    0      0    0    0    0
text1_4    0    0      0      0      0    0      0    0    0    0
text1_5    0    1      1      0      0    1      1    0    0    0
text1_6    0    0      0      0      0    0      0    0    0    0
[ reached max ndoc ... 6,800 more documents, reached max nfeat ... 5,017 more features ]
```

2.2. Classification descendante hiérarchique

Une double classification est généralement réalisée pour optimiser la robustesse des résultats en faisant varier la taille minimum des segments de texte et en ne gardant que les classes communes aux deux classifications. En effet, la scission en segments d'environ 40 mots est souple et la suppression des mots-outils ou des mots rares modifie le contenu des segments à analyser. En faisant varier le nombre minimum de mots par segments, on remanie les frontières

de ces segments. Le fait de garder les classes communes aux deux classifications fournit les classes les plus robustes. Le choix de la taille des segments est une démarche empirique qui dépend des résultats des classifications. Dans le cas présent, les seuils ont été fixés à 8 et 12. Dans l'optique d'une analyse sémiotique, on garde en général trois ou quatre classes pouvant correspondre aux dimensions peirciennes. Le taux de classement en trois classes atteint ici 84 % avec 1 113 segments non classés (« NA » sur la figure 1). Malgré la proportion importante de segments non classés, c'est cette partition qui a été gardée car elle donnait le meilleur compromis entre la répartition et l'interprétabilité des classes. Les segments non classés sont exclus des analyses ultérieures.

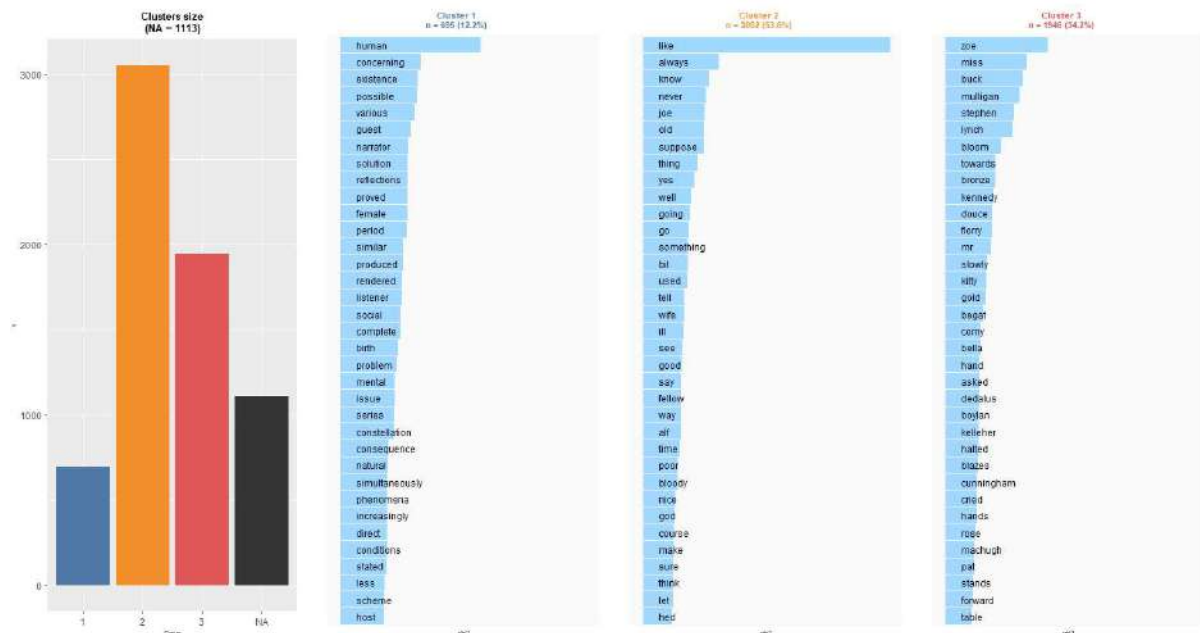


Figure 1. Double classification méthode Reinert : trois classes de segments de texte caractérisées par les mots pleins

2.3. Traitements post-CDH

A ce stade, nous disposons de trois classes de segments de texte caractérisés par les mots pleins. Le package *rainette* permet d'obtenir un fichier contenant les segments de texte, les variables du fichier de départ (parties, chapitres) auquel on peut ajouter la classe sémiotique sous forme de variable nominale. A partir de ce nouveau fichier, toutes les analyses et illustrations graphiques de R sont possibles, notamment les traitements décrits par Reinert dans ses analyses *Alceste* (Reinert, 2001, 2003, 2007).

Le package *wordcloud2* est utilisé pour représenter visuellement les mots caractéristiques des classes. Contrairement aux nuages de mots habituels basés sur la fréquence des mots, la taille des mots est ici liée à leur caractérisation au sens du khi-deux (Escofier et Pages, 1998) : plus la valeur-test est élevée, plus le mot est caractéristique d'une classe et plus sa taille est grande. Les gradients observés dans les nuages de mots résultent donc d'un comparatif des mots entre les différentes classes. Pour une meilleure lisibilité, on ne garde que les 50 mots pleins les plus caractéristiques d'une classe. Les mots-outils étant moins nombreux, on les garde tous. Deux exemples de phrases caractéristiques sont listés pour chacune des classes. Des traitements supplémentaires, non présentés ici, peuvent compléter cette caractérisation : une analyse factorielle des correspondances sur le tableau classes x mots permet le positionnement relatif des classes selon les mots qu'elles contiennent et la visualisation des mots contributifs d'un axe. Une analyse des correspondances multiples sur le fichier contenant en ligne les segments

et en colonne les classes, les parties, les chapitres, et les mots pleins en illustratif, permet d'observer l'évolution des classes tout au long du récit.

3. Résultats de la méthode Reinert

3.1. Description des classes

Chaque classe est décrite par les mots pleins et les mots-outils caractéristiques sous forme de nuages de mots. L'ordre des mots des nuages ci-dessous et des listes de la figure 1 peut différer selon la fonction utilisée : catdes du package FactoMiner (utilisée ici), textstat_keyness du package quanteda, ou autre. La méthode Reinert étant appliquée dans une perspective sémiotique, les classes sont identifiées selon leur dimension peircienne. Cet « étiquetage » est à la fois analytique et synthétique : il s'appuie sur le contenu des classes suggéré par les mots pleins et procède du fait que telle classe est séparée des autres. Une certaine subjectivité entre nécessairement en jeu, liée à la manière de chacun de résonner à ce que Max Reinert nomme les « mondes lexicaux ».

3.1.1. Classe 1 : dimension symbolique : réflexion et métaphysique

Cette classe avec 695 segments de texte est la plus petite, se différencie la première et est la plus différente des autres. Le nuage des mots pleins (cf. figure 2 gauche), avec notamment le mot le plus caractéristique « human » qui en donne d'emblée sa tonalité, illustre la dimension symbolique de cette classe. Différents univers sont évoqués (par les mots de la figure 2 et/ou par ceux identifiés par les analyses factorielles non présentées ici) :

- un univers mental, la conscience, des problèmes sociaux à résoudre : « conditions », « social », « female » ;
- des réflexions philosophiques avec une argumentation logique : « proved », « reflections », « solution », « possible », « scheme », « conclusions » ;
- une notion très importante dans la vie de James Joyce, celle de départ et d'itinérance, suggérée par les termes « guests », « host », « acquaintances » ;
- une dimension cosmologique : « constellation », « terrestrial », « solar » ;
- une dimension temporelle et cyclique à un niveau planétaire ou personnel : « period », « former », « successively », « ultimate », « extending », « temporary », « years », « future », « final », « previously », « development », « origin », « birth », « existence », « death » ;
- les écrits et leur transmission, évoqués par les mots « narrator », « listener », « book », « prospectus ».

Le nuage des mots-outils, figure 2 (droite), indique également une temporalité « been », « did », des comparaisons (« between », « other »).

Les valeurs-test des mots pleins s'échelonnent de 12,68 pour « human » à 7,27 pour « conclusion », celles des mots-outils de 25,71 pour « of » à 2,24 pour « each ».

Tableau 1. Exemples de segments de texte caractéristiques de la classe 1

By various reiterated feminine interrogation concerning the masculine destination whither, the place where, the time at which, the duration for which, the object with which in the case of temporary absences, projected or effected. What moved visibly above the listener's and the narrator's invisible thoughts? (text9_571)
There remained the generic conditions imposed by natural, as distinct from human law, as integral parts of the human whole: the necessity of destruction to procure alimentary

sustenance: the painful character of the ultimate functions of separate existence, the agonies of birth and death: (text9_252)



Figure 2. Classe 1 symbolique. A gauche les mots pleins, à droite les mots-outils

3.1.2. Classe 2 : dimension indiciaire réel-affect : émotion, intimité, monologues intérieurs

La classe 2 est la plus grande classe avec plus de la moitié des segments de texte (3 052). Le nuage des mots pleins de la figure 3 (gauche) montre les mots les plus représentatifs dénotant l'aspect émotif et affectif de cette classe. Le mot « like » est le mot le plus caractéristique avec 596 occurrences, toutefois il a une double signification : il correspond au verbe « aimer » (souvent associé dans ce texte au domaine alimentaire) mais est également un comparatif (« like that », « look like »). Cette classe 2 comprend des adverbes de fréquence évoquant une approbation ou un rejet (« always », « never », « ever »), des adjectifs positifs (« well », « good », « nice », « sure ») ou négatif (« bad »). Les adjectifs « poor » ou « little » sont souvent employés dans un contexte affectif (« poor papa too », « poor fellow », « poor little naughty boy »). Les dialogues sont évoqués avec les termes « tell », « say », « told », « talking », « says » et les monologues avec les mots « know », « think », « suppose », « wonder », « imagine », « thought », « remember », « forget ». Les mots « thing », « something », « nothing » caractéristiques de cette classe seront décrits ultérieurement. Les mots-outils de la figure 3 (à droite) évoquent la notion indiciaire d'altérité avec les pronoms personnels et réfléchis (« I », « my », « myself », « you », « he », « they ») ainsi que les pronoms démonstratifs (« that », « those »).

Tableau 2. Exemples de segments de texte caractéristiques de la classe 2

Those are nice things, says the citizen, coming over here to Ireland filling the country with bugs. So Bloom lets on he heard nothing and he starts talking with Joe, telling him he needn't trouble about that little matter till the first but if he would just say a word to Mr Crawford. (text4_328)



Les valeurs-test des mots pleins s'échelonnent de 19,06 pour « like » à 5,81 pour « ever », celles des mots-outils de 19,67 pour « I » à 5,27 pour « theres ».

Le nuage des mots pleins de la figure 4 (gauche) montre la prépondérance des noms propres dans cette classe 3 : Bloom (Mr Bloom, Léopold Bloom, Marion Bloom), Stephen (Dedalus), Buck Mulligan et Lynch, les colocataires de Stephen Dedalus. Les femmes évoquées dans cette classe sont les prostituées de la maison close : Bella (Cohen), Zoe (Higgins), Kitty et Florry, ou bien les serveuses-sirènes de l'hôtel Ormond : miss Kennedy et miss Douce. D'autres personnages sont mentionnés : Blazes Boylan l'amant de Marion Bloom, le professeur MacHug, Martin Cunningham l'ami de Leopold Bloom. Dans cette classe, les hommes et les femmes se rencontrent dans un cadre social (café, hôtel, maison close) qui évoque des rapports en lien avec le corps : les fonctions physiologiques ou sexuelles.

Tableau 3. Exemples de segments de texte caractéristiques de la classe 3

Encore, enclap, said, cried, clapped all, Ben Dollard, Lydia Douce, George Lidwell, Pat, Mina Kennedy, two gentlemen with two tankards, Cowley, first gent with tank and bronze Miss



Les valeurs-test des mots pleins s'échelonnent de 11,93 pour « zoe » à 6,33 pour « professor », celles des mots-outils de 12,89 pour « his » à 2,01 pour « the ».

- des parties du corps : « eyes », « face », « lips », « hands », « fingers », « feet », « shoulder », « neck » ;
- la fonction visuelle suggérée par la fréquence du mot « eyes » et les verbes « see », « watched », « peered », « glanced » — comme un regard cherchant à percer le brouillard. Comment ne pas évoquer James Joyce et ses problèmes de vue ?
- des descriptions vestimentaires suggérant le gentleman : « hat », « silk hat », « cap », « ashplant », « cane » ;
- la description de métaux : « bronze », « gold », « silver », « azure » . Les mots « marble », « copper », « mosaïc », « coral », « cristal », « diamond », « stone » sont fortement contributifs de l'axe porté par cette classe ;

- la mention de parties de bâtiment, de passages évoquant la ville : « doorway », « door », « steps ». Les mots « bridge », « porch », « staircase », « halldoor », « quay » ont également une forte contribution sur l'axe associé à cette classe 3 sur les graphiques factoriels.
- des couleurs sont mentionnées mais elles apparaissent comme des miroitements métalliques (« gold », « blue », « white », « gray », « green »).

La présente analyse a été réalisée en 2022. Des analyses complémentaires ont été effectuées en 2026 en complétant la liste des mots-outils par certains mots utilisés ici en tant que mots pleins. Les résultats sont améliorés dans le sens d'un rééquilibrage des classes (moins de NA et plus de classe 1 symbolique). Une lemmatisation a également été testée. Ces essais fournissent les mêmes dimensions symbolique, indice-réel-affect et indice-réel socius. Une CDH a été appliquée à la seule classe 3 qui contient des traces iconiques. Les résultats d'une partition en trois ou quatre classes évoquent tous la dimension indiciaire réel-socius. Il semble donc bien que cette scission symbolique/indice-réel affect/indice-réel socius soit stable et constitutive des postures énonciatives de l'auteur.

4. L'hypothèse de Jacques Lacan et la sémiotique de Peirce

On a donc une réponse à la question relative à l'émergence d'une classe iconique : cette dimension du corps, des sens, des résonances avec le passé ne se différencie pas de manière autonome par des répétitions marquées de mots qui signifieraient une posture franche de l'auteur vis-à-vis de cette dimension. Elle apparaît plutôt mêlée avec la dimension du réel-socius, en relation avec « des autres ». Ce résultat rejoint l'hypothèse lacanienne du brin Imaginaire déficient et dénoué du nœud borroméen.

Dans la sémiotique de Peirce, un objet s'impose au sujet via une priméité (icône). Le sujet réagit à cette priméité par une rupture et une réaction (secondéité : indice). Pour atténuer les affects, comprendre la situation, prendre de la distance par rapport à l'expérience vécue, une terciéité (interprétant, symbole) élabore des lois et des habitudes, nomme les objets grâce au langage, aux écrits, à l'information, élabore des sciences et des philosophies.

Dans *Ulysses*, la présence de l'objet a des effets réels sur le corps qui, bien que non exprimés iconiquement, provoquent des réactions et la captation d'un regard : l'auteur cherche à voir dans la classe 3, à ressentir/dire dans la classe 2. Cette classe 2 montre bien une posture cherchant à définir quelque chose du réel à travers les mots caractéristiques « thing », « something », « nothing », « anything », « everything », « that », « mean », « meaning », « meant ». C'est finalement dans la classe 3 que l'objet sera le mieux perçu, notamment visuellement avec le désir de percer une réalité opaque et à travers une dimension sociale (réel-socius).

L'observation de l'évolution des classes tout au long du récit (Tissot, 2022) montre une constante oscillation des classes 2 et 3 dans la partie centrale (chapitres 4 à 15) entre l'objet pressenti et l'objet vu, l'objet en soi et l'objet hors de soi, et une recherche de l'objet en sortant hors de soi (exil). C'est finalement un positionnement symbolique final qui permet une prise de distance et une re-situation du sujet dans son histoire, parallèle à celle du cosmos dans un cycle de répétitions.

5. Discussion

La méthode Reinert cherche à définir le sujet d'un discours et l'articulation de postures changeantes plutôt que l'objet d'un contenu. Elle ignore la construction des phrases et leur structure grammaticale (on parle de « sac de mots ») et ne prend en compte que les répétitions de co-occurrences les plus marquées. Les mots rares sont donc éliminés dès le départ ce qui

ampute de texte de Joyce de tous les mots-valises qu'il créait avec délectation ou des mots empruntés à des langues étrangères, et qui supprime également tous les effets sonores ou les jeux de mots liés à la linéarité des phrases. Si la méthode Reinert ne permet pas non plus d'identifier l'absence de message lié au branchement direct du Réel sur le Symbolique se traduisant par un texte énigmatique tel que décrit par Jacques Lacan, elle peut mettre en évidence l'absence de dimension iconique peircienne, cohérente avec l'hypothèse lacanienne d'un dénouage du brin Imaginaire. Les trois dimensions icône, indice, symbole ne sont pas identifiables à la lecture d'un texte court, d'un poème et, à plus forte raison d'un long roman comme *Ulysses*. En faisant émerger les répétitions de co-occurrences les plus fréquentes, la CDH révèle ce que Max Reinert nomme les « mondes lexicaux » et, par analogie, les trois dimensions peirciennes du sens et les trois brins du nœud borroméen. La convergence de différentes disciplines comme l'analyse textuelle, la sémiotique, la psychanalyse, constitue un guide dans l'exploration et la tentative de compréhension des conditions d'écriture d'un texte ou de celles de toute énonciation.

Références

- Balat M. (2000). *Des fondements sémiotiques de la psychanalyse. Pierce après Freud et Lacan*. L'Harmattan.
- Escofier B., Pages J. (1998). *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation* ». Paris. Dunod, 3e édition, 284 p.
- Lacan J. (2005). *Le séminaire livre XXIII, le sinthome*. Champ freudien. Texte établi par Jacques-Alain Miller. Éditions du Seuil.
- Peirce C.-S. (1978). *Écrits sur le signe*. Traduit et commenté par G. Deladalle. Edition du Seuil.
- Reinert M. (1983). *Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte*. Cahiers de l'analyse des données. Volume 8, Numéro 2.
- Reinert M. (2001). *Alceste, une méthode statistique et sémiotique d'analyse de discours ; application aux Réveries du promeneur solitaire*. La Revue Française de Psychiatrie et de Psychologie Médicale, V, 49, pp. 32-36.
- Reinert M. (2001). *Approche statistique et problème du sens dans une enquête ouverte*. Journal de la société statistique de Paris. Tome 142, n°4 : 59-71.
- Reinert M. (2003). *Le rôle de la répétition dans la représentation du sens et son approche statistique par la méthode Alceste*. Semiotica 147-1/4 : 389-420.
- Reinert M. (2007). *Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours*. Éditions de la Maison des sciences de l'homme. Langage et société, n°121-122 : 189 à 202.
- Reinert M. (2008). *Mondes lexicaux stabilisés et analyses statistiques de discours*. JADT 2008 : 9ème Journées internationales d'analyse statistique des données textuelles.
- Tissot C. (2022). *Textmining et sémiotique*.
<https://clairetissotsemiotique.wordpress.com/2022/06/11/analyse-textuelle-semiotique-du-roman-ulysses-de-james-joyce-destructurer-la-destructure>.

De la mesure des paragraphes à la classification des visualisations : portée épistémologique

Claude Patricia Tardif

Université Paris 8 – claudeclpro@gmail.com

Abstract

The visual segmentation of text by its paragraphs remains underexploited in quantitative analyses. This contribution proposes to explore it through the classification of data visualizations, conceived not as a mere final illustration, but as a heuristic investigative framework. By classifying these graphical objects across two text corpora, we demonstrate how the classification of data visualizations serves as a lever to reveal text properties and support research hypotheses. Our protocol proceeds in three phases: measuring paragraph lengths in a text via the *Narra 2.0* software, algorithmically generating visualizations from these metrics, and classifying them by their morphology. The first corpus of visualizations (244 literary texts by 16 authors) reveals a specific textual rhythm driven by paragraph length distribution, which results from the frequency of changes, offering a diachronic view of sixteen authorial dynamics. The second corpus of visualizations (1,255 texts by 162 authors) shifts to a macro-analytical scale, demonstrating that the visual text segmentation by its paragraphs correlates with both literary genre and historical period. Ultimately, identifying rhythmic regimes at intra- and inter-authorial levels opens new quantitative lens for stylometry and examination of this uncovered narrative rhythm.

Keywords: Textual Segmentation, Paragraph Length, Distant Reading, Data Visualization, Classification, Rhythm.

Résumé

La segmentation visuelle du texte par ses paragraphes demeure une caractéristique textuelle sous-exploitée dans les analyses quantitatives. Cette communication propose de l'explorer à travers la classification de visualisations de données, conçue non comme une simple illustration finale, mais comme un dispositif d'investigation heuristique. S'appuyant sur deux corpus, nous démontrons comment la classification de visualisations constitue un levier pour révéler des propriétés des textes et étayer des hypothèses de recherche. Le protocole opératoire se déroule en trois phases : la mesure des longueurs des paragraphes d'un texte *via* le logiciel *Narra 2.0*, la génération algorithmique des visualisations à partir de ces mesures, puis leur classification morphologique. Le premier corpus de visualisations de 244 textes littéraires de 16 auteur.e.s révèle un rythme textuel en fonction de la distribution des longueurs des paragraphes, qui résulte de la fréquence des changements, offrant une vue diachronique de seize dynamiques auctoriales. Le second corpus de visualisations de 1 255 textes de 162 auteur.e.s à une échelle macro-analytique démontre que cette segmentation visuelle des textes par leurs paragraphes est corrélée au genre littéraire ou à la période d'écriture. En identifiant des régimes rythmiques aux échelles intra- et inter-auctoriales, la classification de visualisations de données textuelles ouvre de nouvelles perspectives quantitatives pour la stylométrie et l'analyse de ce rythme narratif mis au jour.

Mots clés : Segmentation textuelle, longueur des paragraphes, Distant Reading, visualisation de données, classification, rythme.

1. Introduction : enjeux heuristique et démonstratif de la classification de visualisations de données

La segmentation visuelle du texte littéraire par ses paragraphes demeure une caractéristique textuelle négligée dans les analyses quantitatives. Cette contribution propose de l'explorer à travers la classification de visualisations de données, conçue comme un instrument d'investigation doté d'une double portée épistémologique, qui s'articule autour d'une :

(i) fonction heuristique, fondée sur une première classification de visualisations de données issues d'un corpus de 244 textes de 16 auteurs qui révèle des caractéristiques inexplorées du texte telles qu'un rythme visuel généré par les rapports de la longueur/durée de ses segments (les paragraphes) ainsi qu'une éventuelle structure compositionnelle. Ces propriétés mises au jour, qui se manifestent tant à l'échelle de l'œuvre qu'à celle diachronique de l'auteur, échappent aux méthodes de fouille textuelle classiques.

(ii) capacité démonstrative, fondée sur une seconde classification de visualisations de données issues d'un corpus de 1 255 textes de 162 auteurs qui offre une perspective macro-analytique de cette dimension textuelle et permet d'éprouver des hypothèses de recherche, celles d'une corrélation entre la longueur des paragraphes et le genre littéraire ou l'époque.

Chaque visualisation est un texte, dans sa totalité, réalisée en fonction de son séquençage visuel, soit de sa segmentation paragraphique. Le texte est appréhendé en tant que séquence de segments (paragraphes) et sa forme en tant que processus dynamique se déployant au rythme de la longueur/durée variables de ses paragraphes, c'est-à-dire en tant que série temporelle. L'étude explore ainsi les rapports de longueurs/durées des paragraphes d'un texte.

1.1. Fondements terminologiques et conceptuels des notions mobilisées

Nous définirons les notions clefs de texte, paragraphe et rythme en les distinguant des concepts de visualisation et de classification employés dans leur acception courante.

L'usage du terme « texte » se limite ici à sa matérialité écrite et celui de « littéraire » aux romans, nouvelles et contes, à l'exception des textes tels que *La Maison des feuilles* de Mark Z. Danielewski (2002). Si son étymologie remonte au latin *textus*, « tissu ; trame ; enchaînement », du verbe *texere* « tisser », qui « récuse le dogme de la linéarité » (Dahan-Gaida, 2023, p. 298), nous nous intéressons pourtant à sa linéarité, mais au sens d'enchaînement de segments visuellement repérables sans lire le texte (les paragraphes – nous ne traiterons pas des chapitres). Dès lors que l'on a un enchaînement de longueurs ou de durées variables non permutables, la notion de temps intervient inévitablement. Aussi, l'approche initialement visuelle des textes littéraires s'est déplacée vers leur dimension temporelle.

Par « paragraphes », nous entendons les paragraphes narratifs et dialogaux, marqués par une « fin de ligne » souvent appelé « saut de ligne ». Afin de préserver la linéarité textuelle, nous avons choisi de ne pas les distinguer, car cela créerait deux groupes textuels discontinus. Les paragraphes ne sont pas qu'un signe graphique ; ils sont une ponctuation qui signale un changement. « Les repères de paragraphes sont informatifs » (Stark, 1988, p. 299). En effet, les paragraphes obéissent à la logique selon laquelle « la pensée [...] commande le découpage » (Castille, 2012, p. 150). Le paragraphe narratif est, par nature, une unité thématique. « Si les paragraphes nous permettent de "voir" des thèmes, c'est parce que les thèmes, n'"existent" pleinement qu'à l'échelle du paragraphe » (Algee-Hewitt *et al.*, 2015, p. 215). « Si on utilise la modélisation thématique pour analyser la littérature, le paragraphe fournit une unité plus efficace que tout autre segment "mécanique" » (Moretti, 2016, p. 196). « C'est le principe thématique qui détermine d'une manière décisive la segmentation opérée » (Achard-Bayle &

Pešek, 2020, p. 20). En outre, « la disposition graphique du texte en reflète la structure compositionnelle ("rhétorique") qui est ainsi signalée et donc communiquée au récepteur du texte. » (*Ibid.*, 2020, p. 30). L'étymologie du terme « paragraphe » porte déjà ce sens : du latin *paragraphus*, il est une "marque pour distinguer les différentes parties d'un exposé" [...] Le mot [...] signale la fin d'un développement et désigne, en grec tardif, la section d'un texte terminée par un tel signe, qui correspond à une unité du développement » (Rey, 2001, p. 2428).

La notion de « rythme » fait référence à la conception qui prévalait chez les Présocratiques, telle que décrite par Émile Benveniste (1966). « Rythme, dérivé de *rhein* « couler », désigne « littéralement "manière particulière de fluer" » (1966, p. 333). Il s'entend ici de « *rhuthmos*, au sens de "modalité d'un accomplissement" » (Michon, 2018, p. 25), de « perception d'une forme dynamique » (Wulf, 2010, p. 28), de forme temporelle distinctive en son mouvement, avant que la tradition platonicienne y intègre une structure réursive identique à elle-même telle que conçue encore aujourd'hui, entre autres, par Pierre Sauvanet qui le théorise selon au moins deux des qualités suivantes : « structure, périodicité, mouvement » (1992, p. 238).

1.2. Aperçu généalogique de l'atlas et présentation du corpus textuel des deux classifications de visualisations

La classification d'images s'inscrit dans une longue tradition savante qui, à partir du XVIII^e siècle, considère l'image comme un vecteur de connaissance susceptible d'être organisée et comparée. Malgré l'étendue du savoir humain, elle n'a toutefois pas été abandonnée. Citons, par exemple, *Galaxy*, un projet du programme *Zoo Sloan Digital Sky Survey* qui classe les galaxies selon leur forme et dont l'objectif est de valider les modèles galactiques proposés par les scientifiques et le projet *Human Cell Atlas* dont l'objectif est de cartographier toutes les cellules du corps humain pour enrichir notre compréhension de l'unité de vie la plus fondamentale et ses principes. Mentionnons une analyse de la littérature, proche de notre concept, qui se fonde également sur la visualisation de données (Keim & Oelke, 2007).

Le corpus de la première classification se compose de 16 *leporellos* comprenant au total, sous la forme d'un sismogramme, 244 textes parmi les plus célèbres de 16 auteurs occidentaux. Chaque *leporello*, un par auteur, inclut entre 7 à 26 textes classés par ordre chronologique de leur première publication.

Le corpus de la seconde classification est un atlas comprenant, sous la forme d'un spectre de couleurs, 1 255 textes littéraires de 162 auteurs, la plupart occidentaux et célèbres, dont la liste est présentée dans les premières pages de l'*Atlas*. 72 % des textes de l'*Atlas* ont été publiés entre 1800 et 2000, 4 % avant 1800 et 14 % après 2000. Le corpus des textes se répartit ainsi selon les genres littéraires suivants :

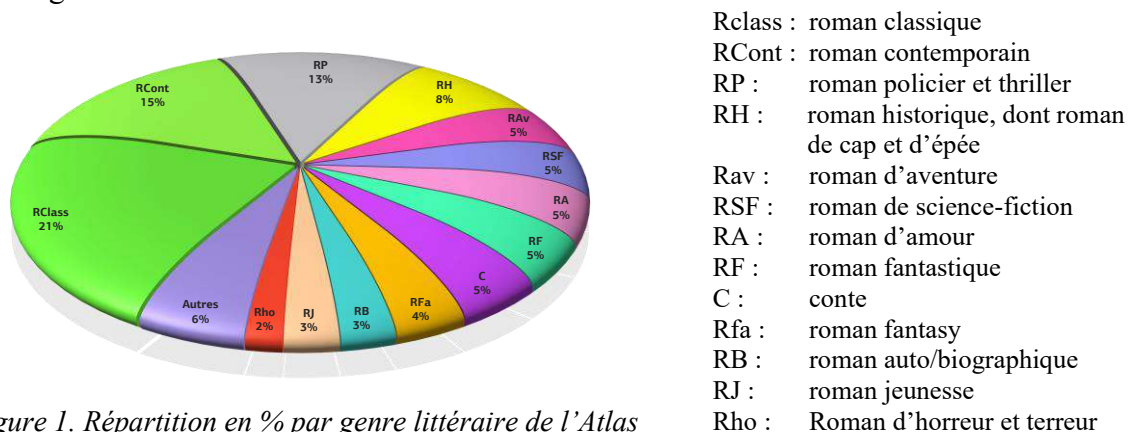


Figure 1. Répartition en % par genre littéraire de l'Atlas

2. Protocole numérique opératoire : de la quantification des données aux visualisations et leur classification

Le protocole numérique se déroule en trois phases.

2.1 Mesurer la longueur des paragraphes d'un texte à l'aide d'un logiciel Narra 2.0

La mesure de la longueur des paragraphes est effectuée avec le logiciel applicatif *Narra 2.0* en nombre de caractères (ou signes), unité plus précise que le mot. Pour une appréhension immédiate de leur taille, ces mesures sont ensuite converties en nombre de lignes sur l'équivalence de 60 caractères vaut une ligne. En effet, par exemple, 2 lignes est plus parlant que 120 caractères. Que l'on prenne 50, 60 ou 80 caractères pour une ligne, les rapports entre les mesures restent identiques.

En entrée, l'interface du logiciel propose de choisir un texte (sous formats .ePub ou .doc, les seuls acceptés par le logiciel) qu'il va parcourir pour générer en sortie :

- en arrière-plan de l'interface, un fichier csv comprenant les mesures de chaque paragraphe du texte sélectionné – autant de nombres que de paragraphes –, ainsi que des données statistiques : des valeurs brutes, des indicateurs de tendance centrale, un indicateur de dispersion ;
- dans l'interface du logiciel, un histogramme en fonction de la longueur des paragraphes du texte sélectionné et une courbe des fréquences comparée à celle de la loi de Zipf.

2.2. Visualiser le texte selon deux modes

À partir du fichier csv obtenu par texte sélectionné, qui comprend toutes les mesures, les visualisations des deux classifications sont réalisées chacune avec un algorithme *ad hoc*.

2.2.1. Le sismogramme pour la première classification : une visualisation séquentielle de la longueur des paragraphes

Les sismogrammes qui composent la première classification comprennent autant de lignes verticales que de paragraphes dans le texte, dont la largeur est fixe à 1 pixel et dont la hauteur est proportionnelle à la longueur de chaque paragraphe. La largeur du sismogramme dépend ainsi du nombre de paragraphes dans le texte. Un sismogramme court peut être un texte long si ses paragraphes sont très longs, et inversement.

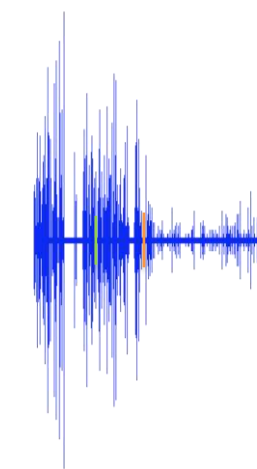


Figure 2.
Sismogramme
Accident nocturne

La lecture de *Sismogramme Accident nocturne - Modiano* s'effectue de gauche à droite, du premier paragraphe au dernier. Elle invite à la prudence. En effet, le centre du graphique, situé au niveau du paragraphe central (signalé par un trait orange) ne correspond pas au le milieu du texte (marqué par un trait vert), plus à gauche, puisque les paragraphes les plus longs se situent au début du texte.

Outre un rythme spécifique du texte mis au jour, le sismogramme met en évidence une structure temporelle en deux ou trois temps. Deux, si l'on considère les deux grandes amplitudes comme formant une unité, d'autant que leur séparation (une suite de paragraphes de 1 ligne) correspond à une liste. Le changement radical entre paragraphes longs et courts se produit exactement au niveau du paragraphe central, soit aux trois quarts du texte. Cette rupture visuelle coïncide avec un moment de basculement dans le récit, celui de la seconde rencontre – tant attendue –, où la part des dialogues est prépondérante.

2.2.2. Le spectre pour la seconde classification : une visualisation fréquentielle des paragraphes selon leur longueur

Chaque texte s'identifie par sa « fréquence spectrale ». Les spectres qui composent la seconde classification sont générés à partir de la distribution fréquentielle des paragraphes en fonction de leur longueur. Contrairement aux sismogrammes, quels que soient le nombre de paragraphes et la taille du texte, un spectre mesure 7,5 cm de long. La couleur est ici la seule variable visuelle. Elle permet d'appréhender immédiatement, par différence chromatique, la proportion des paragraphes courts, moyens ou longs dans un texte sur la base d'un code couleur prédéfini selon lequel le jaune-vert, point médian du spectre de la lumière visible, correspondra à la longueur moyenne des paragraphes, soit de 6 lignes sur un échantillon de 250 textes, donnant : paragraphe de 1 ligne = rouge..., 3 lignes = orange ..., 5 lignes = jaune ..., 7 lignes = vert ..., 9 lignes = turquoise ..., 10 lignes et plus = bleu.

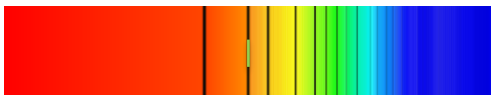


Figure 3. Spectre Accident nocturne - Modiano

La lecture de *Spectre Accident nocturne - Modiano* s'effectue en fonction des couleurs et des raies noires qui les délimitent. La moitié du spectre (marqué d'un trait vert), soit la moitié du nombre total des paragraphes, se compose de paragraphes très courts de 1 et 2 lignes, 40 % de paragraphes de 1 ligne et 20 % de longs paragraphes de plus de 10 lignes.

2.3. Classifier les visualisations obtenues

La première classification se fonde sur les sismogrammes, regroupés par auteur dans un *leporello* (ou livre accordéon) et classés selon l'ordre chronologique de première parution des textes. Au total, cette classification comprend 16 *leporellos*, un par auteur. La hauteur de chaque *leporello* dépend du plus long paragraphe de tous les sismogrammes de l'auteur et sa longueur du nombre de sismogrammes contenus ainsi que du nombre de paragraphes par texte, puisqu'un sismogramme peut s'étendre sur deux pages ou, au contraire, une page peut contenir deux sismogrammes.

La seconde classification se fonde sur les spectres, classés dans un atlas, lequel est ainsi nommé en référence aux atlas des spectres stellaires. Un atlas désigne un recueil organisé de données visuelles (ici, les spectres) servant de référentiel pour l'analyse comparative. Cette classification s'appuie sur l'observation des caractéristiques spectrales de l'ensemble des textes. Deux critères ont été retenus pour leur pertinence : la proportion des paragraphes très courts de 1 ligne (rouge) et celle des paragraphes très longs de plus de 10 lignes (bleu).

3. Résultat : deux classifications des visualisations obtenues

Les deux classifications présentées ci-après seront analysées dans la partie « Discussion ».

3.1. La première classification : un leporello par auteur

Seize *leporellos*, en version numérique, ont été réalisés et, leur version papier exposée à la Maison des Sciences Humaines et Sociales (MSH) de Paris Nord par ordre chronologique de naissance des auteurs (de gauche à droite), soit selon cet ordre : Hugo, Sand, Dostoïevski, Tolstoï, Zola, James, Zweig, Woolf, Cartland, Higgins Clark, Dick, King, Follett, Djian, Lévy et Nothomb (*Figure 4*). Ensuite, afin d'appréhender la structure matérielle et la composition d'un *leporello*, un exemplaire est montré déplié partiellement (*Figure 5*). Enfin, cinq *leporellos* en version numérique sont présentés ouverts afin de pouvoir comparer les auteurs et d'en souligner les singularités (*Figure 6*).

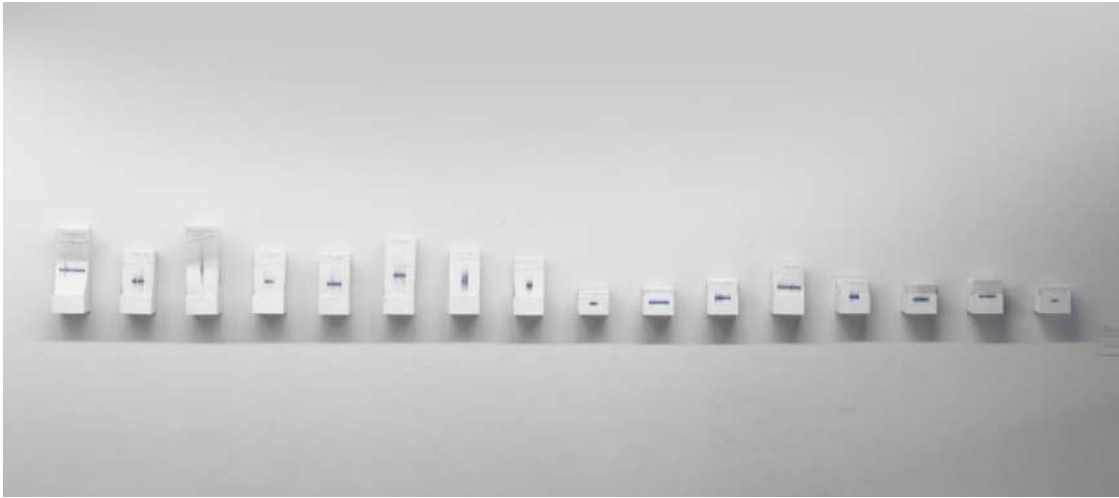


Figure 4. Vue partielle de l'exposition Narra, la face cachée des textes littéraires. 16 leporcellos placés par ordre chronologique de la date de naissance des auteurs. MSH Paris Nord, 2021.

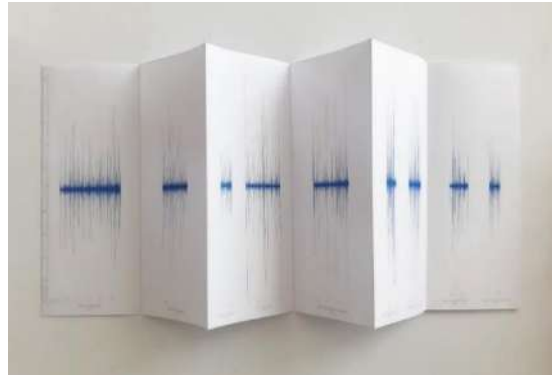


Figure 5. Leporello Henry James partiellement déplié. 168 x 27 cm.

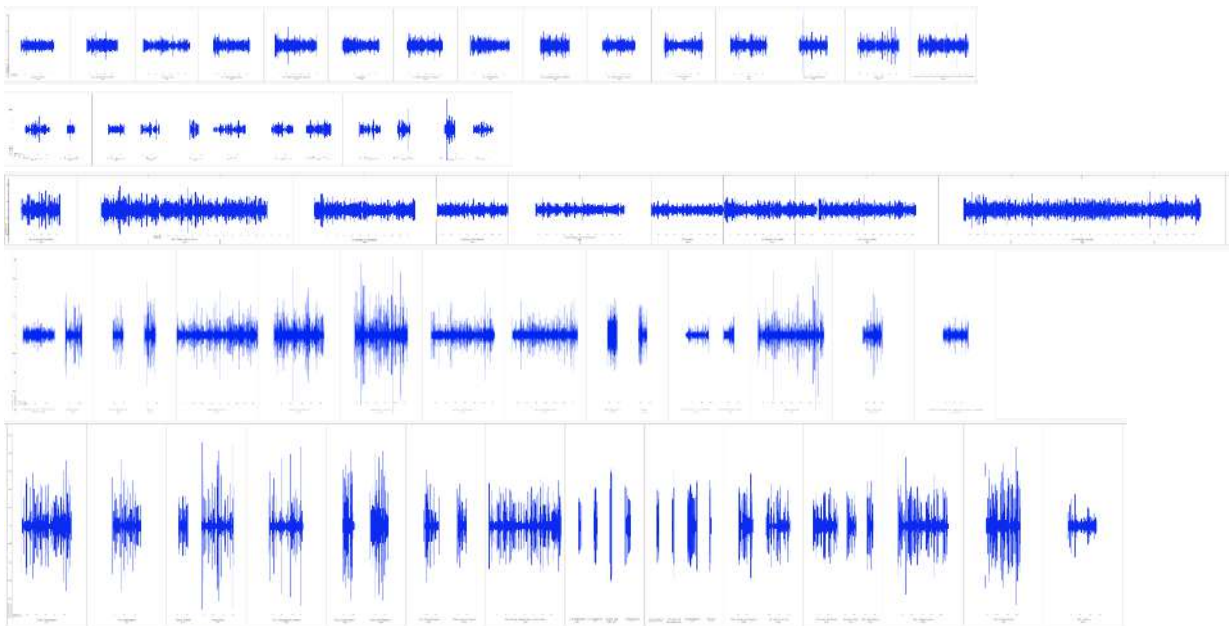


Figure 6. Version numérique intérieure de 5 leporcellos sur les 16 réalisés. De haut en bas : Philip. K. Dick, hauteur 9cm ; Amélie Nothomb, hauteur 9 cm ; Ken Follett, hauteur 10 cm ; Léon Tolstoï, hauteur 22 cm ; Henry James, hauteur 27 cm

3.1. La seconde classification : un atlas des spectres de textes littéraires

Un atlas de spectres de textes littéraires, comprenant 1 255 textes sous leur forme spectrale de 162 auteurs (104 pages), comprend sept classes : Classe I des spectres comprenant plus de 55% de paragraphes de 1 ligne, Classe II entre 55 et 45 %, Classe III entre 54 et 33 %, Classe IV entre 33 et 21 %, Classe V moins de 21 %, auxquelles s'ajoutent la Classe VI des spectres dont la distribution est équilibrée et Classe VII des spectres comptant très peu de paragraphes de longueur moyenne (4-7 lignes). À l'intérieur de chaque classe, les spectres sont ensuite classés selon le pourcentage des paragraphes de 1 ligne et selon la proportion de paragraphes très longs.

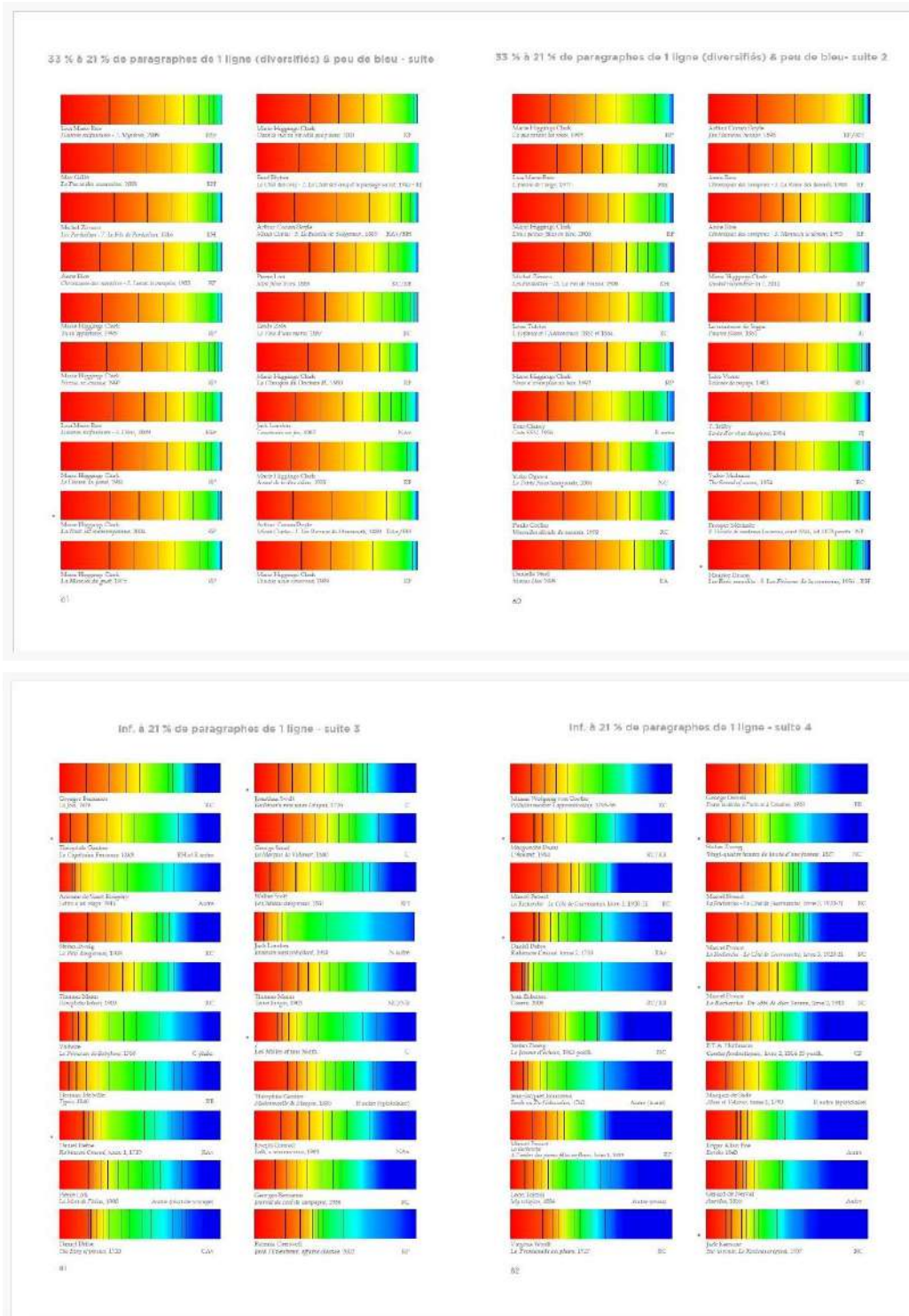


Figure 7.
Vue
numérique
intérieure de
l'Atlas, pages
63 et 64,
incluses dans
la Classe
IV. Spectres
comprenant
entre 33 % et
21 % de
paragraphes
de 1 ligne et
très peu de
longs
paragraphes

Figure 8.
Vue numérique
intérieure de
l'Atlas, pages
83 et 84,
incluses dans
la classe V.
Spectres
comprenant
moins de 21 %
de paragraphes
de 1 ligne

4. Discussion et conclusion

Si une interprétation qualitative des résultats demeure possible, la classification des visualisations substitue au discours interprétatif la force probante de l'image issue de données mesurables, lui conférant une double portée, heuristique et démonstrative.

4.1. Une portée heuristique : la première classification révèle la morphologie rythmique du texte et celle de la dynamique diachronique de l'auteur

La visualisation permet de condenser un phénomène diachronique comme un texte en une forme synoptique immédiatement saisissable. Les sismogrammes exposent un rythme du texte modulé en fonction de la fréquence des changements, de thème étant vu en introduction qu'un paragraphe narratif est une unité thématique et de locuteur pour un paragraphe dialogal. « Les changements sont la trame du temps [...] nous estimons souvent le temps par la quantité de changements qui s'y produisent » (Fraisie, 1961, p. 326). La façon dont les unités de sens se développent, donc s'enchaînent plus ou moins rapidement selon leur longueur ou leur durée, fait partie des rythmes du texte. Ce qui le caractérise est le « cours [des] idées [...] et la façon de les lier les unes aux autres » (Chartier, 2009-2010, p. 685). Roland Barthes relève que le langage n'est pas qu'un message, qu'il est aussi un flux, un rythme. « J'écoute l'emportement du message, non le message lui-même » (1970, p. 16) ; d'où *Le Plaisir du texte* (1973).

Par suite, le regroupement chronologique par auteur des sismogrammes dans une classification autorise une étude comparative intra-auctoriale, offrant un éclairage inédit sur le style de l'auteur qui complète les analyses textuelles traditionnelles. Ce dispositif synoptique permet d'évaluer leur évolution en mettant en lumière soit une permanence rythmique, soit au contraire des variations rythmiques significatives au fil de leur carrière.

L'examen du *leporello* de Tolstoï, par exemple, met d'emblée en évidence une hétérogénéité rythmique significative, infirmant l'hypothèse d'un style monolithique. La classification fait également apparaître un groupe d'œuvres (*L'Enfance et l'Adolescence*, *The Power of Darkness*, *Maître et Serviteur* et *Une Tourmente de neige et autres nouvelles*) caractérisé par des paragraphes assez brefs. Cette dynamique narrative est corrélée aux spécificités de genre : le registre théâtral pour *The Power of Darkness* et la forme brève (nouvelles) pour les trois autres. À l'inverse, l'essai *My Religion* se singularise par une prédominance de longs paragraphes, révélant un rythme posé propre au genre démonstratif. Concernant *Guerre et Paix*, la visualisation révèle une grande amplitude de ses unités thématiques, signe d'une dilatation du rythme narratif. À l'opposé, la proximité formelle des deux volumes d'*Anna Karénine* témoigne d'une cadence narrative plus vive qu'induit la fréquence plus élevée des changements. Enfin, la visualisation de *L'Enfance et l'Adolescence* révèle une rupture structurelle marquée où le rythme de *L'Enfance* est plus lent que celui de *L'Adolescence*. En conclusion, Tolstoï module son rythme en fonction des genres et, vraisemblablement, de la densité dialogale de ses textes. Son *leporello* donne à voir une variation scripturale remarquable, qui se traduit par une diversité rythmique désormais spatialisée et quantifiable.

En élargissant l'analyse au-delà de l'échelle intra-auctoriale, la classification permet une étude comparative inter-auctoriale qui met en lumière deux caractéristiques : une tendance baissière de la hauteur des *leporellos*, soit de la longueur des paragraphes au fil du temps, et des signatures rythmiques contrastées. En effet, ces visualisations révèlent d'un seul coup d'œil des écritures différentes : vive pour Dick et Nothomb, accentuée pour cette auteure par la brièveté de ses textes, déployée pour James et Tolstoï, dense pour Follett. Elles permettent de mesurer instantanément les écarts du rythme narratif, entendu ici comme la manière de développer les unités thématiques ou dialogales et qui résulte de la fréquence des changements.

En outre, si les *leporellos* de Dick, Nothomb et Follett mettent en évidence une homogénéité rythmique resserrée et régulière, ceux de James et surtout de Tolstoï, d'époques différentes, témoignent d'une écriture nettement plus variée avec une alternance de paragraphes courts et longs. Cette distinction souligne la capacité de la méthode à identifier des invariants ou des ruptures stylistiques, par-delà les écarts chronologiques des corpus considérés.

4.2. Une portée démonstrative : la seconde classification confirme l'hypothèse d'une corrélation entre la longueur des paragraphes et le genre littéraire ou la période

L'*Atlas des spectres de textes littéraires* a démontré l'existence d'une corrélation entre la longueur des paragraphes et le genre littéraire ou la période d'écriture (voir [analyse](#)). Il existe néanmoins des disparités (voir *Figure 9*) qui s'expliquent par l'hétérogénéité de certains genres littéraires étudiés comme les romans historiques, comprenant des romans de cap et d'épée, et les contes, comprenant des contes fantastiques, philosophiques et de jeunesse.

Pour autant, les regroupements observés confirment l'hypothèse de départ selon lequel la segmentation paragraphique est corrélée au genre littéraire. En effet, en l'absence de corrélation, la répartition des signatures spectrales des textes dans les classes de l'*Atlas* correspondrait à une distribution aléatoire. Ces résultats valident ainsi l'*Atlas* comme un instrument d'analyse fiable.

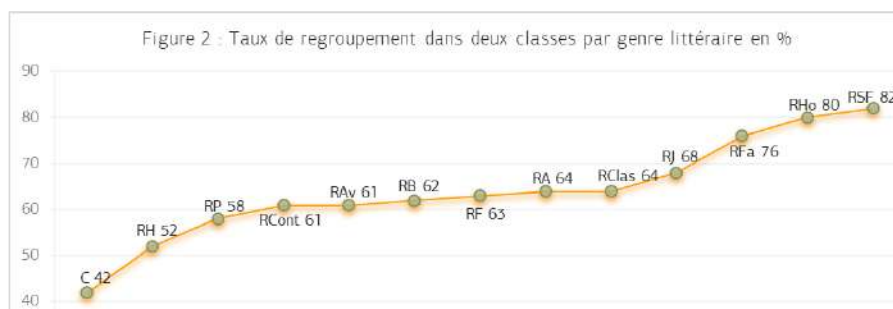


Figure 9. Taux de répartition dans deux classes par genre littéraire

Ces observations montrent que la visualisation et leur classification ne se limitent pas à une description formelle, mais constituent un instrument d'analyse capable de discriminer des régimes rythmiques à l'échelle intra- et inter-œuvres, intra- et inter-autoriales.

Sur le plan épistémologique, la présente classification numérique de visualisations de données est redevable du catalogue Henry Draper des spectres stellaires (1918-1924). Les historiens des sciences, Lorraine Daston et de Peter Galison, estiment que les images des atlas scientifiques sont « des images "au travail" » (2007/2012, p. 30) et que ces atlas « apprennent à voir l'essentiel et à négliger l'accessoire, à distinguer des objets types de ceux qui présentent des irrégularités, et à considérer l'étendue et les limites de la variabilité dans la nature » (*Ibid.* p. 33). À l'instar de cet atlas pionnier qui classait les étoiles selon leur signature spectrale pour comprendre l'univers, notre démarche transpose cette logique en classant les textes selon leur signature rythmique afin d'élargir notre compréhension des textes ainsi que de leur segmentation.

Dans une prochaine étape, pour une vingtaine d'auteurs, nous prévoyons de confronter trois types de classification de visualisations de données ou modalités d'analyse par auteur : séquentielle (*leporellos* suivant l'ordre d'apparition des paragraphes), fréquentielle (spectres de

distribution selon la longueur des paragraphes) et distributionnelle (les diagrammes en boîte pour identifier les *outliers*, soit les paragraphes distinctivement longs dans le texte, et les diagrammes en violon pour saisir la densité des paragraphes selon leur longueur).

Par la suite, l'articulation entre l'analyse statistique et le close reading visera notamment à identifier des patterns rythmiques propres à chaque auteur et à repérer visuellement, en un regard, sur les sismogrammes, les moments clés du récit.

Bibliographie

- Achard-Bayle, G., & Pešek, O. (2020). Modèles d'organisation thématique des paragraphes et entre les paragraphes, à l'épreuve de la Rectorique de Cicéron. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (26). <https://doi.org/10.4000/discours.10794>
- Algee-Hewitt, M., Heuser, R., & Moretti, F. (2015). On Paragraphs. Scale, Themes, and Narrative Form – Stanford Literary Lab. *Stanford Literary Laboratory*. <https://litlab.stanford.edu/LiteraryLabPamphlet10.pdf>
- Barthes, R. (1970). *Sade, Fourier, Loyola* (Autre tirage: 1980, 2003). Éditions du Seuil.
- Barthes, R. (1973). *Le plaisir du texte*. Éditions Points.
- Benveniste, É. (1966). *Problèmes de linguistique générale, I*. Gallimard.
- Castille, J.-F. (2012). La poétique du paragraphe dans l'écriture réaliste de Flaubert et Maupassant. *Questions de style, Réalisme(s) et Réalité(s)*, (9), 147-159. <https://hal.science/hal-00779676>
- Chartier. (2010). *Résumés annuels*. <https://www.college-de-france.fr/site/roger-chartier/resumes.htm>
- Dahan-Gaida, L. (2023). *L'art du diagramme : Sciences, littérature, arts*. Presses universitaires de Vincennes.
- Daston, L., & Galison, P. L. (avec Latour, B.). (2012). *Objectivité* (S. Renaut & H. Quiniou, Trad.; Vol. 1-1). les Presses du réel.
- Fraisse, P. (1961). *Influence de la durée et de la fréquence des changements sur l'estimation du temps*. <https://doi.org/10.3406/psy.1961.26817>
- Keim, D. A., & Oelke, D. (2007). Literature Fingerprinting: A New Method for Visual Literary Analysis. *2007 IEEE Symposium on Visual Analytics Science and Technology*, 115-122. 2007 IEEE Symposium on Visual Analytics Science and Technology. <https://doi.org/10.1109/VAST.2007.4389004>
- Michon, P. (2018). Sommes-nous en train d'assister à l'émergence d'un nouveau paradigme scientifique : Le paradigme rythmique ? In *Rythmanalyse(s) : Théories et pratiques du rythme ontologie, définitions, variations* (p. 25-42). Jacques André éditeur.
- Moretti, F. (2016). *La littérature au laboratoire* (V. Lëys, Trad.). Ithaque.
- Rey, A. (avec Le Robert). (2001). *Dictionnaire historique de la langue française : Contenant les mots français en usage et quelques autres délaissés, avec leur origine proche et lointaine* (Nouvelle éd. augmentée par Alain Rey). Dictionnaires Le Robert.
- Sauvanet, P. (1992). « Le » rythme : Encore une définition ! (p. 233-243). Centre culturel international de Cerisy Éd. l'Harmattan.
- Stark, H. A. (1988). What do paragraph markings do? *Discourse Processes*, 11(3), 275-303. <https://doi.org/10.1080/01638538809544704>
- Wulf, J. (2010). Rythmes sémantiques (Richelieu et la fronde). In P. Petitier (Dir.), *Michelet, rythme de la prose, rythme de l'histoire* (p. 23-33). Presses universitaires du Septentrion.

De l'analyse de données en cartes : Visualisations géographiques et chronologiques de corpus textuels

Marie Pérès-Leblanc¹, Anne-Lise Humain-Lamoure², Jean-Marc Leblanc³

¹ Univ Paris Est Créteil, LAB'URBA, F-94010 Créteil, France – marie.leblanc@u-pec.fr

² Univ Paris Est Créteil, LAB'URBA, F-94010 Créteil, France – al.humain@u-pec.fr

³ Univ Paris Est Créteil, CEDITEC, F-94010 Créteil, France - jean-marc.leblanc@u-pec.fr

Abstract

The analysis of a large corpus drawn from national and regional press in France, spanning from the late 1990s to 2024, has enabled us to identify temporal evolutions in the use of clothing-related terms associated with Islam, as well as the media outlets, authors, and territories most frequently mobilized. It has also allowed us to examine how the perceived undesirability of certain populations is shaped by stereotyped representations, and even by moral panics partly constructed within and by the press media. This research, situated at the intersection of disciplines, has raised methodological issues, particularly related to the limitations of visualization tools in textual analysis when it comes to combining textual variables with geographical and chronological ones. Based on an experiment using a tool commonly employed in geography (Magrit, <https://magrit.cnrs.fr/>), this article presents proposed developments for the TexObserver tool aimed at improving the representation of such data

Keywords: Textual data, vizualizations, maps, chronologies.

Résumé

L'analyse d'un vaste corpus issu de la presse nationale et régionale en France de la fin des années 1990 à 2024 nous a permis de cerner les évolutions au cours du temps de l'usage de termes vestimentaires associés à l'Islam et des médias, auteurs, territoires les plus mobilisés et de cerner comment l'indésirabilité de certaines populations se nourrit de représentations stéréotypées voire de paniques morales construites en partie dans et par les médias de presse. Cette recherche qui s'inscrit au carrefour des disciplines a ouvert des problématiques méthodologiques notamment liées aux limites des propositions de visualisation des outils d'analyse textuelle lorsqu'il s'agit de croiser des variables textuelles avec des variables géographiques. A partir d'une expérimentation sur un outil couramment utilisé en géographie (Magrit <https://magrit.cnrs.fr/>) notre article présente les perspectives de développement de l'outil TexObserver proposées afin de mieux représenter ces données.

Mots clés : données textuelles, visualisations, cartes, chronologies

1. Introduction

Cet article repose sur l'analyse d'un vaste corpus de presse menée dans le cadre du projet de recherche « La fabrique urbaine de l'indésirabilité »¹ (2023-2025). Ce projet, coordonné par Claire Hancock (Lab'Urba) et Muriel Froment-Meurice (LAVUE) en réponse à un appel de la défenseure des droits, vise à documenter dans toute leur diversité les modes de gestion de l'espace public conduisant à l'éviction de populations désignées comme « indésirables », en

¹ <https://laburba.univ-gustave-eiffel.fr/recherches/projets-de-recherche/la-fabrique-urbaine-de-lindiserabilite>

confrontant l'appareil normatif et les mises en application effectives, et ce à différentes échelles et sur des terrains diversifiés dans toute la France.

Nous avons constitué un corpus à partir de la base Factiva dans le but de repérer de manière inductive des représentations stéréotypées, voire de paniques morales² construites en partie dans et par les médias de presse sur la période allant de 1998 à 2024.

Nous exposerons ici notre méthodologie de recueil et d'analyse ainsi que nos principaux résultats. A la suite de cette présentation nous expliciterons les écueils rencontrés liés aux outils textométriques et leurs limites, les méthodes de création de carte avec un outil open source, gratuit, en ligne et interopérable (Magrit³) et proposerons un développement d'outil pour le logiciel TextObserver permettant d'ajouter une visualisation géographique dynamique des résultats d'analyse.

2. Méthodologie de constitution et d'analyse du corpus

Notre corpus d'articles de presse a été constitué à partir de la base Factiva accessible depuis les portails universitaires. Cet agrégateur de contenu créé dans un objectif de veille informationnelle économique internationale regroupe près de 30 000 sources compilées provenant de 200 pays et en 32 langues. Nous avons retenu la presse généraliste nationale et régionale en France et en langue française (ce qui inclut la plupart des grands médias, magazines, suppléments et sites web inclus quelle que soit leur périodicité ou leur mode de publication). L'objectif était de retenir des textes auxquels une grande partie de la population a accès. On note que le Monde et le Dauphiné libéré ne figurent pas dans la base. Nous avons écarté les blogs et les dépêches des agences de presse et retenu dans les résultats les catégories "informations et politiques générales", "à la une/tendances/analyse" écartant les rubriques spécialisées (comme "sport") dans la mesure où il s'agit d'un sujet que l'ensemble de la population ne suit pas forcément.

Il est à noter que la base Factiva comprend peu d'articles de la période antérieure au milieu des années 1990 et que certaines sources cessent d'apparaître ou apparaissent en cours de la période d'analyse. De la même manière le nombre de publications et leurs canaux de diffusion s'accroissent (un même sujet est traité en brève sur un site web de quotidien et sur la version papier). Tous ces éléments doivent être pris en compte dans l'analyse diachronique des résultats. En dépit des nombreux filtres activés un nettoyage important a été nécessaire au niveau des sources et des articles.

La principale difficulté dans la constitution de ce corpus a été la construction de la requête autour de l'idée très large de l'indésirabilité de certaines populations sans préjuger des discours, mais à l'aide de mots clé. Le corpus utilisé ici se fonde sur des termes couramment utilisés pour désigner des vêtements féminins de l'Islam : « voile », « foulard (islamique) », « hijab », « niqab », « abaya », « burqa », « burkini ». Cette entrée par les vêtements féminins se révèle intéressante à plusieurs titres. Les tests effectués montrent qu'en dépit d'une multiplicité des termes employés les occurrences s'étendent sur plus de 30 ans⁴. Des recherches montrent

² Ce concept forgé par le sociologue Stanley Cohen (« moral panic »), désigne une réaction disproportionnée de l'opinion publique vis-à-vis d'un comportement perçu comme une menace et attribué à un groupe donné – perception amplifiée par les médias dans l'exemple analysé. Cohen, S. (1972), *Folk devils and moral panics*, London: Mac Gibbon and Kee.

³ <https://magrit.cnrs.fr/>

⁴ Ce n'est pas le cas pour d'autres expression de type « prière de rue » ou de mots clés comme « barbu ».

comment le vêtement féminin est devenu ces dernières décennies un signe distinctif dans l'espace public de l'islam et peut donc concourir à des catégorisations (Abouddrar, 2017). Immédiatement visibles, ils peuvent être un signe d'affirmation de la part des populations qui les portent autant qu'un signe discriminant les exposant à des attitudes ou des comportements hostiles (Aranguren & Madrisotti, 2019). Enfin, le voile, le burkini, plus récemment l'abaya ont été l'objet de vifs débats en France dès les années 1980 (Pelletier, 2005) et visés par différents dispositifs d'interdiction tout autant controversés (loi contre le voile à l'école en 2004, arrêtés municipaux contre le port du burkini en 2016, circulaire d'interdiction de l'abaya à l'école en 2023). Le port de vêtements assimilés à des signes religieux fait aussi l'objet de polémiques et de bataille juridique (crèche Baby Loup en 2008). Le choix de ces mots-clés permet donc d'appréhender la fabrique de l'indésirabilité dans l'espace public (physique et métaphorique) dans sa complexité.

Le corpus finalement constitué comporte près de 14 000 articles issus de la presse nationale et régionale en France de 1997 à 2024, pour 10 200 000 occurrences et nous a permis d'analyser comment se constitue une catégorisation dans le temps long, par les modes de désignation. C'est cet aspect de notre recherche que nous présentons ici.

3. Présentation d'une partie des résultats de l'analyse

Nous présentons ici les grands traits de notre analyse afin de démontrer les enjeux en termes de visualisation⁵. Nous avons soumis le corpus au logiciel Iramuteq (Marchand, Ratinaud, 2015) en mobilisant la méthodologie Alceste (Reinert, 1993). Dans la configuration par défaut il a donc été lemmatisé. Le balisage comprend l'année de parution, le titre de la publication (sont confondus en un seul titre la publication papier, web et suppléments le cas échéant), le ou les auteurs et la diffusion territoriale. Cette dernière variable a dû être construite hors corpus à partir d'extractions rapportant chaque publication à une territorialisation au regard de la diffusion affichée par les médias et les groupes de presse dont ils dépendent (Bousquet, Amiel, 2021). Elle se décline en plusieurs modalités : France métropolitaine, régions et département. Les territoires ultramarins n'ont pas été retenus car nous disposions de trop peu de sources les concernant.

L'analyse a fait émerger des caractéristiques lexicales en diachronie mais aussi une récurrence de certains médias, locuteurs et territoires.

Un premier résultat d'analyse basé sur les fréquences relatives des termes vestimentaires pris dans leur ensemble a permis de mettre en évidence une périodicité des usages sous formes de pics qui traduisent un effet d'emballlement du discours médiatique. Ces emballlements correspondent aux moments de vifs débats et controverses (par exemple le voile à l'école en 2004, le débat autour de la crèche Baby Loup en 2008, ou encore celui autour des prières de rue en 2010). La périodicité régulière du thème de l'islam dans les espaces publics permet de supposer une mobilisation régulière du terme pour provoquer des polémiques en périodes préélectorales présidentielles (2012, 2016-2017, 2019) et des formes de paniques morales électoralistes.

Nous avons également identifié une périodicité d'usage selon les vêtements. Cette succession des termes peut être une manière de renouveler la polémique et faire grandir les paniques morales notamment en multipliant les termes ces dernières années en arabe, voire de lier ces

⁵ Une analyse plus fouillée, centrée sur les résultats est en cours de soumission.

Nos observations sur la répartition des médias éditeurs montre des différences très marquées. A cette répartition des titres s'ajoute celle des auteurs. Si 85% des signataires ne produisent qu'un à trois articles sur la période⁶ en revanche 2% des auteurs (soit une cinquantaine) représentent 20% du total des articles et ne participent généralement qu'à un seul média (parfois deux). Cet aspect nous intéresse particulièrement en raison de la corrélation entre média nationaux ou régionaux et type d'énoncés.

[illegible]

⁶ Un même article (et donc signataire) peut parfois être repris à l'identique pour un supplément ou le site web d'un média papier.

La deuxième classe qui regroupe 29% des énoncés relève des chroniques violentes et est construite à partir des déictiques spatio-temporels (jour, horaires, mois, rue, type d'habitation) et des termes renvoyant à la famille, à la mort et au terrorisme. Ce contenu est plutôt constant sur la période même si les années 2012 et 2015 sont fortement et logiquement associées. Là encore ce sont les médias régionaux de grande audience avec quelques journaux locaux de grandes villes qui y sont plus particulièrement associés.

La troisième classe qui représente 13% des énoncés couvre le champ lexical de la politique et des campagnes électorales. On y retrouve toutes les échelles électorales et l'ensemble des présidents de la République ainsi qu'une bonne partie des premiers ministres. Contrairement aux précédents ces énoncés sont le fait des médias quotidiens et hebdomadaires nationaux sans effet de surreprésentation de région (si ce n'est PACA). Il est à noter qu'on ne trouve ces contenus qu'à partir de 2016 ce qui permet de souligner que relier ainsi des signes distinctifs à des énoncés de politique générale est une tendance nationale très récente.

Les deux dernières classes se regroupent au centre du plan factoriel (Figure 1, zone 4) et représentent chacune 12% des énoncés. Elles ont respectivement pour thématique les analyses internationales et les débats de société. Les espaces cités sont à l'échelle des Etats ou d'organisation supranationales. Si ce contenu se déploie sur toute la période les analyses internationales ont des thématiques liées à la chronologie alors que les débats de sociétés sont surtout visibles à partir de 2019. Ces deux contenus sont quasiment exclusivement le fait de médias nationaux.

On note ainsi qu'une grande part des énoncés sont le fait de la presse régionale et locale avec de fortes variations géographiques (**Erreur ! Source du renvoi introuvable.**). De plus, certaines classes de discours sont significativement associées à minima à des régions, notamment au sud de la France, dans le Grand Est et en Ile de France avec deux types de contenus. D'une part, les débats autour de l'interdiction de signes et donc de vêtements musulmans dans les espaces publics, d'autre part des récits et chroniques qui associent fortement un vêtement, la djellaba⁷ (et le fait d'être barbu) à des actes de violences en particulier du terrorisme. Ces contenus s'étendent de la fin des années 1990 jusqu'en 2024 et expliquent en grande partie les pics réguliers d'usage de ces termes, de même que la répartition géographique observée qu'il faut néanmoins nuancer à plus fine échelle. Si la carte ci-dessous peut être reconstituée ex-post à partir d'une extraction de la base des balises, il est en revanche plus complexe de cartographier les classes de discours observés.

Si les balises de date permettent de façon très classique en textométrie d'aborder facilement les notions chronologiques en termes de partitions elles ne permettent pas de croiser aisément ces données avec des informations spatiales. Les visualisations géographiques au sein de ces outils n'existent pas et supposent des manipulations très artisanales qui consistent en l'extraction des données sous forme de tableau des balises, à retravailler pour créer une variable spatiale lisible pour un logiciel tiers de cartographie. Ceci rend l'analyse particulièrement fastidieuse en mode exploratoire et nécessite une extraction à chaque modification du corpus. Par ailleurs, nous aurions souhaité pour cette analyse géographique croiser les résultats d'analyse des données textuelles avec des données géolocalisées classiques (démographiques, sociologiques, électorales par exemple) à différentes échelles (ici les régions et les départements). Or la spatialisation, le croisement entre analyse de discours et données tierces permet de mettre en

⁷ Le terme ne faisait pas partie de la requête, il a néanmoins émergé de manière constante tout au long de la période d'étude. Il a donc été à ce titre pris en compte dans l'analyse.

lumière des dynamiques territoriales différenciées dans la fabrique de l'indésirabilité, montrant ainsi des effets de contextes. Nous supposons dans cet exemple de recherche que des types de discours sont associés à des espaces principalement ruraux de faible densité en marge des grandes aires urbaines et où le vote en faveur de l'extrême droite et/ou l'abstention progressent plus rapidement qu'ailleurs. Les outils actuels ne nous ont pas permis dans le temps imparti du programme de tester ces hypothèses tant l'analyse est entravée par des manipulations peu satisfaisantes.

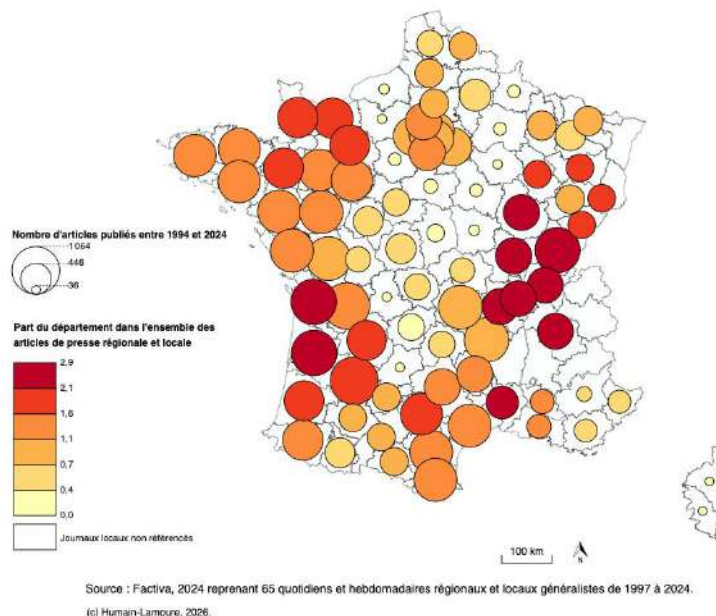


Figure 2 : Carte des occurrences par article et par département dans la presse régionale et locale entre 1997 et 2024

4. Expérimentations de visualisation avec Magrit

La carte proposée à la Figure 2 a été produite grâce aux données extraites de notre corpus afin de palier un manque dans les visualisations de résultats proposées par les outils textométriques. Il existe différents logiciels de cartographie thématique simples permettant la visualisation géographique de données. Parmi ceux-ci, notre choix s'est porté sur Magrit⁸ car il est en open source, gratuit, en ligne et interopérable. Produit par l'unité d'appui à la recherche RIATE (CNRS, Université Paris Cité), ce logiciel est en constant développement via la plateforme GitHub.

Nous explicitons ici son fonctionnement afin d'éclairer comment on pourrait utiliser plus directement les données issues des analyses textométriques afin de les visualiser sous forme de carte.

Comme l'ensemble des logiciels de ce type, il suppose de relier grâce à un identifiant commun une base de données (en format .xlsx, .csv, .txt, etc. voir Figure 3) et un fond cartographique où chaque unité est identifiée dans une table attributaire (en format .shape ou .geojson souvent produit à partir d'Illustrator, Figure 3). La jointure ici est réalisée à partir de l'identifiant « code » départemental. La table attributaire du fond cartographique est enrichie des données

⁸ Pour rappel : <https://magrit.cnrs.fr/>

du fichier initial (Figure 4). Il est alors possible de réaliser une représentation cartographique des données. Le choix de la fonctionnalité dépend du type de donnée dans la sémiologie graphique classique telle que fixée dès les années 1960 (Bertin, 1967) et de l'objectif de la carte. Le choix se porte ici sur les symboles proportionnels adaptés à une variable quantitative de stock – ceux-ci sont déclinés selon une variable couleur en fonction de la part que chacun représente dans l'ensemble (variable quantitative de ratio). L'application permet de choisir le type de symbole, la taille maximale et pour la variable de ratio, la discrétisation (la mise en classe) la plus adaptée au regard des paramètres de la distribution.

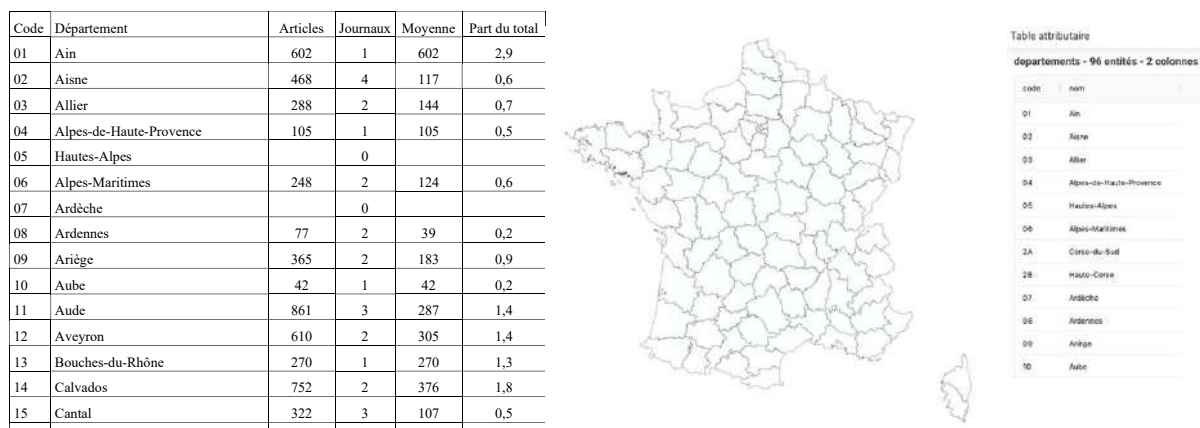


Figure 3 : Fond cartographique géoréférencé et fichier excel contenant le tableau

La carte est créée dans une nouvelle couche qui a sa propre table attributaire et sur laquelle de nouvelles opérations sont possibles. D'autres couches peuvent être ajoutées soit pour l'habillage de la carte, soit pour mettre en relation visuelle d'autres données.

Joindre à une couche géospatiale

Cette fonctionnalité permet de joindre les données d'un tableau à une couche géospatiale. Les données du tableau sont ajoutées à la couche géospatiale en fonction d'un champ commun aux deux jeux de données. Seuls les champs contenant des valeurs uniques peuvent être utilisés pour réaliser la jonction.

Sélectionner la couche cible

departements

Paramètres de normalisation

Invisible à la casse (ne pas tenir compte de la casse)

Normaliser le texte (ignorer les signes diacritiques, les espaces et les signes de ponctuation)

Champs de jonction

Charger departements

Table Carte

code

1 01

2 02

3 03

4 04

5 05

Informations sur le résultat de la jonction

Géométries ayant une correspondance : 100%

Entrées du tableau ayant une correspondance : 100%

Utiliser un préfixe pour les champs joints ?

Sélectionner les champs à joindre ?

Confirmer Annuler

code	nom	Région	Articles	Journaux	Moyenne	Part du total
01	Ain	Auvergne-Rhône-Alpes	602	1	602	2,8518022198400876
02	Aisne	Hauts-de-France	468	4	117	0,6010542968078677
03	Allier	Auvergne-Rhône-Alpes	288	2	144	0,6831413018886807
04	Alpes-de-Haute-Provence	P Provence-Alpes-Côte d'Azur	105	1	105	0,4389240091063420
05	Hautes-Alpes	P Provence-Alpes-Côte d'Azur		0		
06	Alpes-Maritimes	P Provence-Alpes-Côte d'Azur	248	2	124	0,5882633003415122
2A	Corse-du-Sud	Corse	38	1	38	0,17078805042574878
2B	Haute-Corse	Corse	36	1	36	0,17078805042574878
07	Ardèche	Auvergne-Rhône-Alpes		0		
08	Ardennes	Grand Est	77	2	38,5	0,1628462395898296
09	Ariège	Occitanie	365	2	182,5	0,86179805813789107
10	Aube	Grand Est	42	1	42	0,19321043882137122
11	Aude	Occitanie	861	3	287	1,3616444953205724
12	Aveyron	Occitanie	610	2	305	1,4489377816715367
13	Bouches-du-Rhône	P Provence-Alpes-Côte d'Azur	270	1	270	1,03086678183101
14	Calvados	Normandie	752	2	376	1,7837884332807596
15	Cantal	Auvergne-Rhône-Alpes	322	3	107,33333333333333	0,6081846598878807
16	Charente	Nouvelle Aquitaine	1012	8	253	1,201368847078886
17	Charente-Maritime	Nouvelle Aquitaine	369	2	184,5	2,3804901132082252

Figure 4 : jonction à une couche géospatiale et table attributaire

La dernière étape consiste en la mise en page – l'habillage de la carte – comportant le titre, une légende plus explicite, la source, l'échelle, l'orientation, les crédits (Figure 6, le menu déroulant sur la droite permet ces ajouts). L'export de la carte réalisée en format .svg ou .png se fait grâce au dernier menu. Le projet peut être enregistré au format propriétaire .mjson (en haut à droite).



Figure 5: Options de symboles et paramètres de discrétisation

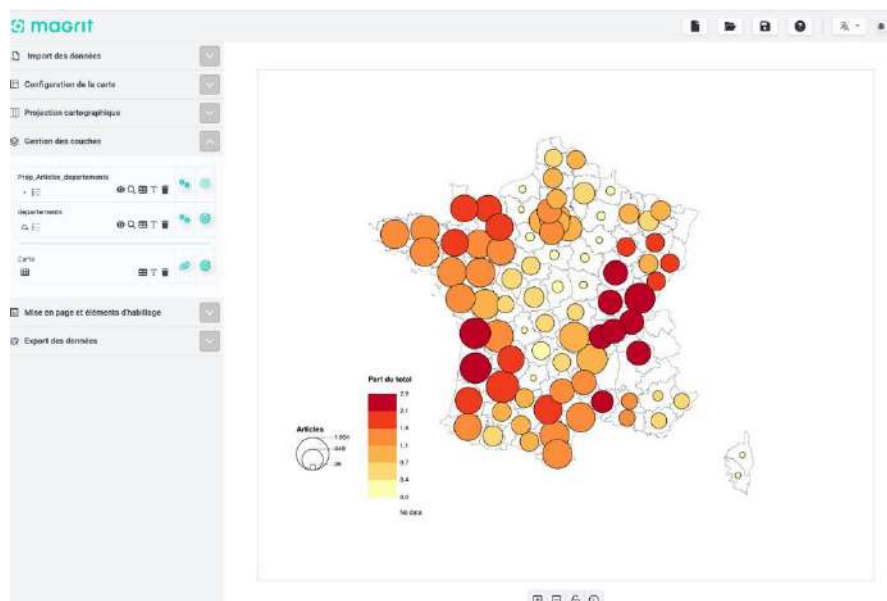


Figure 6 : carte version exploratoire

Cet outil est simple à prendre en main, ergonomique (les menus sur la gauche nous guident à chaque étape). Cela fait partie des éléments de choix pour une intégration à notre maquette. En constant développement, cette application commence à intégrer des outils d'analyse spatiale de plus en plus élaborés le rapprochant ainsi d'un système d'information géographique (SIG).

5. Propositions de développement de l'outil Textobserver

Notre propos ici est de proposer une maquette de développement pour un module de visualisation des données à caractère géographique au sein du logiciel TextObserver. L'exemple décrit ci-après permettra d'éclairer cette nouvelle fonctionnalité. Une maquette dynamique illustrant les manipulations possibles est disponible en ligne⁹.

Si la couche de fabrication de la carte reprend les étapes proposées par le logiciel Magrit nous souhaitons ajouter des éléments au traitement des données qui permettra de récupérer les informations à caractère géographique directement dans le corpus à la fois en s'appuyant sur le balisage mais aussi sur un dictionnaire qui permettrait d'identifier les lieux dans le texte. L'outil

⁹ Voir <https://textobserver.textopol.org/JADT2026/>

Unitex (Kyriacopoulou et al. 2014) par exemple pourrait être utilisé pour ce faire. L'introduction d'une balise *geo* peut par exemple - au-delà de sa fonction première de partitionnement - permettre directement dans le corpus de cibler une aire géographique (pays, région, département, ville, quartier). Nous souhaitons également proposer lorsque le balisage le permet d'ajouter une variable chronologique à cette visualisation (en permettant de visualiser de manière dynamique une période choisie par l'utilisateur).

La possibilité d'utiliser des fonds de cartes par défaut nous semble intéressante pour s'adresser à une communauté de non-géographes souhaitant un résultat rapide (proposition d'un fond lié à l'aire géographique déduite du corpus : un balisage par région affichera un fond de carte région). Magrit a l'avantage de proposer de nombreux fonds librement utilisables (l'ensemble des mailles administratives en France et en Europe sont ainsi disponibles). De nombreux fonds géolocalisés sont disponibles en ligne grâce notamment à l'opendata des collectivités. Nous souhaitons également proposer à l'utilisateur d'intégrer ses propres fonds si le fond nécessaire n'est pas disponible ou que l'utilisateur souhaite un changement d'échelle personnalisé. Cette proposition pourra-être accompagnée de la possibilité de faire appel à la communauté afin d'enrichir ce fond en ligne au moyen de liens vers les bases de données disponibles dans les réseaux de Géographes et référencées sur Textopol.

Si le corpus soumis à TextObserver ne possède pas de balise *geo* on devra ajouter une entrée d'équivalence par fichier csv. Ce fichier permettra ici par exemple de qualifier les journaux selon leur diffusion géographique et donc de distribuer par exemple les articles citant des vêtements de l'Islam dans la presse régionale et locale sur la carte de France à l'échelle régionale. Si le fichier comprend à la fois des références de niveau régional et départemental la carte produite contiendra deux couches et le basculement de l'une à l'autre permettra une visualisation dynamique.

La visualisation peut se faire à partir d'une table lexicale unique mais aussi de plusieurs tables issues du même corpus. Dans notre corpus on peut ainsi par exemple ouvrir simultanément la table par publication (Journal) et celle par locuteurs (Aut) afin de combiner ces deux couches d'information lors de la création de la carte. L'ajout de la table Date aux deux précédentes permettra au chercheur de faire varier la période considérée et de visualiser le résultat de manière dynamique dans le module.

De la même manière l'exemple évoqué ci-avant nous a permis de mettre en lumière l'importance de rapprocher des données issues de l'analyse du corpus d'autres données « externes ». On pourra ainsi ajouter sur notre visualisation une couche permettant de relier le nombre d'abonnements de la PQR par département ou encore les résultats électoraux des régions/département selon l'élection ciblée. La combinaison de ces différentes informations issues du corpus mais aussi externes permet alors de configurer une cartographie permettant de rapprocher des informations difficilement saisissables directement lors de l'analyse outillée du corpus.

D'autre part nous souhaitons ici reprendre un aspect fondateur du logiciel TextObserver qui est celui de l'implémentation du mouvement permettant de mieux se saisir des différentes variations en permettant une expérimentation dynamique dans le module de visualisation cartographique. Le schéma en ligne explicite l'ensemble des fonctionnalités envisagées (<https://textobserver.textopol.org/JADT2026/>).

Un dernier point nous semble ici essentiel. Il s'agit de proposer à l'utilisateur une navigation fluide entre tous les éléments du dispositif rendant l'ensemble du processus dynamique et

modifiable : de la carte au corpus, du corpus à la carte mais aussi au module d'annotation¹⁰ et aux résultats textométriques.

7. Conclusion

L'analyse du corpus, dont on a présenté ici les principaux résultats, a ouvert des problématiques méthodologiques en montrant les limites des méthodes textométriques lorsqu'il s'agit de prendre en considération des données ou des variables spatiales. Cette contribution propose ainsi un prolongement vers de l'analyse des données spatiales et géographiques, et donc un croisement de disciplines autour d'objets textuels par ailleurs fréquemment explorés par la textométrie.

La rencontre entre linguiste, spécialiste de la visualisation et géographe s'avère ainsi particulièrement riche et ouvre des perspectives, en permettant de confronter des cadres d'analyse, des échelles d'observation et des modes d'interprétation complémentaires. La statistique, mobilisée à la fois par les géographes et les textométriciens, illustre cette convergence. Si les usages diffèrent selon les traditions disciplinaires, elle constitue un socle commun favorisant le dialogue et l'enrichissement mutuel des approches.

Au-delà du simple aspect technique et des fonctionnalités logicielles, la confrontation des méthodes et les regards croisés portés sur un même objet éclairent notre compréhension du corpus.

Bibliographie

- Abouddrar B.N. (2017), *Comment le voile est devenu musulman*, Flammarion, coll. Champs-Essais.
- Aranguren, M. & Madrisotti, F. (2019), « Des regards aux égards : Effets du foulard islamique sur le comportement visuel des passagers du métro de Paris », *Hommes & Migrations*, 1324, pp.39-47. <https://doi.org/10.4000/hommesmigrations.8276>
- Barats, C. (2019). Médiatisation et usages socio-discursifs du « classement de Shanghai », *Revue française d'administration publique*, Vol. 2019/1 N° 169, pp.101-120.
- Becker H., (1985, 1963), *Outsiders. Études de sociologie de la déviance*, Éditions A.-M. Métailié.
- Bertin J. (1967), *Sémiologie graphique*, Paris, Mouton et Gauthier-Villars.
- Bousquet F. et Amiel P. (2021), *La presse quotidienne régionale*, Paris, La Découverte, Coll. Repères.
- Kyriacopoulou T., Martineau C., Martinez C. and Schoen A. (2014), Traitement automatique des langues : extraction d'information dans les textes lié au logiciel Unitex. In *Convention Accès Brevets Innovation dédiée au BigData*, Jun 2014, Paris, France.
- Marchand and P. Ratinaud, P. (2015), Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014), *Mots, Les Langages Du Politique*, 108, pp.57-77, <https://www.cairn.info/revue-mots-2015-2-page-57.htm>.
- Pelletier, D. (2005), L'école, l'Europe, les corps : la laïcité et le voile, *Vingtième Siècle. Revue d'histoire*, n°sup 87, pp.159-176. <https://doi.org/10.3917/ving.087.0159>.
- Ratinaud, P. (2016), Étude d'un large corpus textuel avec IRaMuTeQ : Twitter et le hashtag #mariagepourtous, In Lo Monaco G., S. Delouvé, & P. Rateau (Eds.), *Les représentations sociales. Théories, méthodes et applications*, De Boeck, pp.107-112.
- Reinert, M. (1993), Les "mondes lexicaux" et leur logique, *Langage et société*, 66, pp.5-39.

¹⁰ Ce module de TextObserver permet pour chaque mot d'ajouter des annotations qui peuvent être utilisées par exemple pour désambiguïser un usage polysémique et en permettant notamment d'effectuer une concordance sur le mot ou sur l'annotation.

Étude des cooccurrences du verbe *Dire* dans un corpus d’entretiens à propos d’un événement traumatique

Anna Colli¹, Delphine Battistelli², Marie Chagnoux³

¹Laboratoire Modyco, CNRS, Université Paris Nanterre - Laboratoire CEMTI, Université Paris 8 Saint-Denis — acolli@parisnanterre.fr

²Laboratoire Modyco, CNRS, Université Paris Nanterre — dbattist@parisnanterre.fr

³Laboratoire CEMTI, Université Paris 8 Saint Denis — marie.chagnoux@univ-paris8.fr

Abstract (in English)

In this article, we analyze the various uses of the verb *dire* (*to say*) within a corpus on the November 13, 2015 attacks in Paris and Saint-Denis. Using co-occurrence analysis, we identify the uses of *dire* that are favored depending on the speakers’ distance from the event. In particular, we observe an overall tendency among speakers to “stage” conversations through reported exchanges and inner speech

Keywords : corpus linguistics, textometry, cooccurrences, semantics, verb *to say*, terrorist attacks

Abstract (in French)

Dans cet article nous analysons les différents emplois du verbe *dire* à l’intérieur d’un corpus sur les attentats du 13 novembre 2015 à Paris et à Saint Denis. Nous faisons émerger, à l’aide de l’analyse des cooccurrences, les emplois de *dire* qui sont privilégiés selon la position du locuteur par rapport aux lieux de l’événement traumatique. Nous remarquons une tendance globale des locuteurs à la “mise en scène” des conversations à l’aide du dispositif des échanges rapportés et du *dire* intérieur.

Mots clés : linguistique de corpus, textométrie, cooccurrences, sémantique, verbe *dire*, attentats

1. Introduction

Divers travaux de recherche (par exemple, (Charretier et al., 2024) et (Lacoste et al., 2024)) s’intéressent à la manière dont les récits sur un événement traumatique peuvent différer en fonction de la place du locuteur par rapport à cet événement (ex. : victime, secouriste, habitant des quartiers touchés, habitants plus éloignés, ...). Dans ce cadre, nous avons déjà étudié les marqueurs modaux en tant que marqueurs privilégiés pour appréhender les différentes attitudes du locuteur vis-à-vis d’un contenu propositionnel (Colli et al., 2024). La méthodologie d’analyse modale retenue dans nos analyses est celle de Gosselin (2010) qui considère que le verbe *dire* relève de cette catégorie des marqueurs modaux, postulat qui guide la présente étude. En outre, Gosselin (2010) affirme que sans être directement désignée comme telle, la catégorie de la modalité est pourtant convoquée dans les travaux en psychologie et en psycholinguistique qui s’intéressent au repérage de marqueurs pertinents pour l’analyse de discours post-traumatiques (i.e. les verbes liés à des processus cognitifs ou les marqueurs d’émotions positives ou négatives). Une première analyse de la modalité selon cette approche dans un corpus qui renvoie à un événement traumatique, le corpus 13/11 qui sera présenté plus loin, a permis de constater que le verbe *dire* est le marqueur le plus fréquemment employé dans le corpus.

Nous proposons ici de nous pencher de manière plus précise sur les différents emplois de ce marqueur dans le corpus 13/11 en cherchant à mieux décrire en quoi ce marqueur participe de la variation d'un récit en fonction de la place du locuteur par rapport à un événement traumatique. Nous présentons tout d'abord en Section 2. les différents emplois et natures sémantiques du verbe *dire* reconnues généralement dans la littérature ainsi que notre corpus de travail. En Section 3., nous présentons la méthodologie et les résultats de notre analyse. Nous terminons avec une discussion de nos résultats (Section 4.) et la conclusion (Section 5.).

2. Corpus de travail et cadre théorique

2.1. Corpus de travail

Notre corpus d'analyse, 13/11, est composé de 1700 entretiens semi-directifs retranscrits, menés consécutivement aux attentats du 13 Novembre 2015 à Paris et Saint Denis. Ils ont été collectés, à deux reprises, en 2016 et 2018, dans le cadre de l'étude 1000 du Programme 13 Novembre¹ et correspondent à la réponse de l'interviewé à la question ouverte « Racontez-moi votre 13 novembre 2015 », posée par l'enquêteur². Les textes de notre corpus varient entre 2.044 et 48.869 tokens avec une moyenne de 13.781 tokens par entretien. Les interviewés sont divisés en quatre cercles selon une distance plus ou moins grande par rapport aux lieux des attentats. Ainsi, le cercle 1 (697 interviewés) regroupe les survivants, les témoins, les intervenants professionnels (médecins, policiers, etc.), ainsi que les proches et les endeuillés. Le cercle 2 (271 interviewés) rassemble les habitants et les usagers des quartiers directement touchés. Le cercle 3 (271 interviewés) correspond quant à lui aux habitants de l'Île-de-France, tandis que le cercle 4 (531 interviewés) comprend les habitants des villes de Metz, Caen et Montpellier.

2.2. Le verbe *Dire*

Authier-Revuz (2020), Vincent and Perrin (2001) et Rendulic (2015) proposent des analyses sémantiques fines du verbe *dire* qui visent à identifier différents types de discours rapportés introduits par ce verbe. Authier-Revuz (2020) identifie plusieurs types de discours rapporté introduit par *dire*, tels le discours direct et indirect et le *dire* intérieur ("*je me suis dit*"). Vincent and Perrin (2001) focalisent leur analyse sur le *dire* intérieur. Rendulic (2015), pour sa part, identifie les mêmes emplois mais aussi la macrostructure des échanges rapportés (abordés en Section 3.2.). Authier-Revuz (2020) évoque également certains emplois du verbe *dire* que certains désignent sous le terme de "marqueurs discursifs" e.g. (Anscombe, 2024) et qu'Authier-Revuz qualifie de marqueurs d'auto-représentation du *dire* (marqueurs ARD), dont, par exemple : *disons*, *on va dire*, *je veux dire*, *comment dire*, etc. Il s'agit de commentaires produits par l'énonciateur à propos de son propre *dire*, particulièrement employés à l'oral. Ces marqueurs sont caractérisés par un figement plus ou moins accentué sur les propriétés syntaxiques (Anscombe, 2024). Par exemple, pour les marqueurs ARD formés avec aller+dire nous retrouvons *on va dire* mais aussi *j'allais dire*, *je vais dire*, *je ne vais pas dire* ou *je vais presque dire*. Selon Gosselin (2010), *dire* est un marqueur modal polysémique. Il introduit un discours rapporté (direct ou indirect) et, selon le contexte, peut marquer une modalité épistémique — exprimant alors une évaluation subjective du locuteur — ou déontique, renvoyant plus précisément à une obligation (exprimée à travers la construction *dire + de + infinitif*). Par ailleurs, (Gosselin, 2010, p.100) ajoute que ce marqueur est toujours affecté d'une modalité aléthique parce qu'il décrit objectivement un

¹ www.memoire13novembre.fr. Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-10-EQPX-0021 Programme 13-Novembre.

² Nous travaillons ici sur l'ensemble des entretiens de phase 1 et 2.

acte de langage. En revanche, en ce qui concerne *dire* en tant que marqueur ARD, du moment que ces emplois ne décrivent pas un acte de langage, la valeur modale marquée, d'un point de vue sémantique, est exclusivement celle de la modalité épistémique. Elle peut ensuite se combiner avec les modalités exprimées par les autres marqueurs qui accompagnent *dire* comme, par exemple, la modalité déontique avec valeur d'auto-permission marquée par *pouvoir* dans *si je peux dire*. Selon la définition de Gosselin (2010), les marqueurs modaux, dont le verbe *dire*, permettent au locuteur de valider ce dont il parle selon six modalités différentes, c'est-à-dire six attitudes différentes : aléthique, épistémique, appréciative, axiologique, boulique et déontique. Dans cette perspective, nous avons appliqué une ressource lexicale (Battistelli and Étienne, 2025) qui permet de repérer automatiquement ces marqueurs et les classer selon les six modalités identifiées dans Gosselin (2010). L'application de cette ressource à notre corpus a permis de mettre en lumière une différence d'emploi des marqueurs modaux selon les cercles et de constater que le verbe *dire* est le marqueur modal le plus fréquemment employé dans le corpus 13/11 (Colli et al., 2024). En outre, l'analyse textométrique présentée dans Lacoste et al. (2024) identifie le dispositif citationnel, dont le verbe *dire* constitue un marqueur, comme une spécificité du cercle 1 par rapport aux cercles 3 et 4. Pour cette raison, nous nous proposons dans cet article d'analyser le verbe *dire* dans les 4 cercles de notre corpus à travers ses cooccurrents à l'aide de l'outil TXM (Heiden, 2010). L'objectif est de pouvoir faire émerger, en regroupant les cooccurrents plus fréquents en différents ensembles, de faire émerger différents *clusters* des emplois de *dire* et d'observer alors quels sont ceux qui sont privilégiés par tel ou tel cercle.

3. Analyse des cooccurrents de Dire

3.1. Méthodologie

Afin d'analyser les emplois du verbe *dire* dans le corpus, nous nous appuyons sur la méthode de l'analyse des cooccurrences³, qui permet d'identifier, à l'aide d'un indice statistique, les associations significatives entre deux mots ou deux lemmes dans une fenêtre de mots déterminés. Comme Mayaffre (2014) l'indique, le cooccurrent d'un mot permet de le co(n)textualiser en le sémantisant. Nous avons décidé ici d'analyser de manière indépendante le contexte gauche (fenêtre de 5 mots) et le contexte droit (fenêtre de 9 mots) parce qu'ils nous permettent d'acquérir des informations différentes sur ce verbe : le contexte gauche va permettre a priori d'identifier les occurrences des marqueurs ARD et certains types spécifiques de discours rapporté (p.e. le *dire* intérieur et les échanges rapportés, dont nous parlerons en section 3.2.), alors que le contexte droit va fournir a priori des informations sur d'autres types de discours rapporté (discours direct vs indirect) et sur le contenu du discours rapporté. Afin d'interpréter les résultats de manière uniforme, nous avons décidé de prendre en compte dans notre analyse les 15 premiers cooccurrents par cercle. Nous examinons les marqueurs qui sont présents dans les listes et leur ordre d'apparition. Nous complétons les résultats avec des exemples extraits du corpus à l'aide de l'outil *concordances* de TXM.

³ Calcul des cooccurrences sur les lemmes, indice de spécificité => 2, fréquence => 2.

3.2. Analyses

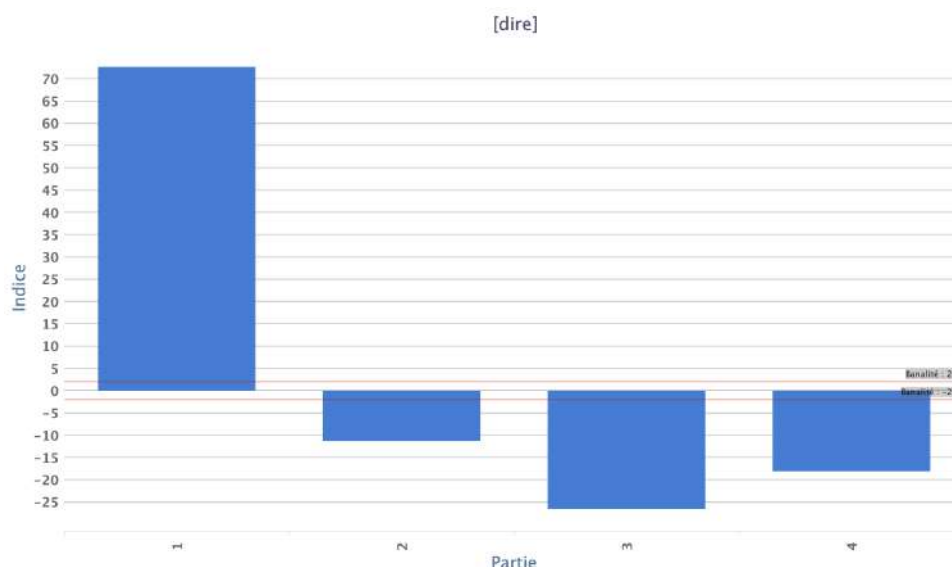


FIGURE 1 : Indices de spécificités du lemme *dire* dans le corpus 13/11 par cercle.

Pour commencer notre analyse du verbe *dire* à l'intérieur de notre corpus, nous menons une analyse des spécificités⁴ qui nous permet de constater que le verbe *dire* est sur-employé en cercle 1 par rapport aux cercles 2, 3 et 4 (cf. Figure 1). Nous présentons la liste des 15 premiers cooccurents à gauche de *dire* en table 1.

CERCLE 1			CERCLE 2			CERCLE 3			CERCLE 4		
n°	Cooccurrent	Indice	n°	Cooccurrent	Indice	n°	Cooccurrent	Indice	n°	Cooccurrent	Indice
1	comment	292	1	se	288	1	être suivre	278	1	lui	249
2	aller	289	2	comment	205	2	se	244	2	se	229
3	et	267	3	on	163	3	comment	185	3	comment	193
4	"	218	4	je	127	4	on	177	4	être suivre	191
5	pouvoir	181	5	me	127	5	vous	127	5	je	185
6	comme	180	6	vous	127	6	vouloir	127	6	vous	127
7	appeler	153	7	être suivre	127	7	me	127	7	me	127
8	avoir	127	8	lui	127	8	lui	127	8	vouloir	127
9	je	127	9	vouloir	127	9	je	127	9	aller	127
10	on	127	10	elle	116	10	aller	113	10	pouvoir	118
11	me	127	11	nous	87	11	pouvoir	101	11	nous	104
12	vous	127	12	pouvoir	84	12	elle	77	12	comme	64
13	se	127	13	avoir	65	13	nous	64	13	incapable	44
14	elle	127	14	comme	39	14	\$	39	14	elle	35
15	être suivre	127	15	et	38	15	comme	31	15	nous	24

TABLE 1 : Liste des 15 premiers cooccurents à gauche du verbe *dire* (contexte de 5 mots) dans les 4 cercles du corpus 13/11.

Parmi les cooccurents gauche de *dire*, en cercle 1, nous remarquons un premier *cluster* de lemmes qui renvoient, comme attendu, aux emplois de *dire* en tant que marqueur ARD. Nous

⁴ L'indice de spécificités de Lafon (1984) consiste à "mesurer le caractère attendu ou exceptionnel de la fréquence d'un mot dans une partie du corpus, au regard de sa fréquence dans l'ensemble du corpus et de la taille de la partie" (Pincemin, 2022, p.3)

notons cet ensemble d'emplois de *dire* **ClusterDire_ARD**. Il correspond à l'emploi des ad-
verbes *comme* et *comment*, ainsi qu'à celui des verbes *aller*, *pouvoir* et *vouloir*. Une analyse
des cooccurrences dans leur contexte large à l'aide de l'outil *concordances*, dont nous présen-
tons ici quelques extraits en guise d'exemples (1-3), permet de le confirmer.

(1) *Sauf que suite à cet événement, j'ai eu un surmenage, on va dire, d'alcool très fort, donc ça veut dire quand même, grosso-modo à boire un litre, voire un litre cinq de bière.* (PAR0086)

(2) *Enfin très vite quand quand effectivement, enfin ça se fait dans la dans la soirée de la veille, j'allais dire la soirée du 13 novembre.* (PAR0389)

(3) *bah là où on était assis, donc elle voulait prévenir et, enfin comment je pourrais dire, et peut-être rassurer les gens qui cherchaient à la joindre, donc je suis allé le chercher, donc ça m'a permis de voir ...* (PAR0491)

Nous remarquons parmi les occurrences les plus fréquentes pour *aller+dire* les marqueurs *j'allais dire*, *j'allais pas dire*, *j'allais presque dire*, *on va dire*, *on allait dire*, *je vais dire*, pour *vouloir+dire* les marqueurs *je veux dire* et *ça veut dire*, pour *comme+dire* les marqueurs *comme on dit*, *comme je dis*, pour *comment+dire*, les marqueurs *comment dire* et *comment dirais-je* et pour *pouvoir+dire* les marqueurs *si je peux dire* ou *pouvez-vous me dire*, un marqueur qui est à la fois un marqueur ARD et de discours rapporté, toujours prononcé par l'intervieweur afin de relancer l'interviewé. Nous remarquons aussi des combinaisons de marqueurs ARD comme dans le cas de *comment je pourrais dire* (cf. Exemple 3).

Parmi les cinq premiers cooccurents, nous notons la conjonction *et* et les guillemets. L'analyse des concordances nous permet d'identifier ces deux cooccurents comme des indices d'un deuxième *cluster* que nous noterons **ClusterDire_echanges** qui indique un type spécifique de discours rapporté que sont les échanges rapportés (Vincent and Perrin, 2001). Il s'agit d'une macrostructure qui englobe plusieurs occurrences (deux minima) de discours rapporté, surtout des discours directs, et qui est définie dans (Vincent and Perrin, 2001, p.187) comme « [un ensemble de] tours de parole ordonnés chronologiquement et causalement de manière à reconstituer une conversation qui a eu lieu dans le passé ». Cette macro-structure apparaît donc comme un enchaînement de discours directs, coordonné par la conjonction *et* et rythmé par la répétition du verbe *dire* qui se trouve à proximité des guillemets clôturant le discours direct précédent.

(4) *Et je leur ai dit : " Non, c'est...enfin c'est une impasse, du coup, on a, on a continué à courir sur le boulevard, et là il y a quelqu'un qui a... qui nous disait : « Rentrez, rentrez ». C'était un bel... bel immeuble haussmannien avec des grosses portes solides en fer, c'était agréable. Et quand je suis rentrée, il y avait ma soeur et mon frère, et il n'y avait plus mon père et Julien. Donc je leur ai dit : " Mais ils sont où ? " Ils me disent " bah je sais pas ". et tout, donc on ... je suis ressortie et en fait ils étaient en train de, ils avaient déjà traversé, ils étaient en train de remonter vers la maison. Et donc, ils ont dit : " Mais qu'est-ce que tu fous, et tout ? ". Alors je leur ai dit : " Mais venez, venez! " (PAR0003)*

(5) *Il nous demande ce qui se passe, donc on lui demande, on lui dit : " Bah il y a un attentat au Bataclan " et il nous dit : " Ah vous étiez au restaurant ", et on lui dit : " Non non on était au ... à la salle de concert ", et ... il dit : " Non, non il y a un a ... l'attentat c'est ... " (PAR0452)*

Dans (4) et (5), deux survivants interviewés mettent en scène des conversations qui ont eu lieu le soir du 13 novembre. Dans (4), PAR0003 raconte la mise à l’abri dans un immeuble après la fuite du Bataclan avec sa famille. Il est intéressant de noter comment plusieurs échanges avec différents interlocuteurs s’enchaînent les uns après les autres, en suivant les mouvements du locuteur. PAR0452 (exemple 5), de son côté, met en scène une conversation avec un chauffeur de taxi à la sortie du Bataclan. Les discours directs coordonnés par des conjonctions ou des pauses (...), la répétition du verbe *dire* et les verbes au présent permettent la “monstration” des tours de parole qui contribue à rendre la narration plus vivante et dramatique (Rendulic, 2015).

Nous comparons maintenant les listes des cooccurents des autres cercles, pour voir si les mêmes *clusters* définis en cercle 1 apparaissent également et si oui, dans quelle mesure. En ce qui concerne le **ClusterDire_ARD** nous retrouvons les mêmes cooccurents : *comment*, toujours parmi les trois premières positions, *pouvoir*, *aller* et *comme* avec l’ajout de *vouloir*. À la différence du cercle 1, ils apparaissent plus distribués dans la liste (en C1, 3 sur 5 marqueurs ARD occupent les 5 premières positions, alors que en C2, C3 et C4 un seul). Concernant le **ClusterDire_echanges**, nous remarquons, pour toutes les cercles, la disparition des guillemets et la descente du *et* en bas de la liste en cercle 2, 3 et 4, ce qui suggère une diminution de la fréquence d’emploi du dispositif des échanges rapportés. Nous notons également des différences concernant les marqueurs de positionnement des pronoms personnels et des auxiliaires (*avoir* et *être*|*suivre*⁵). Si ces marqueurs constituent un bloc unique en C1 (ils se trouvent tous en bas de la liste avec le même indice de cooccurrence), nous observons que ce bloc se défait dans les autres cercles. En particulier, nous remarquons que l’auxiliaire *être*|*suivre* cooccure ici plus fréquemment par rapport à *avoir* tout comme le pronom *se* (toujours en première ou deuxième position). L’analyse des concordances nous a permis de dégager un quatrième *cluster* comprenant ces deux marqueurs ainsi que le pronom *me*, tous les trois constituant des indices du *dire* intérieur (Authier-Revuz, 2020) que nous noterons ici **ClusterDire_interieur**. Il convient de noter que, avec le pronom *se*, le seul sujet accepté avec le verbe *dire* est le pronom *on* : il s’agit par conséquent toujours d’un *dire* intérieur. Ce n’est pas le cas pour *me*, qui accepte toute personne comme sujet (“*je me dis*”, mais aussi “*tu me dis*” ou “*ils me disent*”). En revanche, une analyse manuelle des cooccurrences à gauche de *me-dire* a confirmé que, dans environ 65% des cooccurrences en cercle 2, 3 et 4, *me-dire* renvoie à un *dire* intérieur. Dans les exemples (6) et (7) les locuteurs mettent en scène leurs pensées en obtenant, comme dans le cas des échanges rapportés, un effet de dramatisation de la narration qui passe à travers une monstration construite des tours de parole. Il est intéressant noter en (7), le passage de la première personne (*je me suis dit*) à la troisième personne (*on se dit*) : au moment où le locuteur entre en contact téléphoniquement avec une nouvelle personne son *dire* intérieur devient partagé, commun.

(6) *Et du coup, je l’ai appelée elle, parce que je me suis dit : " Si y a une personne qui peut savoir vite, 'fin qui peut obtenir vite des informations sur ce qui se passe, c’est elle en fait. Elle, elle pourra me dire euh ... si, si, 'fin, voilà, elle pourra trou ... elle pourra savoir en fait. ", parce que du coup on savait pas, on savait pas du tout ce qui se passait, à ce moment-là on, on pensait pas, en tout cas moi je pensais pas ... attentat quoi.* M-mh. * Je me souviens qu’on s’est dit : " C’est peut-être, il y a peut-être un truc euh ... ". (PAR0026)*

(7) *Au début, ils me disaient : " Ben bah, on va sortir etc. " donc au début, j’aurais dit : " Bon, je vous préviens, je suis pas trop trop chaude " et puis bah, au final, on se dit : " On va pas s’arrêter de vivre quoi! ", donc, on est sorti (PAR0132)*

⁵ Le l indique que TreeTagger hésite pour la lemmatisation entre le lemme *être* et *suivre*. Un contrôle manuel a confirmé qu’il s’agit entièrement d’occurrences du verbe *être*.

CERCLE 1			CERCLE 2			CERCLE 3			CERCLE 4		
n°	Cooccurrent	Indice	n°	Cooccurrent	Indice	n°	Cooccurrent	Indice	n°	Cooccurrent	Indice
1	putain	310	1	tu	192	1	tu	143	1	que	222
2	oh	283	2	"	127	2	ah	140	2	"	127
3	vous	265	3	:	127	3	que	127	3	:	127
4	toi	236	4	te	117	4	"	127	4	tu	110
5	!	218	5	aller	109	5	:	127	5	dire	88
6	Ecoutez	190	6	falloir	96	6	bon	112	6	te	85
7	si	182	7	bon	91	7	aller	104	7	mais	74
8	tien	170	8	que	86	8	ben	96	8	bon	68
9	il	163	9	mais	81	9	te	93	9	falloir	66
10	écoute	149	10	ah	70	10	bah	89	10	ben	60
11	bêtise	132	11	ben	68	11	falloir	85	11	aller	57
12	OK	127	12	bah	55	12	mais	67	12	tien	50
13	que	127	13	il	44	13	il	66	13	heure	47
14	mais	127	14	oh	43	14	tien	54	14	bah	47
15	"	127	15	putain	37	15	oh	53	15	il	46

TABLE 2 : Liste des 15 premiers cooccurrents à droite du verbe *dire* (contexte de 10 mots) dans les 4 cercles du corpus 13/11.

Nous présentons les 15 premiers cooccurrents à droite du verbe *dire* dans les 4 cercles en table 2. Nous remarquons que la plupart des cooccurrents à droite en cercle 1 sont des marqueurs indices du discours rapporté de type direct : des interjections (*putain*, *oh*), des signes de ponctuation (!) et d'autres marqueurs, parfois mal lemmatisés, employés comme introducteurs de discours direct (*Ecoutez*, *écoute*, *tien pour tiens*, *OK*, *bon*, *non*) qui forment un cinquième *cluster*, appelé **ClusterDire_discours_direct**, qui renvoie au discours direct. La liste des 15 premiers cooccurrents de cercle 2, 3 et 4 est, comme pour cercle 1, dominée par le **ClusterDire_discours_direct** des marqueurs indices de discours direct (*bah*, *bon*, " ", : , *ah*, *ben*, *oh*, *Ecoute*, *putain*, *tien*, *tenir*, *ah*). En revanche, nous remarquons ici l'apparition du cooccurrent *que* (distance moyenne de 3 entre *que* et *dire*) en cercle 3 et 4, indice du discours rapporté de type indirect introduit par *dire* ; ce marqueur constitue tout seul le **ClusterDire_discours_indirect**. Nous remarquons également la présence du verbe *falloir*, qui indique une nécessité ou une obligation exprimée soit par le locuteur lui-même (*dire* intérieur), soit envers son interlocuteur, ou encore par l'interlocuteur envers le locuteur. L'absence de la préposition *de* parmi les cooccurrents, et par conséquent de la construction *dire* + *de* à valeur déontique, indique que l'obligation déontique est ici davantage exprimée par un discours rapporté contenant un verbe déontique (*falloir*) que par l'utilisation directe du verbe *dire* avec une valeur déontique (cf. Exemples (8) et (9)).

(8) *Je sais pas, c'est vrai que ma collègue elle dit : " oui, il faut pas regarder les médias sont ... "* (PAR0263)

(9) *Je me dis : " Faut surtout pas que je, je parte dans ce délire et que je, et que ça, ça " ampute " sur ma, sur ma vie privée "* (CAE0121)

4. Discussion

L'analyse des 15 premiers cooccurrents de *dire* pour chaque cercle nous a permis d'identifier 5 *clusters* de marqueurs, résumés en Table 3. Afin de quantifier la présence de chaque *cluster* dans les 4 cercles nous avons calculé le pourcentage de couverture des indices des marqueurs par *cluster* sur le total des indices des 15 premiers cooccurrents (listes présentées en table 1 et 2) ; la catégorie "autre" comprend tous les marqueurs (dans les listes) qui n'appartiennent pas à un *cluster* identifié. Nous présentons les résultats en figure 2. En cercle 1, par exemple, le total

cluster	marqueurs
ClusterDire_ARD	<i>comme, comment, aller, pouvoir et vouloir</i>
ClusterDire_echanges	<i>“”, et</i>
ClusterDire_interieur	<i>me, se, être</i>
ClusterDire_discours_direct	<i>mais, te, bah, bon, OK, tien, ...</i>
ClusterDire_discours_indirect	<i>que</i>

TABLE 3 : Les clusters des emplois de dire définis en Section 3.2. à partir des 15 premiers cooccurents gauches/droite.

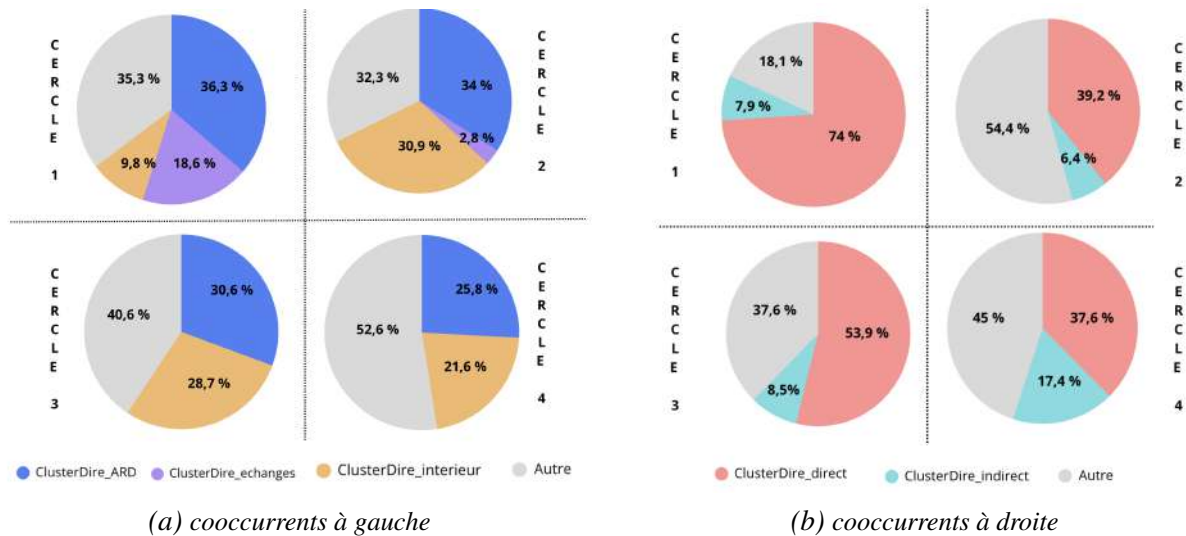


FIGURE 2 : % de couverture des indices des marqueurs par cluster sur le total des indices des 15 premiers cooccurents. ⁶

des indices des cooccurents faisant partie du *ClusterDire_ARD* couvre 36,3% du total des indices de cooccurrence des marqueurs de la table 1. La comparaison des listes pour chaque cercle a permis de mettre en lumière l’emploi préférentiel en cercle 1 (18,6%) par rapport aux autres cercles, du **ClusterDire_echanges** qui ressemble les indices des échanges rapportés et qui disparaît en cercle 3 et 4. Le locuteur préfère donc directement “montrer” la conversation à travers un enchaînement de tours de parole sous la forme de discours directs. Le résultat obtenu est cohérent par rapport à celui présenté dans Lacoste et al. (2024) d’une spécificité de l’emploi de la ponctuation citationnelle en cercle 1 par rapport aux cercles 3 et 4. Le dispositif de la monstration est mobilisé aussi en cercle 2, 3 et 4 par l’emploi du *dire* intérieur (**ClusterDire_interieur**), représenté par une cooccurrence majeure par rapport au cercle 1 (9,8% en cercle 1 vs 30,9% en cercle 2, 28,7% en cercle 3 et 21,6% en cercle 4) de l’auxiliaire *être* plutôt que de *avoir* ainsi que des pronoms *se* et *me*. Le locuteur fait part de ses réflexions en les mettant en scène à l’aide d’un discours auto-adressé. Le résultat obtenu semble s’écarter de celui présenté dans

Lacoste et al. (2024) L’emploi préférentiel des échanges rapportés, pour cercle 1, ainsi que du *dire* intérieur pour cercle 2, 3 et 4 reflètent la manière dont le locuteur a vécu l’événement qu’il raconte. En effet, si en cercle 1, au moment des attentats, la plupart des interviewés (survivants, intervenants professionnels et témoins) se retrouvent entourés des gens au bar ou dans la salle du concert, les interviewés de cercle 2, 3, 4, sont pour la plupart chez eux, en train d’aller se coucher ou devant la télé. Pour cette raison, le locuteur voulant mettre en scène la situation dont il parle montrera, en cercle 1, les conversations eues avec ceux qui l’entouraient alors qu’en

cercle 2, 3 et 4, il s'agira de ses réflexions intérieures. Nous faisons l'hypothèse, à tester par la suite, que les entretiens des survivants du Bataclan, par rapport aux autres victimes, sont caractérisés par une présence majeure du *dire* intérieur, du moment que la plupart a été obligée d'attendre, en silence, allongée dans la fosse avec les terroristes.

L'analyse du contexte droit montre un emploi préférentiel, dans toutes les cercles, du discours direct (**ClusterDire_discours_direct**), qui, comme le souligne Rendulic (2015), par rapport au discours indirect, permet de "mettre en scène" la conversation. En revanche, nous remarquons en cercle 4 une présence majeure de discours indirect introduit par *dire* (17,4% en cercle 4 vs 7,9% en cercle 1, 6,4% en cercle 2 et 8,5% en cercle 3). Cette "mise en scène" à l'aide du discours direct et des échanges rapportés en cercle 1 et du *dire* intérieur en cercle 2, 3 et 4, pourrait renvoyer au phénomène des reviviscences, dont la présence dans les récits d'événements traumatiques est signe d'un niveau plus élevé de troubles de stress post-traumatiques (TSPT). En particulier, Auxemery and Gayraud (2020) définit la reviviscence comme une "récitation de la scène traumatique" (Auxemery and Gayraud, 2020, p.520) qui s'actualiserait dans le langage par la présence de certains marqueurs linguistiques comme le temps présent ou les mots liés aux sensations. Le récit serait alors "une répétition littérale du trauma" (Auxemery and Gayraud, 2020, p.520), ce qui semble se reproduire ici aussi à travers le recours privilégié au discours direct, aux échanges rapportés et au *dire* intérieur.

5. Conclusion

Dans cet article, nous avons analysé les listes des 15 cooccurents les plus fréquents du verbe *dire* dans les quatre cercles de notre corpus afin d'observer comment l'emploi de ce verbe varie dans les récits des attentats du 13 novembre 2015 à Paris et Saint-Denis selon la distance des locuteurs par rapport à cet événement. Parmi les listes de cooccurents plus fréquents, nous avons pu retrouver des indices des emplois attestés dans la littérature tels que les échanges rapportés, le *dire* intérieur, le discours direct, le discours indirect et les marqueurs ARD. L'analyse présentée ici représente une étape préparatoire pour des analyses plus approfondies. En particulier, nous nous focaliserons sur la quantification précise, à l'aide de méthodes automatiques, de ses emplois afin de pouvoir confirmer avec des tests statiques les résultats obtenus par l'analyse des cooccurents. Nous faisons l'hypothèse d'un sur-emploi du discours direct et des échanges rapportés en cercle 1 et du *dire* intérieur en cercle 2, 3 et 4. Pour la suite, nous nous proposons de continuer à analyser les variations d'emplois en prenant en compte d'autres variables comme la phase (2016 vs 2018) ou le rôle à l'intérieur du même cercle (victimes Bataclan vs victimes terrasses, victimes vs secouristes). Nous nous proposons aussi d'explorer les variations d'emplois du discours direct, des échanges rapportés ainsi que du *dire* intérieur chez les individus atteints ou non par TSPT afin de pouvoir davantage explorer le lien entre les marqueurs de modalité et les indices linguistiques pertinents pour l'analyse des discours post-traumatiques.

Références

- Anscombre J.-C. (2024). Les marqueurs de discours. une catégorie linguistique? *Scolia. Revue de linguistique*, (38) :11–34.
- Authier-Revuz J. (2020). *La représentation du discours autre : principes pour une description*. De Gruyter.
- Auxemery Y. and Gayraud F. (2020). Le syndrome psycholinguistique traumatique (split). *L'Évolution Psychiatrique*, 85(4) :509–528.
- Battistelli D. and Étienne A. (à paraître en 2025). Modale, une ressource lexicale de la modalité au

- prisme des émotions. *Marqueurs modaux, énonciation et argumentation, Les Carnets du Cediscor*.
- Charretier L., Eustache F., Laisney M., Dayan J., Fraisse F., de La Sayette V., Gagnepain P., Chakli A., Klein-Peschanski C., Peschanski D., et al. (2024). Narration of a collective traumatic event according to the presence of ptsd and considering the social function. *Frontiers in Psychiatry*, 15 :1390470.
- Colli A., Battistelli D., and Chagnoux M. (2024). Quel usage des marqueurs modaux dans les discours post-traumatiques ? In *Marqueurs modaux, énonciation et argumentation*, Nantes, France. CerLiCO.
- Gosselin L. (2010). *Les modalités en français : la validation des représentations*, volume 1. Brill.
- Heiden S. (2010). The txm platform : Building open-source textual analysis software compatible with the tei encoding scheme. In *Proceedings of the 24th Pacific Asia conference on language, information and computation*, pp. 389–398.
- Lacoste C., Pincemin B., Heiden S., Peschanski D., Klein-Peschanski C., and Eustache F. (2024). Les mots du 13-novembre : Une première approche textométrique de l’«étude-1000». *Questions de communication*, 45(1) :221–244.
- Lafon P. (1984). Dépouillements statistiques en lexicométrie, slatkine. *Champion*.
- Mayaffre D. (2014). Plaidoyer en faveur de l’analyse de données co (n) textuelles. parcours cooccurentiels dans le discours présidentiel français (1958-2014). In *JADT 2014*, pp. 15–32.
- Pincemin B. (2022). Sémantique textométrique. *La sémantique au pluriel. Théories et méthodes*, pp. 373–396.
- Rendulic N. (2015). *Le discours représenté dans les interactions orales. De l’étude des structures en contexte vers la construction de l’image des relations interlocutives*. PhD thesis, Université d’Orléans.
- Vincent D. and Perrin L. (2001). Raconter sa journée : des histoires de conversations dans la conversation. *Revue québécoise de linguistique*, 29(1) :179–201.

EuroParser : un outil open-source pour collecter, structurer et exploiter de grands corpus de presse

Marceau Hernandez¹, Gaël Lejeune¹, Félix Alié¹, Émile Rebours^{1,2}, Julien Bezançon^{1,3}

¹CERES, STIH, Sorbonne Université – {prenom}.{nom}@sorbonne-universite.fr

²CELSA, GRIPIC, Sorbonne Université – {prenom}.{nom}@sorbonne-universite.fr

³LISN, Université Paris-Saclay – bezancon@lisn.fr

Abstract

The constitution of large corpora can be a tedious task. One would need to find a way to extract, through the different document sources, the various fields of interest, like the body, the author, and so on. A tediousness that increases with the number of sources and fields of interest. To help alleviate these issues, this paper introduces a tool that automatizes the corpus creation from a well known news collator outputs, EuroParser. The tool manages the extraction of the fields of interest, the normalization of these fields' values, and the generation of multiple outputs. This tool allows its users to convert the collator outputs to a structured format. This format is then used to provide multiple conversions that either allow the use of tabular tools (CSV and XLSX), textometry tools (TXM, Iramuteq, and Hyperbase), or other programmatic means (JSON and Markdown). Moreover, EuroParser also provides detailed statistics and visualizations on the extracted data. In order to provide an easy user experience, a web graphical interface was made with an emphasis on the ease of use. Finally, the project is meant to be easily modifiable : its source code is available online under the *AGPL-3.0-or-later* licence, and the structure is thought to ease changes in the main parser or in the downstream converters.

Keywords : newspapers, tool, textometry, parser, web scraping

Résumé

La création de grands corpus est une tâche pouvant s'avérer coûteuse. Cette dernière passe par l'identification et l'extraction, au sein de chacune des sources, des différents champs d'intérêt tels que le corps du texte et l'auteur. Qui plus est, ce coût croît avec la multiplication des sources utilisées et des champs recherchés. Dans l'optique d'amoindrir ce coût, cet article présente un outil automatisant la création de corpus à partir des sorties générées par un agrégateur d'articles de presse couramment utilisé, EuroParser. L'outil s'occupant alors de l'extraction des champs d'intérêt, de la normalisation des champs et de la génération de sorties exploitables. Cet outil offre à ses utilisateurs la possibilité de convertir les sorties de l'agrégateur en un format de données structurées. Ce format permet alors de proposer plusieurs conversions, offrant la possibilité de manipuler ces données dans un tableur (CSV et XLSX), des logiciels de textométrie (TXM, Iramuteq et Hyperbase), ou via d'autres moyens programmatiques (JSON et Markdown). De plus, EuroParser offre également plusieurs statistiques et visualisations sur le corpus constitué. Afin d'offrir la meilleure expérience utilisateur possible, nous avons également créé une application pour navigateur avec une emphase sur la facilité d'utilisation. Pour finir, le projet a été pensé pour être facilement modifiable : son code source est disponible en ligne sous la licence *AGPL-3.0-or-later*. La structure a également été pensée pour permettre d'apporter facilement des modifications à l'analyseur mais également aux convertisseurs finaux.

Mots clés : presse, outil, textométrie, analyseur, moissonnage

1. Introduction

La constitution de corpus constitue un enjeu central pour la linguistique de corpus et l'analyse du discours outillée. Les recherches portant sur les évolutions discursives contemporaines nécessitent des données situées dans le temps, cohérentes en terme de genre textuel et documentées par des méta-données riches et fiables. La presse représente à cet égard une source privilégiée, mais son exploitation systématique pose des difficultés méthodologiques et techniques importantes du fait de l'absence de corpus déjà constitués répondant à ces besoins.

Les ressources web généralistes permettent aujourd'hui de collecter des volumes textuels considérables. Cependant, l'usage direct du web comme corpus soulève des problèmes bien identifiés : hétérogénéité des sources, manque de contrôle sur la sélection des documents et reproductibilité limitée des résultats (Kilgariff, 2007). Les approches *Web-as-Corpus* ont proposé des pipelines de collecte et de nettoyage permettant de mieux encadrer ces pratiques (Baroni et al., 2009), sans toutefois résoudre entièrement la question des méta-données fines propres au genre journalistique. Sans faire une étude exhaustive de ce genre de corpus "attrape-tout", nous pouvons aussi citer l'avènement de ressources telles que Common Crawl (Smith et al., 2013) qui sont une des sources importantes pour l'entraînement des grands modèles de langues mais dont l'hétérogénéité reste problématique (Luccioni and Viviano, 2021). Se restreindre à des corpus plus ciblés (notamment en genre textuel) reste une approche raisonnable, mais qui suppose pour le chercheur d'identifier les pages HTML pertinentes à grande échelle (*crawling*) et d'en extraire automatiquement le contenu pertinent (*scraping*). Des travaux récents ont montré l'intérêt et les limites du *crawling* (Bermès and Moiraghi, 2020; Cartier, 2019) et de la difficulté de disposer d'outils de *scraping* robustes (Barbaresi, 2021) en particulier pour les langues autres que l'anglais (Barbaresi and Lejeune, 2020; Bournonville et al., 2025). Il faut également faire face à la présence de *paywalls*, qui limitent la complétude des données.

À l'inverse, les bases de presse commerciales telles que *Europresse* ou *Factiva* offrent des contenus à la fois contrôlés et fortement structurés, dont le compromis qualité/quantité intéresse les chercheurs en sciences humaines (Ragouet, 2023). Toutefois, l'accès et l'exploitation computationnelle à grande échelle de ces sources reste contraint par les formats disponibles en sortie, bien souvent la sortie proposée est faite pour être lue par un humain (PDF par exemple) et les méta-données sont ainsi computationnellement opaques. Dans ce contexte, nous présentons EuroParser un outil dédié à l'extraction et à la structuration de grands corpus de presse pour l'analyse statistique de corpus. Il vise à offrir une stratégie intermédiaire conciliant passage à l'échelle, reproductibilité et contrôle méthodologique. Cet article examine l'intérêt de cette approche à partir d'un état de l'art sur la constitution de corpus web et leur adéquation aux critères de qualité en textométrie (section 2.). Nous exposons ensuite les choix techniques et les fonctionnalités d'EuroParser (sections 3. et 4.) puis les perspectives de l'outil (section 5.).

2. Collecter des corpus de presse : opportunités et contraintes

2.1. Les exigences de l'analyse sur corpus

Pour l'analyse de corpus, la constitution des données ne relève pas d'une simple opération de collecte, mais d'un processus méthodique et théoriquement informé. Les grandes collectes "opportunistes" de type *Web as Corpus* ou *Common Crawl* ont offert de nouvelles perspectives dans bien des domaines, de la linguistique informatique à l'entraînement d'intelligences artificielles génératives, mais elles présentent d'importantes limites méthodologiques. Si elles offrent un accès massif à des données textuelles, celles-ci sont généralement peu contrôlées, faiblement

structurées et fortement hétérogènes (multiplicité des genres, des langues, des formats, présence de bruit). De tels ensembles de données relèvent davantage d’une logique d’exhaustivité technique que d’une logique de constitution de corpus. Ils rendent difficile, voire impossible, la mise en œuvre de traitements textométriques exigeant une structuration fine des métadonnées, une homogénéité des unités textuelles et un contrôle des conditions de production. Comme le souligne Pincemin (1999), un corpus est un objet construit, dont la validité repose sur l’explicitation de critères de sélection, de délimitation et d’adéquation à une problématique donnée.

2.2. Comment répondre à ces exigences en pratique ?

Il y a donc une exigence sur le contenu textuel lui-même mais aussi sur la structuration des données, à leur homogénéité et à leur reproductibilité ce qui semble disqualifier ce type de collecte à grande échelle. On pourrait aussi post-filtrer ces sources de données, par exemple pour créer un sous-corpus de presse (Spliethöver et al., 2022) mais la bonne représentation des sources n’est nullement garantie, sans parler des textes tronqués car bloqués par des *paywalls*. À l’inverse, une constitution de corpus entièrement manuelle – par exemple par sélection et copie de textes – permet certes un contrôle précis des données, mais se révèle rapidement coûteuse, peu reproductible et difficilement extensible, en particulier dès lors que l’on vise des volumes significatifs ou des mises à jour régulières du corpus. On peut observer que de nombreux collègues extraient donc des corpus à partir de EUROPRESSE ou FACTIVA afin de combiner le meilleur des deux mondes. Une extension pour FACTIVA existe ainsi pour TXM. C’est pour cela que nous avons jugé utile de proposer un tel outil pour EUROPRESSE. Nous avons en effet constaté que les outils existants étaient rarement maintenus, plusieurs n’ayant pas eu de mise à jour depuis plusieurs mois, voire années¹. Ainsi, parmi les outils testés, seul l’un d’entre eux, Cortext (Davi et al., 2022), semblait fonctionner sans cependant offrir de conversion vers les sorties souhaitées.

L’usage de ces bases de données permet de répondre aux critères de cohérence et de pertinence grâce à la sélection explicite. Le *scraping* ne fait ici qu’opérationnaliser une sélection déjà pertinente, issue du système documentaire lui-même. La régularité est également renforcée par la combinaison requête/structuration automatique : les documents extraits présentent une structure éditoriale homogène, ce qui limite fortement le bruit typique des données issues du web ouvert. Le corpus est ainsi non parasité, condition essentielle pour les analyses textométriques. Le scraping de données contrôlées issues de bases documentaires permet de satisfaire de manière conjointe les critères de signifiante, d’acceptabilité et d’exploitabilité définis par Pincemin (cf. tableau 1 page 4). Il assure la cohérence et la pertinence du corpus par un cadrage explicite des sources et des requêtes ; il garantit sa représentativité, sa régularité et sa complétude grâce à la nature éditorialisée et structurée des données ; enfin, il favorise son exploitabilité en produisant des corpus homogènes, commensurables et de volume suffisant pour des analyses textométriques. Cette approche se situe ainsi à mi-chemin entre la collecte massive non contrôlée et la constitution manuelle, en combinant automatisation et maîtrise méthodologique.

2.3. Données computationnellement opaques et analyses automatiques

Une simple collection de fichiers HTML bruts ne suffit pas à toutes les analyses. En effet, la structuration des données, que ce soit en isolant les différentes parties textuelles (titres, légendes ...) ou en extrayant et typant les métadonnées permet d’améliorer la qualité des données et celle

¹ Par exemple : <https://fmoncomble.github.io/press-corpus-scraper> (2025),

<https://docs.cortext.net/data-parsing> (2021), <https://quanti.hypotheses.org/1416> (2018) ou

https://groupes.renater.fr/wiki/txm-users/public/tutoriel_europresse (2018)

	Critère (Définition)	Limite des Approches opportunistes	Valeur Ajoutée (VA) des Agrégateurs de presse	VA du Scraping de données contrôlées
Significativité	Cohérence (objet et angle déterminés)	Agrégation sans unité forte	Sélection explicite (sources, période)	Encodage des méta-données
	Pertinence (adéquation aux objectifs)	Corpus générique peu ajusté	Requêtes ciblées, meilleure adéquation avec les hypothèses	Conservation des observables
Acceptabilité	Représentativité (fidélité à un espace discursif)	Biais implicites liés à la disponibilité	Choix contrôlé de sources, représentativité maîtrisée	N/A
	Régularité (données non parasitées)	Données hétérogènes, bruit (HTML, duplications)	Données éditorialisées	Extraction maîtrisable, dédoublement
	Complétude (ampleur et couverture cohérente)	Corpus partiel, couverture incertaine	Extraction fiable par requête (complétude relative)	Peu de limite à la taille des données
Exploitableté	Homogénéité (données commensurables)	Formats hétérogènes, granularité instable	Unités définies par l'utilisateur	Structuration (JSON/XML/CSV)
	Volume (Taille suffisante pour significativité)	Volume inadéquat (silence ET bruit)	N/A	Volume ajustable via automatisation

TABLE 1 : Limites des approches opportunistes et valeur ajoutée des agrégateurs et du scraping de données contrôlées au regard des critères de constitution des corpus de Bénédicte Pincemin

des analyses qui en découlent :

- **Intégrité textuelle** Retirer le “bruit” autour du contenu textuel d’intérêt² et préserver la structure et l’ordre du texte, la littérature parle parfois de détournage (Lejeune et al., 2015) ;
- **Granularité temporelle** Agréger et comparer les données par jour, mois ou année, afin de favoriser les études longitudinales et diachroniques à différentes granularités ;
- **Normalisation** Agréger des formes équivalentes, cela peut passer par le passage de dates à un format unique ou encore standardisation des noms de source (par exemple, en rassemblant "Le Monde," "LE MONDE," et "lemonde.fr" sous une même égide "Le Monde", le résultat de ce traitement est représenté par la figure 1) ;
- **Métadonnées additionnelles** Comparer quantitativement diverses parties du corpus (par exemple, avec une Analyse Factorielle des Correspondances ou AFC) avec toutes les méta-données extractibles (source, mots-clé, auteur·ice).

La plupart des bases de données commerciales d’articles de presse produisent des exports adaptés à la lecture humaine, au travers de formats HTML ou PDF, pensés pour structurer la donnée à l’aide d’indices visuels et parfois changeants. Par exemple, au sein des fichiers HTML générés par *EuroPresse*, les noms des auteur·ice·s peuvent se retrouver dans une balise “div” avec “docAuthors” comme classe (unique), mais aussi dans l’une (généralement la dernière des balises “p” ayant “sm-margin-bottomNews” dans ces classes (potentiellement plusieurs par document).

Les informations de structure se retrouvent également à plusieurs niveaux de l’arbre sans motivation structurelle apparente mais plutôt pensée pour l’affichage. Cependant, les outils dédiés à l’analyse textuelle requièrent des structurations particulières. Par exemple, pour *Iramuteq* (et *Hyperbase*, nous avons besoin d’avoir un seul fichier ‘.txt’ contenant l’ensemble des articles avec des lignes spéciales de séparation et d’annotation³. *TXM* exploite lui d’autres formats, par exemple des fichiers XML intégrant les métadonnées ou encore des fichiers TXT sans métadonnées complétées par un fichier CSV contenant ces métadonnées. Nous avons déjà pu recenser plusieurs usages d’EuroParser, notamment pour l’analyse : de changements sémantiques (Bordier et al., 2024; Martinc et al., 2024) ; diachronique d’un phénomène (Abiven and Véron, 2025) ; ou d’un cadrage médiatique (Garinat, 2024; Rebours, 2025). Ces usages nous

² C’est-à-dire isoler le corps de l’article des commentaires, options de navigations, publicités, et cætera

³ Exemple d’entête de document dans *Iramuteq* : **** *source_Le-Monde *annee_2023 *titre_[...]

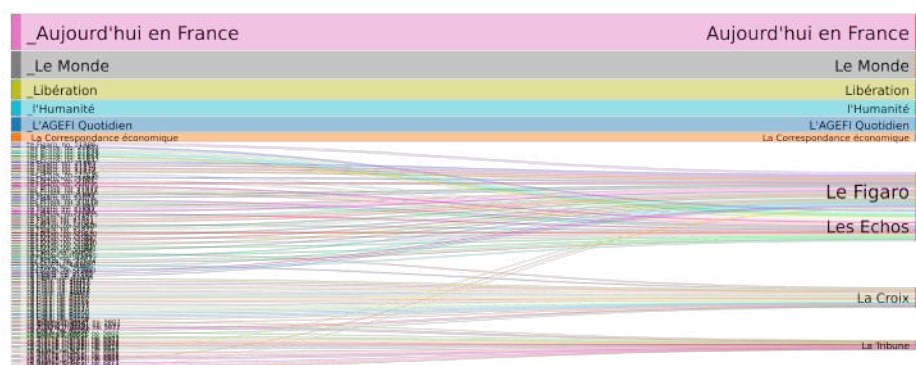


FIGURE 1 : Diagramme de flux exprimant les correspondances entre les noms des journaux dans les fichiers et les noms homogénéisés par EuroParser sur un corpus de test.

ont alors permis de faire évoluer EuroParser en fonction des besoins concrets. Les deux sections suivantes détailleront les choix techniques effectués dans la conception d’EuroParser afin de répondre au mieux aux besoins de corpus de presse pour les recherches en SHS.

3. Fonctionnement d’EuroParser

3.1. Articuler autour du pivot : Un schéma pour tous les structurer

Afin de pallier la complexité de parser des sources, contenant des informations parfois disjointes (notamment l’auteur·ice ou la rubrique), EuroParser utilise une structure de données intermédiaire, un *pivot* indépendant de la structure de la sortie. Ainsi, le système gagne en modularité : l’ajout d’une nouvelle entrée requiert seulement un parseur produisant le pivot depuis cette source. Ceci simplifie la maintenabilité du projet tout en diminuant le coût d’entrée pour d’éventuel·le·s contributeur·ice·s au projet.

Finalement, en plus de nous permettre de découpler l’entrée de la sortie finale, le pivot nous permet également de sauvegarder le résultat du passage. Ainsi, nous pouvons convertir notre collecte vers de nouveaux formats directement à partir du pivot.

3.2. Parser la presse : structuration des articles

L’agrégateur donne en sortie un fichier HTML unique par export. Pour parachever l’extraction, nous utilisons *BeautifulSoup*⁴ afin de naviguer au travers du fichier pour isoler les articles puis y extraire les informations souhaitées (corps du texte et métadonnées). Cette phase d’extraction d’informations est également complétée par le calcul de plusieurs informations, nous procédons alors à (i) l’extraction de mots clés, basée sur Daniel (Lejeune et al., 2010), qui utilise les similitudes entre chapô et corps des articles, (ii) la reconnaissance de la langue, à l’aide de *py3langid*⁵ et (iii) l’ajout d’un identifiant unique, provenant de l’agrégation des titres, dates et noms de journaux. Pendant l’extraction, nous procédons également à un dédoublement en s’assurant qu’aucune entrée ne partage d’identifiant, la collecte produisant naturellement des doublons strictement identiques.

⁴ Lien vers la page du projet : <https://www.crummy.com/software/BeautifulSoup/>

⁵ Lien vers le dépôt du projet : <https://github.com/adbar/py3langid>

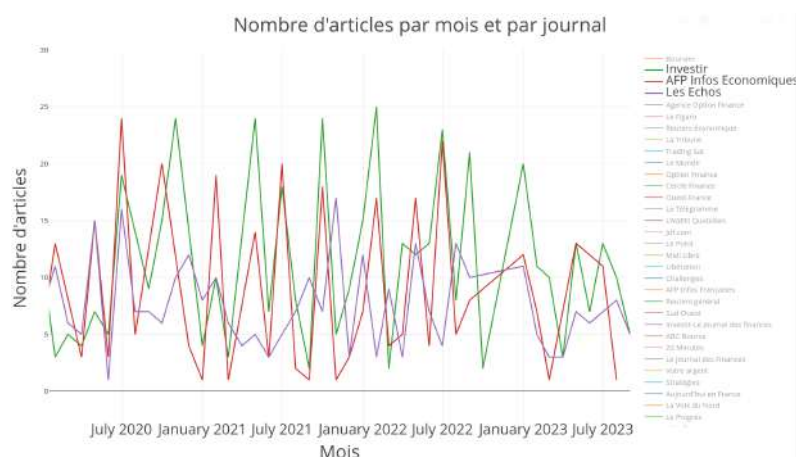


FIGURE 2 : Exemple d’une sortie graphique avec EuroParser sur notre corpus de référence. Seuls trois journaux sur une période restreinte ont été affichés par le biais de la partie interactive.

3.3. Exploiter la presse : plusieurs sorties pour plusieurs usages

EuroParser peut produire plusieurs sorties pour différents usages :

- **Pivot**, c’est une version parsée (en JSON) des articles pouvant être par la suite retransformée, c’est aussi un format pratique pour utiliser les articles extraits programmatiquement ;
- **Données tabulaires**, les formats CSV et XLSX permettent d’importer les articles extraits vers des logiciels de tableurs ;
- **Textométrie**, des sorties adaptées à TXM (Heiden et al., 2010), Iramuteq (Ratinaud, 2009) et Hyperbase (Vanni, 2024) ;
- **Lecture**, un export markdown qui contient des métadonnées, permettant la lecture ainsi que le tri via les logiciels adaptés ;
- **Référencement**, un export Bib_T_EX pour citer facilement les articles utilisés ;
- **Statistiques et exploration**, une description des métadonnées est donnée en JSON et proposée dans des graphiques interactifs en HTML.

De plus, avant de produire une sortie, nous pouvons venir appliquer une série de filtres sur le pivot : soit en exportant le pivot, en le filtrant, et en important les articles dans le format pivot ; soit directement dans l’outil. L’outil, offre alors la possibilité de retirer les mots clés qui n’apparaissent qu’une fois ou bien de les filtrer selon la ou les langues des articles.

3.4. Visualiser la presse : une première entrée dans le corpus à l’aide de statistiques

Comme évoqué ci-dessus, en plus de parser les résultats de l’agrégateur et de convertir ces résultats en divers formats, EuroParser propose également une première analyse du corpus produit. Afin d’illustrer notre propos, nous avons utilisé un corpus récolté par nos soins, ce dernier est composé d’articles de presse nationale française à partir de la requête “Ubisoft”, nous avons ainsi obtenu des articles datant de mars 1995 jusqu’à septembre 2023, au moment de la collecte.

Le transformateur dédié aux statistiques permet de générer un ensemble de métriques clés. Ainsi, sur notre corpus de référence, nous avons accès à plusieurs informations :

- La **distribution temporelle**, le nombre d’articles par mois est globalement en croissance tout au long de notre empan temporel, avec un pic en octobre 2015⁶ à 516 articles ;
- La **distribution en sources**, les journaux les plus présents étant *Boursier* (2 710 articles), *Investir* (2 127), *AFP Infos Economiques* (1 801) et *Les Echos* (1 555) ;
- Les **contributions individuelles**, même si nombre d’articles ne sont pas signés ou avec des signatures génériques (par exemple, “les Échos”), EuroParser permet d’isoler les deux contributeurs récurrents : John W. (684 articles) et Cloé W. (202).

Sorties graphiques En plus de ces mesures statistiques, EuroParser produit également des graphiques interactifs produits à l’aide de la librairie *Plotly*⁷. Ces derniers permettent de rendre compte des distributions calculées. Ces graphiques peuvent alors permettre de constater facilement des variations d’intensité de couverture médiatique pour ensuite en chercher la cause. La figure 2 présente l’un des graphiques obtenus à partir de notre corpus de référence. L’idée des statistiques et graphiques qu’EuroParser produit est d’avoir un aperçu des données d’entrée avant de faire des analyses plus approfondies (par exemple via des outils de textométrie).

4. Utiliser EuroParser

Un des objectifs principaux d’EuroParser est de rendre le traitement de corpus de presse accessible à l’ensemble des utilisateurs, indépendamment de leur profil technique. Cette volonté se traduit par trois types de mise à disposition de l’outil.

4.1. Bibliothèque Python

Premièrement, EuroParser est une bibliothèque Python. Elle expose directement la chaîne de traitements et les modèles de données (*FileToTransform*, *Params*, *Outputs*), permettant d’intégrer EuroParser dans un script ou un programme existant. Les utilisateurs pourraient ainsi alimenter la chaîne de traitements avec des contenus déjà en mémoire, récupérer les résultats sous forme d’objets Python et les manipuler librement en aval. Ce mode d’utilisation s’adresse principalement aux chercheurs ayant une pratique de la programmation : il offre un contrôle fin sur chaque étape du traitement et facilite l’intégration au prix d’une gestion manuelle des entrées et des sorties par le code appelant.

4.2. Ligne de commandes

EuroParser a également une Interface en Ligne de Commandes (*Command Line Tool*, CLI). À partir d’un dossier contenant les fichiers exportés et d’une liste de sorties souhaitées on peut déclencher l’ensemble des traitements. La CLI est un compromis entre la souplesse et l’interfaçabilité de la bibliothèque et la simplicité d’un outil prêt à l’emploi. Elle convient particulièrement aux traitements par lots sur un poste local et ne requiert pas d’écrire de code. Elle s’intègre dans des scripts *shell*, ce qui la rend adaptée à des usages récurrents ou à l’automatisation de tâches périodiques (veille automatisée, mise à jour régulière de corpus, etc.).

4.3. Interface graphique

La troisième et dernière modalité est une interface graphique accessible via un navigateur (figure 3). Nous avons travaillé avec un *designer* afin de réfléchir aux usages attendus et à l’expérience utilisateur. L’objectif de cette interface est alors d’offrir une prise en main facile pour

⁶ Un pic qui s’explique par la couverture du salon *E3* ainsi que par une tentative de prise de contrôle de l’entreprise.

⁷ Lien vers le dépôt du projet : <https://github.com/plotly/plotly.py>

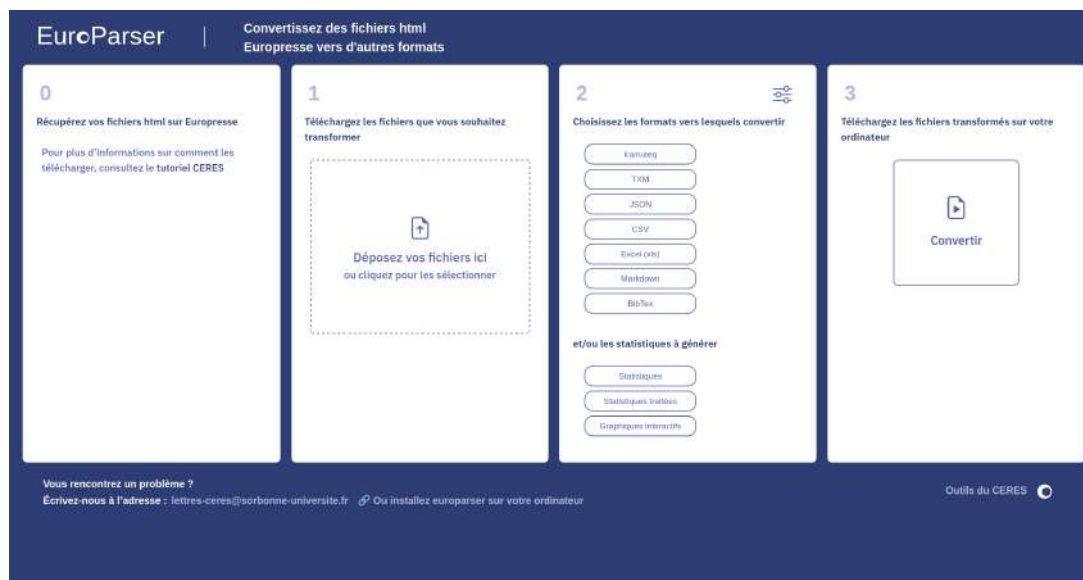


FIGURE 3 : Capture d’écran de l’interface graphique d’EuroParser. L’interface s’ouvre dans un navigateur et se décompose en quatre cadrans : (I) informations sur l’outil ; (II) sélection des fichiers ; (III) choix des options ; (IV) conversion et récupération des résultats.

l’exploitation de corpus de presse sans prérequis technique. Elle offre un parcours guidé : guides de collecte ; dépôt de fichiers ; choix des sorties et paramètres avancés ; puis téléchargement des résultats. Cette interface permet également un usage distant, la partie visible communiquant avec la chaîne de traitements à l’aide de requêtes HTTP. Nous pouvons ainsi proposer EuroParser en tant que service hébergé, accessible à un large public sans aucune installation locale⁸.

5. Conclusion et Perspectives

Dans cet article, nous avons introduit EuroParser, un logiciel libre ayant pour vocation de faciliter l’extraction d’articles de presse. L’outil a deux fonctions : (I) le *scraping* (ou détournage) de fichiers HTML ainsi que (II) la conversion vers des formats structurés avec des méta-données.

En (I) il isole le corps du texte et les autres zones d’intérêt du reste de la page, et la structuration, il récupère les zones d’intérêt dans des champs distincts et étiquetés (titre, date, ...). Ensuite, les métadonnées sont normalisées pour faciliter la comparaison. Par exemple, les dates, naturellement exprimées en plusieurs formats différents, sont normalisées au format ISO 8601. Il en va de même pour les noms de journaux, qui contiennent souvent des informations annexes, comme les numéros de revue ou encore le numéro de publication.

En (II), Afin de faciliter les usages que nous avons pu identifier, EuroParser offre la possibilité de récupérer les articles ainsi extraits dans un ensemble de formats (XML, CSV, JSON, ...). Afin d’offrir une première entrée dans le corpus produit, ces exports multiples sont également complétés par des options de calculs statistiques et de générations de graphiques. De plus, l’architecture a été pensée pour faciliter l’ajout de nouveaux formats d’exports et les contributions au projet.

EuroParser permet ainsi d’obtenir, dans de multiples formats, un corpus structuré d’articles de presse. Cependant, pour ce faire, nous avons besoin d’exploiter les résultats produits par les

⁸ Démonstration en ligne : <https://ceres.huma-num.fr/europarser/>.

agrégateurs de presse à l’aide d’un extracteur spécifiquement conçu pour chaque agrégateur. À ce jour, seul un agrégateur est supporté, *EuroPresse*, limitant alors la portée de l’outil. Une première piste d’amélioration serait alors d’implémenter un extracteur sur de nouveaux agrégateurs, tels que *Factiva*. Nous avons également pensé à ajouter une extraction plus générique basée sur le rendu visuel. Cependant ceci utiliserait significativement plus de ressources (100 fois plus lent selon (Lejeune and Barbaresi, 2020)) rendant le passage à l’échelle très coûteux. Dans une même logique visant à augmenter la flexibilité de l’outil, nous pensons ajouter de nouveaux formats de sortie, notamment pour permettre l’export vers d’autres outils comme *AntConc*, et de nouvelles sorties statistiques et graphiques, en fonction des usages constatés. Nous envisageons également de proposer, en étape intermédiaire, un dédoublonnage des articles qui laisserait le choix aux utilisateur·ice·s des seuils de similarité ainsi que des articles à conserver.

Considérations éthiques

EuroParser est un outil construit et pensé pour la recherche ouverte : l’ensemble du code produit pour ce projet est disponible en ligne sous licence libre⁹ ; l’architecture du projet a été pensée pour permettre la contribution, de nouveaux parseurs et de nouvelles sorties peuvent être ajoutés sans avoir à connaître l’ensemble du projet et peuvent être maintenues indépendamment.

Références

- Abiven, K. and Véron, L. (2025). Du transfuge de classe au class defector : définition, histoire et circulation d’un motif narratif contemporain. *CONTEXTES. Revue de sociologie de la littérature*, (36).
- Barbaresi, A. (2021). Trafilatura : A web scraping library and command-line tool for text discovery and extraction. In Ji, H., Park, J. C., and Xia, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, pages 122–131. Association for Computational Linguistics.
- Barbaresi, A. and Lejeune, G. (2020). Out-of-the-box and into the ditch ? multilingual evaluation of generic text extraction tools. In Barbaresi, A., Bildhauer, F., Schäfer, R., and Stemle, E., editors, *Proceedings of the 12th Web as Corpus Workshop*, pages 5–13, Marseille, France. ELRA.
- Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3) :209–226.
- Bermès, E. and Moiraghi, E. (2020). Le patrimoine numérique national à l’heure de l’intelligence artificielle. *Revue Ouverte d’Intelligence Artificielle*, 1(1) :89–109.
- Bordier, C., Martinc, M., Hernandez, M., Pollak, S., and Lejeune, G. (2024). Plongée dans le lexique du conspirationnisme dans la presse nationale française. In *Proceedings of 17th International Conference Journées d’Analyse statistique des Données Textuelles*, Bruxelles, Belgium.
- Bournonville, A., Dias, G., et al. (2025). Multilingual evaluation of main content extractors for web pages. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2853–2857. Association for Computing Machinery.
- Cartier, E. (2019). Néoveille, plateforme de repérage et de suivi des néologismes en corpus dynamique. *Neologica : revue internationale de la néologie*.

⁹ Sous la licence *AGPL-3.0-or-later* : <https://www.gnu.org/licenses/agpl-3.0.html>.

- Davi, B., Truchet-Aznar, S., Aznar, O., and Aubert, F. (2022). Apports de l’approche bibliométrique et de l’outil Cortext Manager à la revue de littérature scientifique. Une application à la prise en compte de l’environnement en Sciences Régionales (1999-2019). *NOV’AE - Numéros spéciaux*, 9 :1–18.
- Garinat, D. (2024). La pathologisation de l’éco-anxiété dans la presse française. La dépolitisation et la disqualification des discours de la “ jeunesse ”. Master’s thesis, Sorbonne Université.
- Heiden, S., Magué, J.-P., and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In Bolasco, S., Chiari, I., and Giuliano, L., editors, *Proceedings of 10th International Conference Journées d’Analyse statistique des Données Textuelles*, volume 2, pages 1021–1032, Rome, Italy. Edizioni Universitarie di Lettere Economia Diritto.
- Kilgarriff, A. (2007). Last words : Googleology is bad science. *Computational linguistics*, 33(1) :147–151.
- Lejeune, G. and Barbaresi, A. (2020). Bien choisir son outil d’extraction de contenu à partir du web. In *JEP/TALN/RECITAL 2020*, pages 46–49, Nancy, France. ATALA et AFCP.
- Lejeune, G., Brixtel, R., and Lecluze, C. (2015). Évaluation intrinsèque et extrinsèque du nettoyage de pages web. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 95–101.
- Lejeune, G., Lucas, N., and Doucet, A. (2010). Tentative d’approche multilingue en extraction d’information. In *Proceedings of 10th International Conference Journées d’Analyse statistique des Données Textuelles*, pages 1259–1267.
- Luccioni, A. and Viviano, J. (2021). What’s in the box ? an analysis of undesirable content in the common crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 182–189.
- Martinc, M., Laruncet, A., Bordier, C., Hernandez, M., Pollak, S., and Lejeune, G. (2024). The Evolution of Bias in French News Media : How Does Political Orientation Affect Semantic Change ? In *Proceedings of 17th International Conference Journées d’Analyse statistique des Données Textuelles*, Bruxelles, Belgium.
- Pincemin, B. (1999). Construire et utiliser un corpus : le point de vue d’une sémantique textuelle interprétative. In *Atelier Corpus et TAL : pour une réflexion méthodologique, Conférence TALN*, volume 99, pages 26–36.
- Ragouet, P. (2023). Sébastien lemerle, le cerveau reptilien. sur la popularité d’une erreur scientifique. *Sociologie*, 14(2) :263–265.
- Ratinaud, P. (2009). Iramuteq : Interface de r pour les analyses multidimensionnelles de textes et de questionnaires.
- Rebours, É. (2025). L’articulation des notions d’écologie et de punition dans la sphère publique globale : conflictualité médiatique et processus de (dé)légitimation autour de la formule “ écologie punitive ” dans la presse française. Master’s thesis, Sorbonne Université.
- Smith, J., Saint-Amand, H., Plamadă, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st annual meeting of the association for computational linguistics*, pages 1374–1383.
- Spliethöver, Keiff, and Wachsmuth (2022). Commoncrawl news articles by political orientation.
- Vanni, L. (2024). Hyperbase web.(hyper) bases, corpus, langage. *Corpus*, (25).

C05

Italian institutional and cultural corpora

09-lug
17:00 - 18:30
Aula 12

La sfida del lavoro da remoto. Analisi di focus group a responsabili delle risorse umane

Francesca della Ratta-Rinaldi¹

¹INAPP – f.dellaratta@inapp.gov.it¹

Abstract

This paper presents a statistical textual analysis of five focus group transcripts promoted by INAPP (Istituto Nazionale per l'Analisi delle Politiche Pubbliche) and involving HR (human resources) managers from public administrations and private firms, aimed at exploring advantages and challenges of remote work in the Italian post-pandemic context. The corpus consists of full discussion transcripts (56,737 tokens and 6,587 type) and displays typical dialogical features such as polyphony, turn-taking and co-constructed discourse. The analysis combines lexicometric tools and automatic classification techniques. In a first step, Taltac2 was used to examine word frequencies, keywords and lexical specificities across focus groups. In a second step, the corpus was processed with Iramuteq through Reinert's Descending Hierarchical Classification and similarity analysis in order to identify the main semantic structures and discursive polarizations.

Results reveal lexically coherent classes aligned with the composition of the groups (large firms, small and medium enterprises, public administration, consultants) and a relational structure centred on the node "smart working" (the Italian expression used to define the new and advanced telework), polarized between a regulatory-organizational dimension and an experiential-relational dimension. The study highlights the ability of statistical text analysis techniques to formalize complex discursive structures in dialogical corpora and to support comparative interpretation across homogeneous speaker groups.

Keywords: Telework, Text mining, Focus groups analysis

Riassunto

Il contributo presenta un'analisi statistica delle trascrizioni di cinque focus group promossi da INAPP (Istituto Nazionale per l'Analisi delle Politiche Pubbliche) e realizzati con responsabili delle risorse umane di enti pubblici e imprese private, finalizzati a esplorare vantaggi e criticità del lavoro da remoto nel contesto italiano post-pandemico. Il corpus, costituito dalle trascrizioni integrali delle discussioni (56.737 occorrenze e 6.587 forme diverse), presenta caratteristiche tipiche dei testi dialogici, quali polifonia, turni di parola e co-costruzione del discorso.

L'analisi integra strumenti di lessicometria e classificazione automatica. In una prima fase, mediante Taltac2, sono state esaminate frequenze, parole chiave e specificità lessicali dei singoli focus. In una seconda fase, il corpus è stato analizzato con Iramuteq attraverso classificazione gerarchica discendente e analisi delle similitudini, al fine di individuare le principali strutture semantiche e le polarizzazioni discorsive.

I risultati evidenziano una segmentazione in classi lessicalmente coerenti con la tipologia dei gruppi (grandi imprese, piccole e medie imprese, pubblica amministrazione, consulenti) e una struttura relazionale centrata sul nodo "smart working", articolata tra una dimensione regolativo-organizzativa e una dimensione esperienziale-relazionale. Il contributo mette in luce la capacità delle tecniche di statistica testuale di formalizzare strutture discorsive complesse in corpora dialogici e di supportare l'interpretazione comparativa tra gruppi omogenei di parlanti.

Parole chiave: Lavoro da remoto, Text mining, analisi di trascrizioni di focus group

¹ Le opinioni sono espresse dall'autrice a titolo personale e non impegnano l'Istituto di appartenenza.

1. Introduzione

Con la pandemia il mondo del lavoro è stato investito da una profonda riorganizzazione: il “grande esperimento di massa” (Butera, 2020) iniziato nel 2020 con la sperimentazione forzata del lavoro da remoto, è “diventato una componente strutturale della società, ridefinendo le relazioni tra lavoro e vita privata, tra spazi fisici e digitali” (Istat, 2026). Se molti studi hanno approfondito il punto di vista di lavoratori e lavoratrici, sono meno diffusi contributi specifici che approfondiscano il punto di vista aziendale sulle opportunità e le sfide che portano le imprese o le pubbliche amministrazioni a sperimentare (o a concedere) il lavoro da remoto. Per approfondire questi aspetti INAPP (Istituto Nazionale per l'Analisi delle Politiche Pubbliche), nell'ambito di un progetto più ampio sullo studio del lavoro da postazione remota² ha realizzato cinque focus group a responsabili delle risorse umane (HR) di enti pubblici e di imprese private di grandi e medie dimensioni attive in Italia e di consulenti del lavoro, che hanno consentito di esplorare in profondità il punto di vista aziendale sulle sfide poste dall'adozione del lavoro da remoto nelle diverse realtà produttive (in totale hanno partecipato ai focus 38 persone). I settori di attività in cui operano gli intervistati coinvolti sono stati scelti tra quelli con maggiore diffusione di questa forma di lavoro (settore bancario e assicurativo, telecomunicazioni, comparti chimico e metalmeccanico, terziario avanzato). Dei cinque focus, quello dedicato alla pubblica amministrazione e ai consulenti aziendali sono stati svolti in presenza a Roma, mentre gli altri tre sono stati realizzati da remoto, consentendo di ampliare il riferimento territoriale e di conciliare le agende dei partecipanti.

La scaletta di intervista era finalizzata a coprire diversi aspetti: il punto di avvio era la descrizione degli accordi adottati, per passare a una valutazione del contesto attuale, in cui è stato chiesto agli intervistati di indicare vantaggi e svantaggi del lavoro da remoto insieme ai fattori che hanno spinto l'azienda ad adottarlo e a continuare a tenere in piedi le sperimentazioni, con particolare attenzione alle accortezze necessarie per garantire la coesione e la comunicazione aziendale. Un'attenzione particolare è stata dedicata agli spazi di lavoro e alla loro eventuale riorganizzazione. L'intervista si concludeva con la richiesta agli intervistati di fornire una loro visione sugli scenari del lavoro da remoto in futuro, sia per il mondo del lavoro sia per l'impatto generale di questa innovazione sugli spazi, le città e i tempi di vita. Tutti gli intervistati hanno partecipato con entusiasmo alle discussioni, concorrendo a delineare un'elevata qualità delle testimonianze raccolte, con un clima di confronto e scambio particolarmente costruttivo. I risultati emersi con le discussioni sono analizzati in dettaglio in una pubblicazione specifica (della Ratta, 2026); in questo lavoro ci si sofferma sull'analisi testuale delle trascrizioni che ha costituito un passaggio cruciale nell'integrazione tra approccio tradizionale qualitativo e analisi lessicale di tipo statistico.

2. Considerazioni preliminari sulle trascrizioni di focus group

Prima di addentrarci nella descrizione dei risultati è necessaria una premessa metodologica sulla natura del corpus analizzato. I focus group sono interviste di gruppo, svolte con l'ausilio di un moderatore e finalizzate a sviscerare un tema sfruttando l'interazione tra i partecipanti. Da questo punto di vista l'interesse del testo non risiede tanto nell'analizzare chi dice cosa, ma nell'insieme degli elementi che emergono di volta in volta, considerando che ai partecipanti

² Nel Programma Nazionale Giovani, Donne e Lavoro FSE+ 2021-2027, Piano INAPP 2023-2026 - Operazione a titolarità n. 1, attività 5 è prevista la realizzazione di un'indagine statistica sul lavoro da remoto: i focus group sono stati realizzati nella fase preliminare per le attività di costruzione del questionario di indagine.

viene chiesto di aggiungere ulteriori elementi, senza ripetere quello che è già stato detto in precedenza (Corrao, 2000). Pertanto, la trascrizione di focus group presenta caratteristiche peculiari rispetto ad altri materiali testuali comunemente analizzati nell'ambito della statistica testuale, quali documenti, interviste individuali o risposte a domande aperte di questionari, proprio per la sua natura intrinsecamente dialogica e polifonica (per ulteriori approfondimenti si rimanda a: della Ratta, 2005).

Dal punto di vista metodologico, ciò comporta almeno tre implicazioni. In primo luogo, la frequenza delle forme non riflette esclusivamente l'importanza attribuita complessivamente a un tema, ma anche la dinamica interattiva del gruppo: alcune parole possono essere riprese e rilanciate da più partecipanti, amplificando la loro visibilità statistica; altre, pur rilevanti, possono essere enunciate una sola volta ma in modo particolarmente denso.

In secondo luogo, l'unità di analisi non coincide con il singolo parlante, come accade nelle interviste individuali, ma con il discorso costruito dall'insieme dei partecipanti. Per consentire di attribuire i risultati a specifiche categorie di attori, è necessario quindi organizzare preventivamente le caratteristiche dei gruppi in modo da assicurare l'omogeneità in relazione alle variabili di interesse per la ricerca. Nel nostro caso si è scelto di comporre i gruppi in base al tipo di aziende rappresentate, distinguendo tra: a) responsabili HR nelle pubbliche amministrazioni (PA); b) responsabili HR in una grande impresa (GI); c) responsabili HR in una impresa media o piccola (a questo tipo di aziende sono stati dedicati due focus). È poi stato aggiunto anche un focus rivolto a consulenti del lavoro, soggetti specializzati nella consulenza e assistenza legale o fiscale che lavorano soprattutto per le piccole imprese.

In terzo luogo, rispetto alle risposte a domande aperte, il focus group tende a privilegiare l'enumerazione dei temi e la costruzione collettiva di mappe tematiche, piuttosto che l'approfondimento sistematico e ripetuto di un singolo argomento. Questo può tradursi in una maggiore dispersione lessicale, rendendo particolarmente utile l'integrazione tra analisi tradizionale qualitativa e gli strumenti di classificazione automatica finalizzati all'individuazione di nuclei tematici ricorrenti.

L'analisi condotta con Taltac2 (Bolasco, 2013) e Iramuteq (Ratinaud e Dejean, 2009) si inserisce in questa prospettiva integrata: da un lato consente di oggettivare alcune evidenze emerse nell'interpretazione qualitativa, dall'altro permette di individuare connessioni e polarità semantiche che, nella lettura lineare delle trascrizioni, potrebbero risultare meno evidenti. In tal senso, il focus group si configura come un materiale particolarmente sfidante ma al tempo stesso ricco per l'analisi statistica dei dati testuali, in quanto combina complessità interattiva e densità tematica.

3. Il corpus in esame

Il testo (comprensivo degli interventi del moderatore) è di dimensioni medio-piccole e conta nel complesso 56.737 parole, con 6.587 parole diverse e una ricchezza lessicale (rapporto tra vocabolario e parole diverse) media, pari a 11,7%.

Tra le parole che si incontrano nel testo con maggiore frequenza, vi sono naturalmente *lavoro*, *lavoratore* o *lavorare* (480 occorrenze in tutto), *smart working* o *lavoro agile* (366 occorrenze) ma anche *persone* (187) *azienda* (136) *dipendenti* (90) *formazione* (79) *personale* (65), insieme alle locuzioni *a casa* (80) e *in ufficio* (65). Le parole “piene” più frequenti sono visualizzate nella figura 1, che riporta il Wordcloud realizzato con il software Iramuteq.

connesse all'organizzazione del lavoro. Alla prima categoria sono ascrivibili segmenti riconducibili a 4 sottocategorie tematiche: a) il quadro delle regole, 2) gli attori, 3) i riferimenti temporali e 4) i diversi contesti di applicazione; mentre i segmenti riconducibili agli aspetti organizzativi, possono essere distinti in 5 sottocategorie: 1) l'organizzazione del lavoro, 2) gli effetti sui lavoratori e le lavoratrici, 3) il tema della leadership, 4) gli spazi e gli strumenti per aumentare l'efficacia del lavoro da remoto, 5) le sfide e i vantaggi⁴.

Forma grafica	Occorrenze	Scarto	Forma grafica	Occorrenze	Scarto
remoto	86	138,8	persone	187	36,4
privacy	15	120,6	produttività	44	36,3
lavoratore	111	99,9	scrivania	21	35,5
team	36	98,8	leadership	28	32,7
postazioni/e	32	72,8	noi	258	31,6
agile	24	62,0	aziendale	38	31,5
portatili/e	24	58,9	vantaggi	33	31,1
digitalizzazione	7	56,2	collega	34	30,3
telelavoro	22	55,3	affiancamento	6	30,2
portatile	15	46,8	assenteismo	8	29,9
azienda	136	46,3	spazi	39	29,8
dipendenti/e	137	45,4	manager	32	29,3
fisicamente	20	43,9	dirigenza	12	29,1
ferie	21	42,8	gestisce/gestire	49	28,2
lavorare/lavoro	369	40,2	riunioni	20	27,7

Tabella 1 Principali parole chiave del testo (occorrenze in valore assoluto e scarto standardizzato rispetto all'italiano standard)

Dal momento che nell'analisi dei focus group l'unità di analisi non coincide con il singolo parlante, il calcolo delle parole caratteristiche (quelle sopra o sotto-rappresentate rispetto alle modalità delle variabili associate al testo, Lebart e Salem, 1994; Tuzzi, 2024) è stata condotta in riferimento al tipo di impresa rappresentata dai partecipanti (pubblica amministrazione, grande impresa, piccola e media impresa o consulenti del lavoro). L'analisi delle parole specifiche (o caratteristiche) è stata realizzata utilizzando la procedura presente nel programma Taltac che consente di individuare, per ciascuna modalità analizzata, l'elenco delle parole specifiche ordinate per livello di significatività.

In questo modo, è stato possibile caratterizzare le singole discussioni – e quindi la tipologia delle aziende che hanno partecipato a ciascun focus group, attraverso le parole specifiche emerse per ciascuno di essi.

Ad esempio, è soprattutto nel focus svolto nelle grandi imprese che si è fatto riferimento al momento della *pandemia*, per ricordare come nella maggior parte dei casi il processo di sperimentazione del lavoro agile fosse stato già avviato in precedenza; sovra-utilizzati anche i termini specifici che rimandano alle caratteristiche degli *accordi* sindacali, all'*autonomia* concessa e raggiunta dalle persone che lavorano, ai margini di *libertà* dei manager in base alle esigenze delle *strutture*, alla necessità di *cambiare approccio* per raggiungere il successo e in

⁴ Per motivi di spazio non è possibile pubblicare la tabella che riporta i principali segmenti ripetuti (multiword expression), per ulteriori dettagli si rimanda a: della Ratta (2026).

definitiva alla *sfida* rappresentata dallo *smart working*. Interessante che si parli soprattutto delle *persone* che lavorano in azienda, mentre i termini *lavoratore* e *dipendenti* sono sotto-rappresentati rispetto alla media.

Nel focus dedicato al pubblico i termini sovra-rappresentati sono quelli usati per parlare della *dirigenza* e della *leadership* (*dirigente, dirigenti, risorse umane*), del numero massimo di *giornate* da svolgere in presenza e a distanza, al tema delle *ferie* non smaltite, al rapporto coi *sindacati*, alla concessione o meno del *buono pasto* e della possibilità di intervenire sugli spazi. Tra le parole sotto-rappresentate spiccano *vantaggi* e *produttività*.

Anche nelle piccole e medie imprese si parla soprattutto di *persone* (o *dipendenti*), anziché di *lavoratori*, insieme al riferimento agli specifici *progetti*, cui adattare le caratteristiche del lavoro da remoto, dei *vantaggi* per chi ha *bambini piccoli* e della *produttività* assicurata dal raggiungimento degli *obiettivi*, insieme all'importanza di assicurare momenti di *confronto* tra colleghi come quelli davanti alla *macchinetta del caffè*. Tra le parole sotto-rappresentate incontriamo *formazione*, i rimandi al conteggio delle *giornate in presenza* o *a distanza*, e al tema delle *ferie*, elementi che confermano il carattere più informale delle esperienze diffuse nelle *piccole aziende* (Osservatorio Politecnico Milano, 2025).

I consulenti, infine, si caratterizzano per un linguaggio più aderente alla normativa, con il richiamo ai temi della *sicurezza* e della *privacy*, all'*orario di lavoro* e al *controllo*, ai *compiti* di *datore di lavoro* e lavoratori, alle caratteristiche dei *contratti*, alla *formazione* da svolgere e alla possibilità o meno di far rinunciare alla *scrivania* individuale, insieme alle differenze tra *pubblico* e *privato*. Sotto-rappresentati i riferimenti alle *persone* e ai *giorni massimi* del lavoro a distanza.

4. Specificità, cluster analysis e similitudini

Per individuare le principali strutture semantiche, il testo è stato poi processato con Iramuteq al fine di quantificare i diversi elementi presenti nel testo, attraverso la scomposizione del corpus in frammenti (nel caso in analisi composti da 50 parole) per l'individuazione di cluster il più possibile omogenei (Reinert, 1983; Ratinaud e Dejjans, 2009): nel nostro caso 6 classi, sostanzialmente riconducibili a due macrocategorie (figura 2).

I tre cluster riportati sulla destra del dendrogramma (classe 6, classe 1 e classe 5) sintetizzano il 48,4% dei frammenti, e fanno riferimento agli **aspetti regolativi del lavoro agile**. Tra questi il n. 6 (14,5% dei frammenti) consente di riconoscere soprattutto il linguaggio dei consulenti del lavoro, che come appena visto si sono soffermati soprattutto sulle questioni della *privacy*, della *sicurezza* e delle *normative*; il n. 5 (11,6% dei frammenti) condensa invece il linguaggio utilizzato nel focus dedicato al settore pubblico (*amministrazione, pubblico, direzione, orario, risorsa*), mentre il più numeroso, il n.1, raggruppa il 22,1% dei frammenti e raccoglie la terminologia utilizzata per descrivere il quadro di regole: *accordo, smart, sindacale, individuale, giorno, settimana, mese, massimo, sottoscrivere, contrattazione, sede*.

I tre cluster di sinistra condensano invece il linguaggio riferito alle **sfide e vantaggi dello smart working**: il cluster più numeroso, che raccoglie il 23,2% dei frammenti (classe 4), raggruppa i termini riferiti ai *vantaggi* del lavoro da casa, sia per gli individui (specie per chi vive in *grandi città* come Roma, o coloro che hanno *bambini piccoli*), sia per le aziende, che possono veder ridurre i *costi* e mantenere inalterata o aumentare la *produttività*. Vi sono anche termini come *sfida, svantaggi, pensare o vedere*, che ricordano il percorso di ridefinizione organizzativa ancora in fase di evoluzione. Il cluster della classe 3 raggruppa invece le parole utilizzate per discutere dell'importanza di mantenere alta la *coesione* della squadra, attraverso iniziative di

team building o anche assicurando momenti di *incontro* e confronto *in presenza* come quelli passati davanti alla *macchinetta del caffè*, che è importante non perdere. Infine, il cluster n. 2 (13,7% dei frammenti) rimanda al tema della *leadership*, dei rapporti con i *più giovani*, del ruolo svolto dagli strumenti tecnologici (*tool*) e di alcune competenze (*soft skill*).

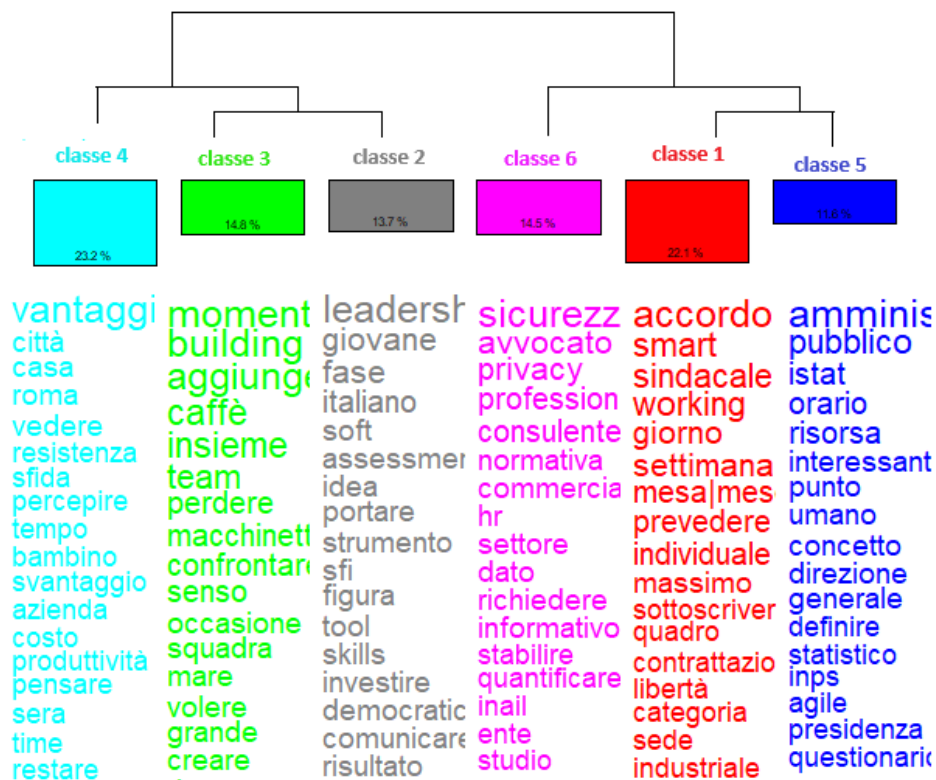
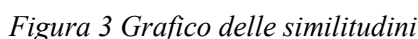


Figura 2 Scomposizione delle trascrizioni in classi: dendrogramma

Infine, il grafico delle similitudini (figura 3), costruito a partire dalle co-occorrenze tra le forme lessicali, consente di visualizzare la struttura relazionale del lessico e le connessioni più stabili tra i termini del corpus. A differenza della classificazione gerarchica discendente, che segmenta il testo in classi mutuamente esclusive sulla base della distribuzione delle forme nei frammenti, l'analisi delle similitudini restituisce la trama reticolare del discorso, mettendo in evidenza le associazioni semantiche più ricorrenti all'interno delle stesse unità di contesto elementare.

Nel nostro caso emerge una struttura fortemente centrata sul binomio lessicale “smart working”, che si configura come vero e proprio nodo semantico dominante. Va innanzitutto considerato che la centralità del binomio ‘smart working’ è in parte influenzata dalla frequente co-occorrenza dei termini smart e working, presenti congiuntamente 245 volte nel corpus. Tuttavia, il grafo consente comunque di mettere in evidenza le principali articolazioni semantiche del discorso: attorno a questo nucleo si articolano infatti due principali poli discorsivi, già in parte riconoscibili nella cluster analysis e qui ulteriormente confermati.

Il primo polo, di natura organizzativo-regolativa, collega *smart* a termini quali *azienda*, *accordo*, *giorni*, *settimana*, *sindacale*, *contratto*, delineando una rappresentazione del lavoro da remoto come oggetto di regolazione formale e di contrattazione. Tale area corrisponde in larga misura ai cluster alla destra del dendrogramma (1, 5 e 6, figura 2), riferiti agli aspetti regolativi del lavoro agile, nei quali prevalgono rispettivamente il lessico degli accordi individuali e collettivi, quello specifico della pubblica amministrazione e quello più tecnico-normativo dei consulenti.



Il secondo polo, più orientato all'esperienza organizzativa e agli spazi, mette in relazione smart working con *casa, persone, ufficio, team, presenza*, evidenziando la dimensione concreta e quotidiana della riorganizzazione del lavoro. Questo asse semantico risulta coerente con i cluster di sinistra (in particolare il 3 e il 4), riferiti alle sfide e vantaggi dello smart working, nei quali emergono, da un lato, i temi della coesione, del team building e dei momenti informali di incontro, e dall'altro i vantaggi e gli svantaggi percepiti, sia per i lavoratori sia per l'azienda.

L'intersezione tra i due poli – resa visibile dalla continuità delle connessioni nel grafo – suggerisce che nel discorso dei partecipanti la dimensione normativa e quella esperienziale non siano contrapposte, ma costantemente intrecciate: la regolazione degli accordi è discussa in funzione dei suoi effetti concreti su persone, spazi e processi produttivi. In questo senso, il grafico delle similitudini non solo conferma i risultati della classificazione, ma ne rafforza

l'interpretazione, mostrando la coerenza strutturale del discorso e la centralità dello smart working come categoria organizzativa complessa, che tiene insieme norme, relazioni e pratiche.

5. Conclusioni

La scelta di realizzare dei focus group rivolti a responsabili HR nasce dall'intento di arricchire le riflessioni sull'impatto del lavoro da remoto a partire dal punto di vista di chi organizza, gestisce e contratta nel concreto le modalità di fruizione del lavoro da remoto, individuando esempi calati in contesti specifici, che hanno confermato la stretta relazione tra la cultura organizzativa, l'adozione del lavoro per obiettivi e l'applicazione efficace del lavoro da remoto. Peraltro, il focus group ha consentito il confronto diretto tra gli interlocutori e si è configurato come uno strumento di ricerca "riflessivo", in grado di attivare processi di scambio e rielaborazione di valutazioni, buone pratiche ed errori da evitare.

Nel complesso, le discussioni hanno consentito di approfondire l'impatto del lavoro da remoto nella gestione delle varie fasi del ciclo produttivo, mettendo in luce la necessità di figure manageriali in grado di adottare nuove modalità di gestione, fondate sulla fiducia, la motivazione e la delega e non più sul controllo. Appare chiaro quindi come la figura del manager rivestirà un ruolo cruciale. Il funzionamento dei processi interni all'organizzazione sta cambiando: è meno sentita l'esigenza di figure intermedie di coordinamento, specie nei contesti che tendono a divenire più orizzontali.

Dato il tipo di testo, è stato necessario integrare un'analisi interpretativa di tipo qualitativo con strumenti di statistica testuale in grado di restituire una formalizzazione delle strutture discorsive. L'impiego combinato di analisi delle specificità, classificazione gerarchica discendente e grafico delle similitudini ha consentito di passare da un corpus dialogico complesso a una segmentazione strutturata in classi lessicali statisticamente fondate e in poli semantici chiaramente identificabili, uno strumento molto utile anche per la definizione del questionario di indagine costruito a valle della realizzazione dei focus group.

Dal punto di vista metodologico, il contributo mostra come l'applicazione delle tecniche di analisi testuale a trascrizioni di focus group possa aiutare a individuare classi semanticamente interpretabili e coerenti con le variabili esterne. Pertanto, l'esperienza qui presentata suggerisce che l'analisi dei focus group mediante tecniche di statistica testuale può contribuire alla costruzione dell'oggetto di studio, rendendo esplicite le strutture latenti del discorso organizzativo. In questa prospettiva, l'integrazione tra metodi rappresenta non solo una scelta tecnica, ma un'opzione epistemologica coerente con la complessità dei processi di trasformazione del lavoro in atto.

Riferimenti bibliografici

- Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma, Carocci.
- Butera F. (2020). Le condizioni organizzative e professionali dello smart working dopo l'emergenza: progettare il lavoro ubiquo fatto di ruoli aperti di professioni a larga banda. *Studi organizzativi*, 1.
- Corrao S. (2000). *Il focus group*, Milano, Angeli.
- della Ratta-Rinaldi F. (2005). L'interpretazione sistematica del materiale derivante da focus group attraverso l'analisi testuale. *Sociologia e ricerca sociale*, n. 76/77, pp.91-104.
- della Ratta-Rinaldi F. (a cura di) (2026). La sfida del lavoro da remoto. Evidenze sui focus group di manager pubblici e privati. *Inapp Paper*, n. 62. Roma, Inapp.
- Istat (2026). Smart working: da necessità a nuovo stile di vita. *Statistiche Today*, 25 febbraio 2026.

- Lebart L., Salem A. (1994). *Statistique textuelle*. Paris, Dunod.
- Osservatorio Smart Working della School of Management del Politecnico di Milano (2025). Lo Smart Working al tempo dell'AI: opportunità e sfide verso il lavoro del futuro. *Atti del convegno Politecnico di Milano*, ottobre 2025.
- Ratinaud P., Dejean S. (2009). IRaMuTeQ: implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. *Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009)*. Toulouse, Le Mirail.
- Reinert M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Cahiers de l'analyse des données*, 8, n.2, pp.187-198.
- Tuzzi A. (2024). *Fondamenti di analisi dei dati testuali*. Roma, Carocci.

Un’analisi testuale dei discorsi parlamentari italiani

Marco Nannetti¹, Matteo Farnè²

¹University of Bologna – marco.nannetti3@studio.unibo.it

²University of Bologna – matteo.farne@unibo.it

Abstract 1 (in English)

Parliamentary debates are a reliable source for text analysis using statistical techniques. They provide a solid foundation for studying a country’s political climate over the years. In particular, they allow for the implementation of techniques to extract topics from texts, analyze the language used by specific individuals or parties, and assess the sentiment conveyed by the text.

This paper applies these techniques to Italian parliamentary speeches contained in the *ItaParlCorpus* dataset (Cova, 2024). The dataset contains all speeches delivered in the history of the Chamber of Deputies (from the 1st to the 18th Legislature). This paper focuses on the 15th through 18th legislatures (between April 2006 and October 2022), with the goal of identifying the main themes of political debate using the Latent Dirichlet Allocation (LDA) model and grouping their content via clustering techniques.

The LDA analysis provides a clear and interpretable representation of the main topics discussed during the legislatures under consideration. Specifically, the latent themes of parliamentary speeches—and thus of Italian politics—during the period in question are extracted from the debates. Minor challenges arise, primarily due to the presence of many procedural speeches and proper names of parliamentarians or individuals. These generate topics that are difficult to interpret or completely meaningless and difficult to capture during the text cleaning phase.

Subsequently, hierarchical clustering techniques are applied to group the numerous topics into homogeneous clusters. The macro-topics are interpreted, and their distribution over time is derived. Finally, the sentiment of the parliamentary speeches under consideration is estimated using the Sentix lexicon (Basile and Nissim, 2013), and an exploratory sentiment analysis is proposed with respect to specific macro-topics, with reference to political parties.

Keywords: Text Mining, Latent Dirichlet Allocation, Sentiment Analysis, Italian Parliament

Abstract 2 (in italiano)

I dibattiti parlamentari sono una fonte affidabile per un’analisi testuale con tecniche statistiche. Essi costituiscono una solida base di partenza per lo studio del clima politico di un Paese nel corso degli anni. In particolare, essi consentono di implementare tecniche per l’estrazione degli argomenti dei testi, per l’analisi del linguaggio di specifiche persone o partiti, e per la valutazione del sentiment che il testo contiene.

Questo elaborato applica tali tecniche ai discorsi parlamentari italiani contenuti nel dataset *ItaParlCorpus* (Cova, 2024). Il dataset contiene tutti i discorsi tenuti nella storia della Camera dei Deputati (dalla Legislatura I alla XVIII). In questo elaborato, si pone l’attenzione sulle legislature che vanno dalla XV alla XVIII (tra aprile 2006 e ottobre 2022), con l’obiettivo prefissato di individuare i principali temi del dibattito politico usando il modello Latent Dirichlet Allocation (LDA) e di raggrupparne il contenuto tramite tecniche di clustering.

L’analisi LDA consente di ottenere una rappresentazione chiara e interpretabile dei principali argomenti discussi durante le legislature considerate. In particolare, si estraggono dai dibattiti le tematiche latenti degli interventi parlamentari, e dunque della politica italiana, nel periodo temporale considerato. Piccole criticità emergono e sono legate alla presenza di molti discorsi procedurali e nomi propri di parlamentari o persone. Questi generano argomenti di difficile interpretazione o completamente senza significato e sono difficili da catturare durante la fase di pulizia del testo.

Successivamente, vengono applicate tecniche di clustering gerarchico per raggruppare i numerosi argomenti in gruppi omogenei. I macro-argomenti vengono interpretati, e viene derivata la loro distribuzione nel tempo. Infine, viene stimato il sentiment dei discorsi parlamentari considerati tramite il lessico Sentix (Basile e Nissim, 2013), e viene proposta una analisi esplorativa del sentiment rispetto a macro-argomenti specifici, con riferimento ai partiti politici.

Parole chiave: Text Mining, Latent Dirichlet Allocation, Sentiment Analysis, Italian Parliament

1. Introduzione

Il Parlamento è molto più di un organo istituzionale: è lo specchio del paese, il luogo in cui prendono vita le voci, le tensioni e le speranze dell'Italia. Qui, giorno per giorno, vengono discussi tutti i temi di rilevanza politica e sociale per tutti i cittadini. I parlamentari che ne hanno ottenuto la fiducia elaborano leggi che permettono di orientare il futuro nazionale per il bene di tutti.

In questo lavoro, consideriamo un dataset che riporta tutti i discorsi parlamentari italiani, *ItaParlCorpus* di Cova (2024). Il dataset è organizzato in quattro gruppi di legislature (LEG I-V, LEG VI-X, LEG XI-XIV, LEG XV-XVIII) e contiene 14 variabili utilizzabili per svariati tipi di analisi. Negli studi recenti sul linguaggio politico italiano, il contributo di Cova con l'introduzione dell'*ItaParlCorpus* rappresenta un importante avanzamento nella disponibilità di dati testuali strutturati sui discorsi parlamentari. Il lavoro, pubblicato su Cambridge University Press all'interno della *Italian Political Science Review*, si colloca nel solco della crescente letteratura internazionale sui parliamentary corpora e sulle applicazioni di metodi computazionali alla scienza politica. Le analisi presentate sono principalmente di natura esplorativa e dimostrativa, includendo distribuzioni lessicali, analisi tematiche tramite topic modeling e primi tentativi di misurazione della polarità e dello stile retorico. Tali approcci si inseriscono in una tradizione consolidata di text-as-data (Grimmer e Stewart, 2013) e di studio quantitativo del discorso politico, ma contribuiscono specificamente al contesto italiano, finora meno coperto rispetto ad altri paesi europei. Il corpus apre inoltre a sviluppi futuri in ambito NLP, quali classificazione automatica, analisi del framing e studi longitudinali sulla trasformazione del linguaggio politico nel passaggio tra diverse fasi istituzionali.

Motivati da questo contesto, applicheremo la Latent Dirichlet Allocation (LDA) di Blei et al. (2003) alla variabile **text** di *ItaParlCorpus*, che riporta il testo del discorso in maniera fedele a come è stato pronunciato, per estrarre gli argomenti di discussione più rilevanti. Ci concentreremo sulle ultime quattro legislature del dataset, ovvero dalla XV alla XVIII (tra aprile 2006 e settembre 2022). Sugli argomenti ottenuti, opereremo poi una cluster analysis gerarchica col metodo del legame medio, per estrarre i macro-argomenti più rilevanti tra quelli identificati dalla LDA. Infine, applicheremo in modo selettivo la sentiment analysis usando il lessico di Sentix (Basile e Nissim, 2013; Basile, 2019), per identificare differenze sistematiche di sentiment tra i partiti principali.

2. Analisi degli argomenti tramite LDA

Restringendo la nostra attenzione alle legislature dalla XV alla XVIII, otteniamo 2613 giornate di discussione parlamentare (tra il 28 aprile 2006 e il 28 settembre 2022). Dapprima, selezioniamo i discorsi con più di 40 parole, escludendo i discorsi del Presidente, che ha come funzione la moderazione del dibattito. Successivamente, al fine di costruire la DTM (Document Term Matrix), ovvero la matrice che conta le occorrenze di ciascun termine del vocabolario in ogni documento, cominciamo la pulizia dei dati.

Come primo passaggio, rimuoviamo 1217 stopwords, che sono un mix di termini comuni del linguaggio italiano (preposizioni, articoli, pronomi, avverbi, congiunzioni e verbi ausiliari), identificato dal pacchetto *quanteda* di R (279), e stopwords parlamentari. Queste ultime (938) sono un elenco di termini generici tipici del linguaggio parlamentare e istituzionale, escluse in modo che non predominassero negli argomenti della successiva analisi LDA, lasciando maggiore spazio ai temi politici. In seguito, vengono tolti i numeri, la punteggiatura ed eventuali spazi multipli per permettere di proseguire con la tokenizzazione. Viene quindi creato il vocabolario delle legislature XV-XVIII, da cui sono rimosse tutte le parole rarissime, che ricorrono meno di 10 volte. Avviene poi la vettorizzazione e la creazione della DTM.

Data la numerosità campionaria, i modelli LDA sono stimati fissando un numero di argomenti compresi tra 50 e 95, con balzi di 5 topic per ridurre l'aggravio computazionale. La scelta è ricaduta sul modello a 70 topic, che presenta il picco di *mean logratio*, calcolata dal pacchetto R *text2vec*

(Selivanov et al., 2023) La Figura 1 riporta il profilo di questa metrica per $K = 50, 55, 60, \dots, 95$. La *mean logratio* è essenzialmente equivalente in questo contesto alla coerenza *UMass* di Mimno et al. (2011).

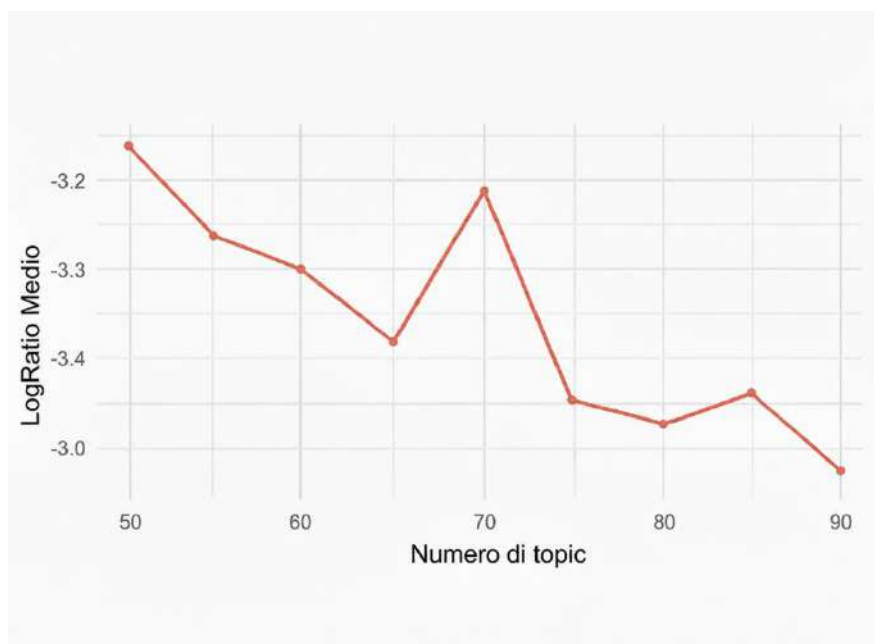


Figura 1: Metrica *mean logratio* del modello LDA applicato su *ItalParlCorpus* ristretto alle ultime quattro legislature (XV-XVIII).

3. Cluster Analysis tramite legame medio

Sui 70 topic estratti, è stata poi applicata una cluster analysis. La misura di distanza utilizzata in questo contesto è la Jensen-Shannon divergence (JS), la più adatta per il confronto di distribuzioni di probabilità (Salton e Buckley, 1988). La matrice in input presenta sulle righe i 70 topic individuati e sulle colonne tutti i termini del dizionario. I valori della matrice indicano la probabilità che una determinata parola appartenga a un topic; pertanto, ogni riga somma ad uno ed esprime la distribuzione di probabilità delle parole in un topic. Il metodo di aggregazione delle unità è il legame medio.

Vengono calcolati i valori di Silhouette media e del Pearson Gamma per confrontare le possibili divisioni in cluster. In particolare, come si nota dalla Figura 2, vengono considerate le possibili partizioni tra 5 e 40 cluster. È stata presa la decisione di dividere i topic in 19 cluster. Questa suddivisione è una via di mezzo che presenta ottimi valori per gli indici di Silhouette media e Pearson Gamma, anche se 19 non è il valore massimo per nessuno dei due indici, permettendo così di avere gruppi più interpretabili rispetto a suddivisioni alternative.

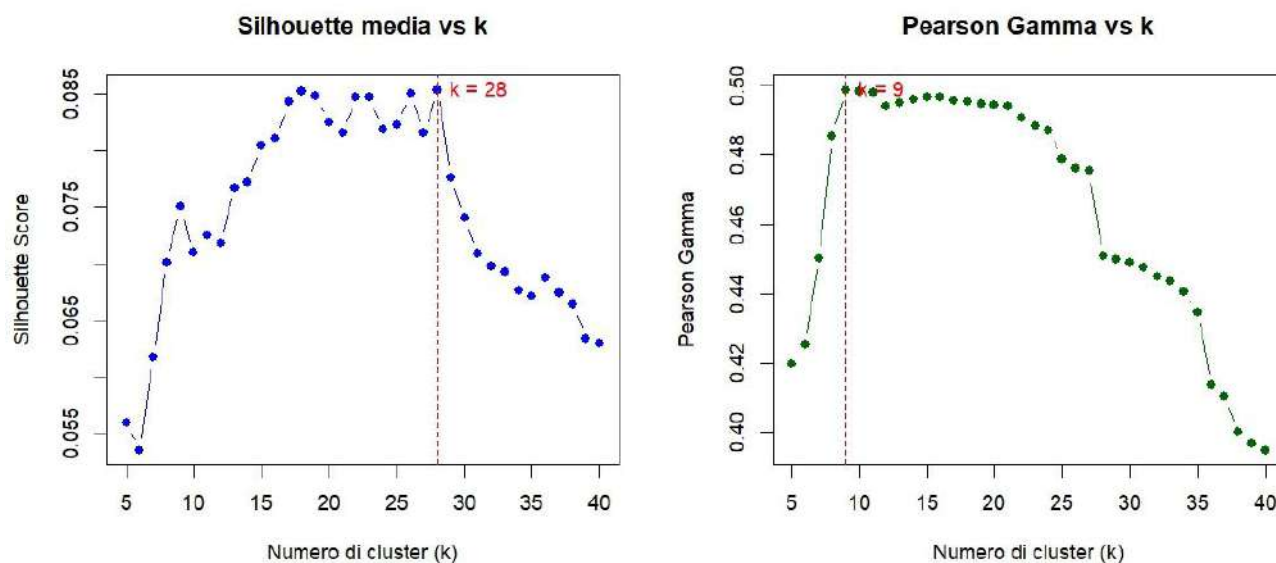


Figura 2: Indici di Silhouette media e Pearson Gamma della partizione a k cluster ($k = 5, 6, \dots, 40$) ottenuta dalla suddivisione LDA a 70 topic, tramite il metodo del legame medio, usando la Jensen-Shannon divergence come misura di distanza.

I macro-argomenti ottenuti sono “stopwords”, “cultura e sport”, “politica internazionale”, “difesa e immigrazione”, “politica territoriale”, “infrastrutture e trasporti”, “ambiente e calamità naturali”, “produzione e difficoltà economico-sociali”, “appalti pubblici e banche”, “bilancio statale e fisco”, “innovazione digitale e monitoraggio”, “scuola e personale pubblico”, “diritti civili di donne, bambini e disabili”, “sanità e covid”, “giustizia e carceri”, “crimini e violenza”, “elezioni e affari costituzionali”, “lavoro parlamentare”, “vittime e ricordo”, “stopwords 2”.

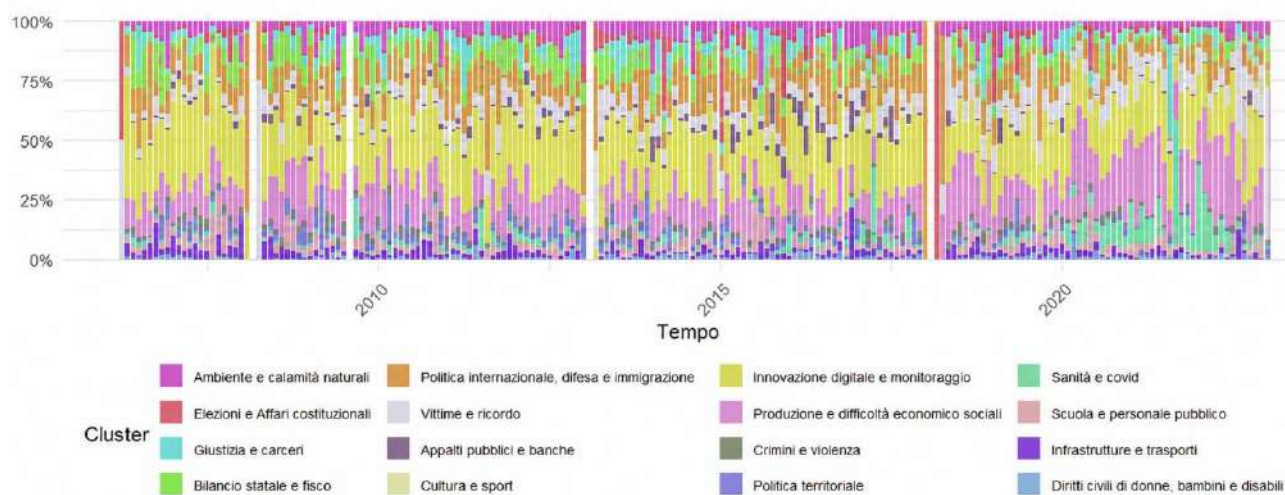


Figura 3: Distribuzione mensile dei cluster di argomenti estratti per i discorsi parlamentari relativi alle legislature XV-XVIII, rimuovendo dal computo i cluster “stopwords”, “stopwords 2”, “lavoro parlamentare”.

La Figura 3 è creata assegnando ad ogni discorso parlamentare il cluster al quale ha la probabilità più alta di appartenere. Vengono tolti i discorsi relativi ai cluster “stopwords”, “stopwords 2” e a “lavoro parlamentare”. Ogni barra verticale indica, con un colore diverso per ogni argomento, la percentuale di discorsi che in un mese vengono tenuti riguardo ai 16 cluster rimasti, senza considerare gli altri 3. Gli argomenti come “produzione e difficoltà economico sociali”, “innovazione digitale e monitoraggio”, “politica internazionale, difesa e immigrazione”, “bilancio statale e fisco”, “giustizia e carceri” e “infrastrutture e trasporti” sono trasversali in tutti gli anni poiché di importanza essenziale

per lo Stato. Il cluster “sanità e covid” ottiene un’impennata vertiginosa a partire dall’inizio del 2020, poiché in quei mesi l’incubo del Covid-19 e le difficoltà che creò colpendo tutta la popolazione divenne uno degli argomenti fondamentali per il quale era necessario prendere decisioni. Si notano tre “buchi” che rappresentano le elezioni di aprile 2008, febbraio 2013 e marzo 2018 ovvero i brevi periodi dove il Parlamento si ferma e lascia votare gli elettori che decidono democraticamente come verrà ricomposto.

4. Sentiment Analysis

La sentiment analysis viene effettuata sullo stesso dataset usato per la cluster analysis, ovvero quello contenente le ultime quattro legislature parlamentari, da cui sono stati tolti i discorsi del presidente. Viene usato principalmente il pacchetto R *udpipe*. Partendo dai discorsi integrali è stata effettuata la tokenizzazione, il POS (Part Of Speech) tagging, ovvero l’assegnazione di ogni parola alla sua categoria grammaticale (nome, verbo, aggettivo, pronome, articolo, ecc.), e la lemmatizzazione. Il modello usato nella sentiment analysis per tokenizzazione, POS tagging e lemmatizzazione è *italian-vit-ud-2.5-191206.udpipe* di Straka, M., e Straková, J. (2019). Questo modello è addestrato sul treebank VIT (Venice Italian Treebank), un corpus annotato italiano usato nel progetto Universal Dependencies (UD) (Nivre et al., 2020), una notevole collezione di framework annotati per più di 150 lingue.

Successivamente, con la funzione *sentix.df.sentiment* (integrata nel pacchetto R *sentixR* di Basile, 2019) viene calcolato il sentiment per ogni discorso parlamentare. Sentix è un lessico in lingua italiana sviluppato da Basile e Nissim (2013), che contiene i termini italiani con punteggio di polarità e calcola il valore del sentiment di ogni lemma moltiplicando polarità e intensità. Il risultato è un punteggio di sentiment compreso tra -1 e +1 per ogni termine. Inoltre, viene automaticamente calcolato anche il sentiment medio del discorso, ottenuto come media dei sentiment dei singoli termini.

Prima di analizzare il sentiment vengono inoltre tolti 38825 discorsi, circa il 20,94% del totale, aventi una probabilità maggiore del 30% di appartenere a “stopwords” o “stopwords 2”, in quanto essi non sono di interesse nell’analisi, perché riguardano votazioni, emendamenti, ordini del giorno parlamentari o rumore generico sui partiti.

In seguito, si sfrutta la divisione dei discorsi nei 19 cluster creati precedentemente per studiare come cambia l’enfasi dei dibattiti a seconda dell’argomento. Inizialmente, si calcola il sentiment per ogni macro-argomento in maniera fuzzy, usando le probabilità di ogni discorso di appartenere ai vari cluster. I risultati sono riportati nel grafico a barre dei risultati in Figura 4 (escludendo i due topic “stopwords”). I topic “scuola e personale pubblico”, “infrastrutture e trasporti”, “innovazione digitale/monitoraggio” e “cultura/sport” sono quelli che presentano il sentiment medio più alto. I topic che in generale vengono affrontati con accezioni più negative risultano essere “giustizia e carceri”, “crimini e violenza”, “ambiente e calamità naturali”.

Si è poi svolta un’altra analisi esplorativa del sentiment dove viene cambiato il metodo di assegnazione dei discorsi ai topic. Ogni riga di testo viene assegnata soltanto al topic che la descrive con probabilità maggiore. Il sentiment di una riga va a sommarsi a quello delle altre righe che ugualmente hanno probabilità maggiore delle altre per un unico topic k . Come si nota dalla Figura 5, i risultati sono molto simili, in quanto i cinque topic migliori e i cinque topic peggiori in termini di sentiment risultano essere gli stessi. Questa suddivisione dà luogo però a differenze di sentiment medio più nette. I punteggi medi dei topic con sentiment più alti si alzano ancora di più e si abbassano i punteggi dei topic più negativi. Anche in questo caso vengono esclusi i sentiment di “stopwords” (a cui erano stati assegnati 68 documenti) e di “stopwords 2” (a cui erano stati assegnati 3423 documenti).

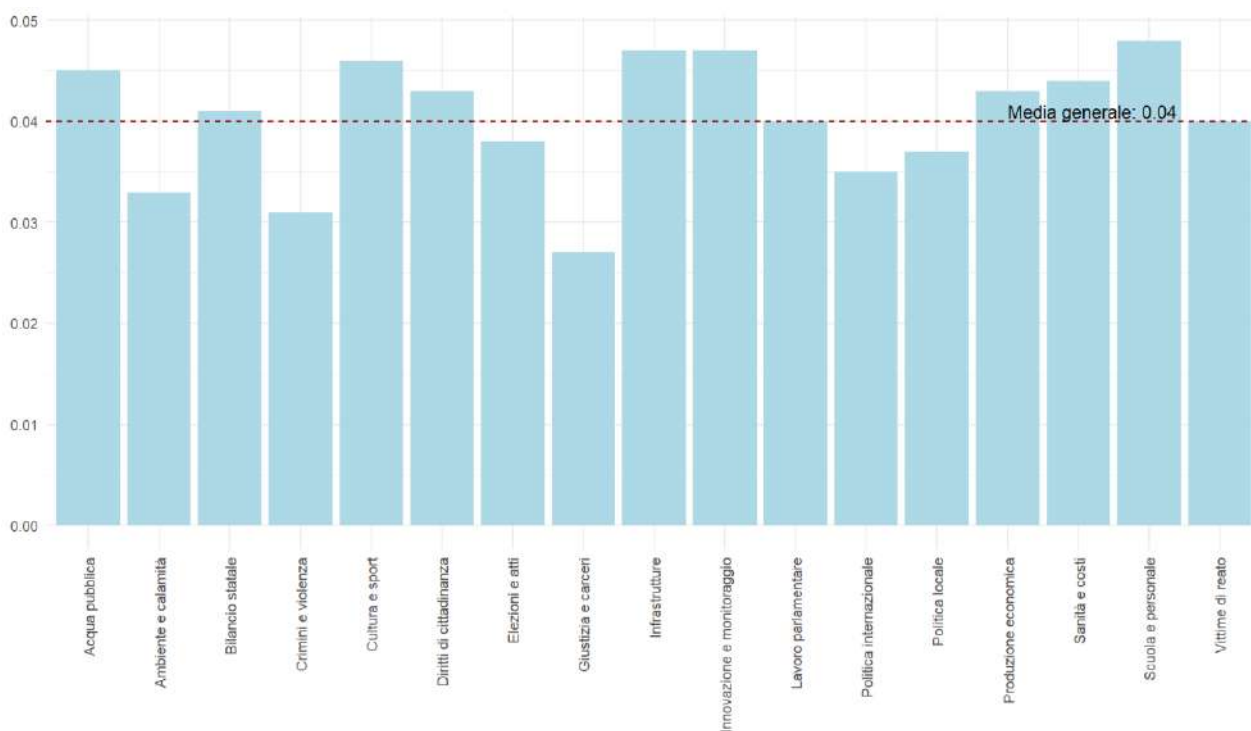


Figura 4: Grafico a barre del sentiment medio dei discorsi per ogni cluster, ottenuto pesando per la probabilità di appartenenza. La linea tratteggiata rappresenta il sentiment medio.

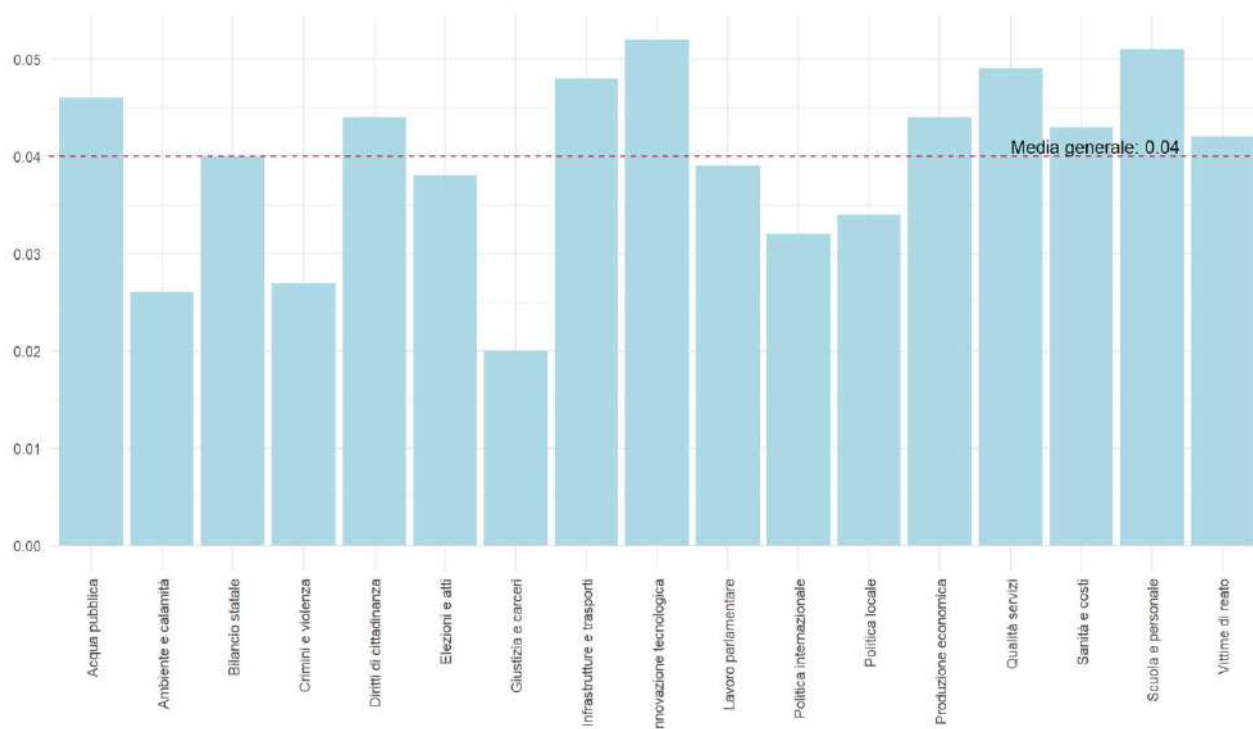


Figura 5: Grafico a barre del sentiment medio dei discorsi per ogni cluster, ottenuto assegnando ogni discorso esclusivamente al cluster che presenta la probabilità maggiore. La linea tratteggiata rappresenta il sentiment medio.

Per la successiva analisi ci si occupa dello studio del sentiment sui dibattiti che riguardano il cluster con argomento “scuola e personale pubblico”. Vengono presi i discorsi che presentano la maggiore probabilità di appartenere a questo cluster. Si analizzano i sentiment per partito e vengono esclusi

dall’analisi i partiti che hanno tenuto meno di 100 discussioni parlamentari a riguardo. I discorsi parlamentari analizzati in questa fase sono 2510. Come si evince in Tab. 1, Il Partito Democratico è il partito più rappresentato con 704 interventi, seguito da Forza Italia – PDL e Movimento 5 Stelle, entrambi superiori a 500 discorsi. Scelta civica con 136 discorsi è il partito meno rappresentato ma anche quello col sentiment medio più alto, pari a +0.0566. La scelta di prendere solamente i partiti ben rappresentati è necessaria per rendere l’analisi leggibile e per poter fare dei confronti partito vs partito.

Tabella 1: statistiche descrittive del sentiment per partito sul tema “scuola e personale pubblico”.

Nome del partito	N	Media	Mediana	SD
Forza Italia - PDL	544	0.0555	0.0554	0.0409
Fratelli d’Italia	191	0.0481	0.0477	0.0300
Italia dei Valori	186	0.0459	0.0473	0.0391
Lega	239	0.0515	0.0464	0.0427
Movimento 5 Stelle	510	0.0474	0.0484	0.0409
Partito Democratico	704	0.0524	0.0511	0.0387
Scelta Civica	136	0.0566	0.0541	0.0362

Il test ANOVA di uguaglianza del sentiment medio tra i diversi partiti ha un p-value inferiore a 2.2e-16. Il corrispondente test non-parametrico di Krusal-Wallis (1952) invece riporta un p-value di 0.0007, dunque estremamente significativo. In seguito, con il test di Dunn (1961) si cercano nei confronti una ad una le coppie di partiti che abbiano una media significativamente diversa del sentiment. Due coppie significative risultano Forza Italia – PDL vs Movimento 5 Stelle (p-value = 0.0051) e Movimento 5 Stelle vs Scelta Civica (p-value = 0.02641). In particolare, dall’analisi svolta risulta che il sentiment riguardo all’argomento “scuola e personale pubblico” sia significativamente più negativo per il Movimento 5 Stelle sia rispetto a Forza Italia – PDL, sia rispetto al partito Scelta Civica.

5. Conclusioni

In questo elaborato, abbiamo effettuato una ricognizione dei temi più ricorrenti dei discorsi parlamentari italiani raccolti in *ItaParlCorpus* di Cova (2024) relativi alle legislature XV-XVIII tramite un’analisi LDA, una cluster analysis, e una analisi del sentiment basata sul dizionario Sentix con riferimento ai gruppi di argomenti e ai partiti politici. L’applicazione dei modelli LDA (Latent Dirichlet Allocation) ha mostrato una buona efficacia nel rappresentare gli argomenti affrontati dai dibattiti politici e le parole più rilevanti che li compongono lungo le diverse legislature. Come illustrato nelle analisi, la modellazione LDA consente una visione rapida e significativa dei temi centrali del dibattito politico italiano nel corso del periodo considerato. L’analisi del sentiment su dati parlamentari mostra aspetti interessanti che possono aiutare a capire gli argomenti che vengono discussi con toni più positivi o più negativi. In particolare, l’analisi riguardo al macro-argomento “scuola e personale pubblico” ha mostrato che deputati di partiti diversi possono usare un tono sistematicamente diverso (positivo o negativo) su un dato argomento, il che può denotare una specifica strategia di posizionamento politico sul tema.

Bibliografia

Basile, V. (2019). sentixR: R package to perform sentiment analysis on Italian using sentix. GitHub.

Basile, V., e Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 100-107.

Blei, D. M., Ng, A. Y., e Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol 3: 993–1022.

- Cova, J. (2024). ItaParlCorpus, <https://doi.org/10.7910/DVN/KUARWD>, Harvard Dataverse, V3.
- Chen, J., Li, K., Zhu, J., e Chen, W. (2015). Warplda: a cache efficient $o(1)$ algorithm for latent dirichlet allocation. *arXiv preprint arXiv:1510.08628*
- Chen, T., e Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.
- Döring, Holger, e Philip Manow (2024). Parliaments and Governments Database (ParlGov): Release 2024. Harvard Dataverse, V1. — DOI: [10.7910/dvn/2vz5zc](https://doi.org/10.7910/dvn/2vz5zc)
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, vol. 56: 52–64.
- Grimmer, J., e Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, vol. 21(3): 267-297.
- Kruskal, W., e Wallis, A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, vol. 47(260): 583–621.
- Mimno, D., Wallach, H., Talley, E., Leenders, M., e McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 262-272.
- Nivre, J., De Marneffe, M. C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., ... e Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.
- Salton, G., e Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing and management*, vol. 24(5), 513-523.
- Selivanov, D., Bickel, M., e Wang, Q. (2023). text2vec: Modern text mining framework for R (Version 0.6.4). <https://CRAN.R-project.org/package=text2vec>.
- Straka, M., e Straková, J. (2019). Universal Dependencies 2.5 Models for UDPipe (2019-12-06) model: “italian-vit-ud-2.5-191206.udpipe”, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). <https://lindat.mff.cuni.cz/repository/items/41f05304-629f-4313-b9cf-9eeb0a2ca7c6>

A Multilevel Analysis of Italy's Three-Year School Planning Documents Using a Lexical and Distributional-Semantic Framework

Pasquale Pavone¹, Maria Teresa Sagri², Elettra Morini², Eniko Tolvay²

¹Department of Literary, Linguistic and Philosophical Studies, Pegaso University –
pasquale.pavone@pegaso.it

²INDIRE – Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa –
{t.sagri, e.morini, e.tolvay}@indire.it

Abstract

The study proposes a multilevel methodological approach for analyzing Piani Triennali dell'Offerta Formativa (PTOF), the official documents through which Italian schools describe their educational projects.

The framework combines two levels of analysis:

Measurement level: texts from PTOF documents are represented through sentence embeddings generated by a BERT model. These representations are then processed using UMAP (dimensionality reduction) and HDBSCAN (clustering) to identify thematically coherent groups in a fully replicable manner.

Interpretative level: text mining and lexicometric techniques are used to characterize the emerging clusters. Specifically, keyness analysis (based on chi-square statistics) is applied to reconstruct the lexicon specific to each cluster, while co-occurrence network analysis is proposed as a subsequent analytical step for exploring the internal structure of selected semantic fields.

The integration of semantic clustering and lexical analysis provides a structured representation of the corpus and lays the groundwork for future analytical outputs, such as interactive dashboards and knowledge graphs, combining distributional semantics and classical text statistics to deliver scalable, interpretable, and methodologically transparent analyses of large institutional text corpora.

Keywords: Distributional Semantics, Sentence Embeddings, UMAP, HDBSCAN, Text Mining, Lexicometry, Three-Year School Planning Documents (PTOF), Large Institutional Text Corpora.

1. Introduction

In the Italian school system, the international dimension has taken on a central role in curriculum construction, driving significant changes across the socio-educational context. In an increasingly globalized scenario, schools face a growing need to innovate their educational provision with a European outlook, acting as key agents of change in the construction of a "European Education Area" (Tarozzi, 2021). Within the Italian system, school autonomy (DPR n.275, 1999) can serve as the operational lever through which European policies are translated into responses consistent with local contexts (Morzenti Pellegrini, 2006; Mulè et al., 2020). Despite this potential, the internationalization of educational offer still appears confined to single project-based interventions, lacking a systemic vision. The literature emphasizes that innovation must take the form of a structural transformation affecting the institution's entire governance and pedagogical-organizational framework. This requires the co-construction of a strategic vision that, through the synergy of funding sources, ensures organizational cohesion and staff motivation, as well as systematic curricular integration. From this perspective, the

Three-Year Educational Offer Plan (PTOF) becomes a key instrument for integrating institutional identity into everyday practices. An evidence-based, data-driven approach — fully integrated with the Learning Organization paradigm (Kools & Stoll, 2016) — ensures that internationalization is a genuine process of improving educational quality rather than a mere label. To support schools along this path, it is essential to implement an integrated Knowledge Management (KM) architecture, capable of transforming a heterogeneous informational heritage into operational knowledge and supporting the transition toward evidence-based decision-making models. Among school documents, the PTOF is the instrument through which institutions communicate their identity, mission, and project planning to all stakeholders. New technological paradigms — particularly NLP techniques applied to PTOF analysis — make it possible to identify latent needs, explicitly stated priorities, and to conduct longitudinal analyses monitoring the evolution of educational provision.

1.1. The analysis of the Three-Year Educational Offer Plans of the national school network

This research action is part of a broader project promoted by INDIRE, aimed at creating a permanent observatory on the internationalization of educational offer. Through an incremental integration of qualitative-quantitative and unstructured data, a geolocated data warehouse was created to reconstruct the project-based behaviors promoted by schools and cross-reference them with PTOF information. The research aims to outline a map of intervention strategies for school model transformation, interpreting processes triggered by the research, training, and experimentation activities promoted by Erasmus+ Agency and INDIRE.

With note no. 17832 of 16 October 2018, the Ministry of Education introduced the possibility for schools to draft the PTOF using the online format available on the SIDI platform. The PTOF drafted in this way is then made public on the school's website and in the specific section of the Scuola in Chiaro portal. The online PTOF is organized into sections and subsections and provides the following structure:

- School and context: describes the school's identity through an analysis of the local area and its needs, including the school's main characteristics, facilities and infrastructure, and a mapping of professional resources.
- Strategic choices: defines priorities and development lines based on the RAV and Law 107/15, including the Improvement Plan, innovation measures, and PNRR actions (Mission 1.4) to tackle early school leaving.
- Educational offer: outlines the core instructional components (curriculum, teaching activities, timetables, expected learning outcomes/PCTO), inclusion measures, digital education (PNSD/DDI), assessment, and initiatives for expansion and ecological/cultural transition.
- Organization: presents the school's operational setup (organizational model, offices and user-facing arrangements), external networks and agreements, and training plans for both teaching and administrative staff (ATA).

1.1.1 Research questions

Consistent with the issues outlined above, the analysis of the PTOF corpus of Italian schools for the 2022–2025 three-year period aims to answer the following research questions:

- What are the project-based behaviors of educational institutions with specific reference to the internationalization of educational offer? What distinctive language do schools use when addressing these themes?

- In what ways do schools implement pathways of improvement and innovation to intervene in internationalization processes (in terms of teaching activities, staff training, collaborations established, etc.)?
- Is it possible to identify recurring characteristics in the planning of schools that share similar features (type, size, geographical location, socio-economic-cultural characteristics of the context in which they are embedded)?
- Is it possible to identify linguistic indicators capable of mapping school contexts in which internationalization still appears as an episodic phenomenon confined to individual projects, distinguishing them from those institutions in which it is systematically integrated into educational offer?

2. Data and Methods

The empirical analysis is based on a corpus of 8,423 PTOFs from 19 Italian regions, excluding the Trentino–Alto Adige region, with an overall size of approximately 209 million tokens. Each document is structured into 113 sections grouped into four macro-areas: School and context, Strategic choices, Educational offer, and Organization. The analysis was conceived as an automated content analysis procedure aimed at identifying latent semantic structures within a large institutional corpus that is typologically homogeneous but heterogeneous in terms of contents, educational priorities, and school contexts. In this respect, the study is grounded in the tradition of statistical text analysis and lexicometry, which assumes that textual content can be investigated through metric representations of language and replicable procedures of classification and interpretation (Bolasco, 2013; Lebart et al., 1998; Tuzzi, 2024). At the same time, it extends this tradition by integrating more recent distributional approaches to meaning to address the semantic dimension of texts on a large scale. The adopted pipeline therefore combines sentence embeddings derived from BERT-based models, dimensionality reduction, cluster analysis, and lexical-textual tools for the interpretation of the identified groups. In order to obtain a detailed semantic classification, the methodology was applied separately to each section of the documents, with particular attention to the descriptive sections containing school-specific content. Within each section, paragraphs were treated as the basic textual units of analysis, since they are sufficiently informative from a semantic point of view while remaining analytically manageable in a large-scale automatic workflow. The following section presents the results of the analysis of the section on activities for improvement plans within the strategic choices of schools.

The methodological pipeline consists of four main steps: semantic representation of paragraphs through sentence embeddings, dimensionality reduction of the resulting vectors, corpus segmentation by means of cluster analysis, and interpretation of the resulting groups through lexical-textual analysis of their characteristic language. In the first step, each paragraph is transformed into a dense vector representation. The study uses sentence embeddings derived from the multilingual Sentence-Transformers model `sentence-transformers/paraphrase-multilingual-mpnet-base-v2`, run locally, which produces a 768-dimensional vector for each paragraph. More specifically, the study adopts a BERTopic-style pipeline in its use of UMAP and HDBSCAN, but not the BERTopic framework itself, since cluster interpretation is based on keyness analysis rather than on cTF-IDF-based topic representation. This choice is not only methodologically appropriate for similarity and clustering tasks, but also crucial in terms of reproducibility and data governance. While more recent proprietary API-based embedding models may provide strong performance, they introduce at least two critical issues in

institutional research settings: first, they may hinder long-term reproducibility because model updates and inference conditions are controlled externally by the provider; second, they require transmitting textual data to private external services. By contrast, the use of a locally executable model with publicly available weights makes it possible to fix versions, dependencies, and processing conditions, and to replicate the workflow end-to-end without access constraints, API-related costs, or the need to share institutional textual data with private companies, even when the documents themselves are publicly accessible. From a theoretical point of view, this choice is grounded in the Transformer architecture introduced by Vaswani et al. (2017) and in BERT, proposed by Devlin et al., (2019), which learns contextual representations of language through bidirectional pre-training on large corpora. More specifically, the procedure follows the perspective of Sentence-BERT, developed by Reimers & Gurevych, (2019), which is explicitly designed to make sentences and paragraphs directly comparable in vector space and is therefore particularly well suited to semantic similarity, clustering, and semantic search tasks. The output of this phase is a dense matrix in which each paragraph is represented as a point in the semantic embedding space, and the distance between vectors can be interpreted as a measure of semantic similarity between textual contents. Since these representations are characterized by relatively high dimensionality, a dimensionality reduction step was subsequently performed using UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018). UMAP is a manifold learning technique designed to project high-dimensional data into a lower-dimensional space while preserving the neighborhood structure among points. Unlike linear methods such as Principal Component Analysis, UMAP can capture non-linear relationships among observations and preserve both local and, to a significant extent, global data structure. In the present study, dimensionality reduction serves a double purpose: on the one hand, it produces a two-dimensional representation useful for the exploratory visualization of the semantic distribution of paragraphs; on the other, it provides a more compact representation of the data to be used as input for the clustering procedure. Cosine similarity was adopted as the distance metric, consistently with the nature of the distributional representations obtained in the previous step, and a fixed random state was used in order to strengthen the reproducibility of the procedure. For reproducibility, UMAP was run with 10 components, 15 neighbors, 0.05 minimum distance, cosine metric, and fixed random state (42), while the 2D visualization used the same UMAP configuration with two components. After identifying one extreme outlier in the 2D projection, the corresponding observation was removed and the 2D visualization was recomputed, while clustering was performed on the outlier-filtered reduced space. The thematic segmentation of the corpus was then carried out by means of HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) applied to the reduced vector representations (Campello et al., 2013, 2015). Unlike agglomerative hierarchical clustering, HDBSCAN does not impose a fixed number of groups in advance, but identifies clusters as areas of higher density in the semantic space, while allowing observations that do not belong stably to any dense region to remain unassigned as noise. This property is particularly useful in a corpus such as PTOFs, where many paragraphs are highly similar or formulaic, while others are more weakly structured and do not form clearly bounded thematic groups. In the present study, clustering was therefore based on the local density structure of the reduced semantic representations rather than on successive pairwise mergers of all observations. The resulting clusters do not derive from a manual classification of texts, but from an automatic procedure that groups paragraphs according to their position in vector space and their density-based neighborhood relations. Clustering thus represents the stage at which the semantic measurement produced by the embeddings is transformed into an interpretable analytical structure. At the same time, the use of a density-based method makes it possible to preserve a clear distinction

between the measurement phase, grounded in a replicable semantic representation and proximity structure, and the interpretative phase, which is developed afterwards through lexical-textual tools. The clustering step was implemented with HDBSCAN using a minimum cluster size of 80, a minimum samples value of 5, Euclidean distance on the reduced UMAP space, the “eom” cluster selection method, and with single-cluster solutions not allowed.

Once the thematic clusters had been identified, their interpretation relied on the reconstruction of the characteristic language of each group through a lexical-textual procedure combining multiword expression detection (Pavone, 2018) and keyness analysis (Gabrielatos, 2018; Scott & Tribble, 2007). This step is particularly important in a specialized corpus such as PTOFs, where meaning is often conveyed by relatively stable nominal sequences, such as *offerta formativa* (educational offer), *piano di miglioramento* (improvement plan), *competenze chiave* (key competences), *inclusione scolastica* (school inclusion), or *curricolo verticale* (vertical curriculum). Treating these expressions as autonomous lexical units makes it possible to build a more precise and interpretable domain-specific terminology, reducing the ambiguity of isolated words. On this basis, a selection of content terms, including both single words and multiword units, was used to construct a paragraph-by-term matrix, later aggregated into a cluster-by-term matrix for keyness analysis. For each term, the observed frequency within a cluster was compared with its expected frequency in the corpus, and a chi-square-based keyness measure was calculated. Compared with TF-IDF, this approach is better suited to reconstructing the distinctive lexicon of already identified groups of texts, since it highlights terms that are significantly overused in a given cluster relative to the corpus as a whole. For lexical characterization, only terms with positive deviation and a standardized residual greater than 2.5 were retained. The result is a characteristic dictionary for each cluster, providing a statistically grounded basis for semantic interpretation rather than relying on intuitive labeling alone.

3. Main Results

The adopted procedure made it possible to identify the main themes discussed in the section devoted to the actions included in the Improvement Plans. Through the representative-selection procedure, 23,592 paragraphs were retained for semantic embedding. After the removal of one extreme outlier, HDBSCAN was applied to 23,591 paragraphs. Of these, 11,644 (49.4%) were left unassigned as noise, while the remaining 11,947 paragraphs (50.6%) were assigned to density-based clusters. In the interpretative phase, the analysis focused on 36 substantively meaningful thematic clusters, derived through characteristic-language profiling based on keyness measures. The relatively high proportion of noise suggests that the section includes a substantial number of weakly structured, generic, or only loosely connected paragraphs alongside a more clearly organized semantic core. Rather than forcing all observations into clusters, the procedure retained only those textual units that formed sufficiently stable density-based groupings in semantic space.

1. Digital innovation and technological mediation – Focus on digital transition, educational technologies, computational thinking, and the organizational use of digital tools in teaching.
2. Basic skills reinforcement in core subjects – Emphasis on strengthening foundational competences in Italian, mathematics, and English, especially reading, writing, and text comprehension.

3. Foreign languages, certification, and internationalization – A thematic area centred on multilingual development, language certification, mobility, and the international opening of the school.
4. Mathematics education and scientific skills – Focus on mathematics teaching, logical-numerical competences, and subject-specific planning in the scientific domain.
5. Social vulnerability and global citizenship themes – A heterogeneous cluster oriented toward wellbeing, fragility, solidarity, risk, and socially relevant global issues.
6. Inclusion and special educational needs – Strong emphasis on disability, BES, support measures, screening, and inclusive educational practices.
7. Internal governance and strategic school roles – A formal organizational register centred on functions, commissions, protocols, and internal coordination roles.
8. Teacher professional development – Focus on staff training, professional competences, reflective practice, and the enhancement of teaching expertise.
9. Bullying prevention and relational safety – A clearly defined area concerning bullying, cyberbullying, prevention, and the promotion of safe school relationships.
10. Final outcomes and performance monitoring – Focus on final assessments, suspensions, examination outcomes, and the monitoring of student performance.
11. Environmental education and sustainability – Thematic emphasis on sustainability, ecological transition, recycling, water, and environmental awareness.
12. Monitoring, evaluation, and internal review – A cluster oriented toward assessment, evidence gathering, school self-evaluation, and internal quality monitoring.
13. Civic education and citizenship learning – Focus on civic education as a formal curricular field linked to citizenship, constitution, and civic competences.
14. Interdisciplinary planning and authentic tasks – Emphasis on integrated curriculum design, cross-disciplinary planning, and competence-based activities such as real-world tasks.
15. Laboratory-based learning and practical settings – A cluster centred on workshops, laboratories, and the use of practical environments for learning.
16. Common assessment and evaluation standardization – Focus on common tests, parallel classes, shared grids, and practices aimed at comparability of results.
17. Remedial support and learning recovery – A thematic area concerning learning gaps, insufficient achievement, remedial courses, and academic realignment.
18. Methodological innovation in teaching – Emphasis on innovative teaching methods, pedagogical renewal, and staff training for methodological change.
19. Legality, rights, and responsible citizenship – A value-oriented civic area focused on legality, democracy, rights, duties, and socially responsible behaviour.
20. Artistic, theatrical, and audiovisual expression – Focus on expressive languages, including art, theatre, cinema, dance, and cultural heritage.
21. Operational coordination of projects and activities – A managerial and organizational cluster centred on project implementation, responsibilities, and operational steering.
22. Music education and performance activities – A distinct thematic area devoted to music, choir, instruments, orchestral practice, and performance-based learning.
23. Science education and STEM orientation – Focus on scientific disciplines and, in part, on orientation toward STEM-related academic and professional pathways.
24. European key competences and competence certification – A cluster framed around European competences, curricular structuring, and the assessment and certification of competences.

25. Vertical curriculum and school continuity – Emphasis on continuity across school levels through vertical curriculum design and coordination between segments.
26. Learning environments, spaces, and equipment – Focus on the physical transformation of educational settings through spaces, furniture, equipment, and innovative environments.
27. INVALSI preparation and standardized testing – A strongly assessment-oriented cluster centred on INVALSI, simulations, test preparation, and the management of standardized evaluation.
28. Transition, continuity, and educational guidance – Focus on transition points between school levels, student placement, continuity projects, and orientation in and out of school cycles.
29. Guidance and student trajectory monitoring – A more analytical cluster concerned with follow-up, student pathways, guidance processes, and outcome tracking.
30. Family involvement and educational co-responsibility – Emphasis on school-family relations, parental participation, communication, and shared educational responsibility.
31. Sport, physical activity, and wellbeing – A thematic area devoted to physical education, sport practice, movement, and bodily wellbeing.
32. School dropout prevention and persistent fragility – Focus on absenteeism, dropout risk, educational vulnerability, and measures for re-engaging fragile students.
33. Psychological support and emotional wellbeing – A cluster specifically centred on counselling, listening services, emotional support, and student wellbeing.
34. Territorial networks and external partnerships – Emphasis on collaboration with local actors, associations, public bodies, and wider territorial networks.
35. School leadership and institutional coordination – A cluster oriented toward headship, collegial bodies, strategic coordination, and institutional management.
36. Cross-curricular reinforcement of key competences – A broader area combining linguistic and logical-mathematical strengthening, framed as transversal support for fundamental competences.

From the analysis of the thematic clusters, internationalization does not emerge as a uniformly distributed dimension across the corpus, but rather as a semantically recognizable theme concentrated in a limited number of clusters. Its most explicit manifestation is found in Cluster 3, whose characteristic lexicon includes terms such as internationalization, mobility, abroad, international, language certifications, native speaker, foreign languages, and specific references to English, French, Spanish, and German. In this cluster, internationalization is framed directly as a field of educational planning connected to multilingual development, certification pathways, student mobility, and the broader opening of the school toward a European and transnational educational space. Alongside this main cluster, the same semantic area reappears in other groups, although in more indirect or partial forms.

Cluster 2, while primarily oriented toward the reinforcement of basic competences, also contains a relevant linguistic component, with terms related to English, foreign languages, and students or families with a foreign background. Cluster 24 reflects a more institutional and curricular articulation of the same theme, through expressions such as European competences, key competences, European citizenship, and competence certification. A further, though weaker, signal emerges in Cluster 36, where multilingualism and language certification are embedded within a broader cross-curricular logic of competence enhancement. Taken together, these results suggest that internationalization is present in the analysed PTOFs as a clearly

identifiable topic, but not as a dominant or fully transversal organizing principle of the educational offer. Rather, it tends to be concentrated in a restricted set of planning domains, especially those related to foreign-language teaching, language certification, mobility opportunities, cultural exchange, and, more indirectly, the adoption of European competence frameworks. Beyond the descriptive list of clusters, the results point to three broader patterns. First, Improvement Plans are strongly organized around a limited set of recurring strategic domains, especially digital innovation, inclusion, evaluation, teacher training, and competence reinforcement. Second, internationalization emerges as a selective rather than transversal planning dimension, concentrated in a small number of semantically cohesive clusters rather than distributed across the corpus as a whole. Third, the semantic mapping produced by the pipeline has practical relevance for institutional monitoring, since it makes it possible to compare schools according to the thematic configuration of their planning discourse and to identify areas of strategic concentration, relative absence, or thematic specialization.

4. Conclusions and future developments

The proposed pipeline shows that large-scale semantic analysis can be effectively applied to PTOF documents to reconstruct the main thematic structures underlying school planning discourse. By combining locally computed sentence embeddings, dimensionality reduction, density-based clustering, and lexical-textual interpretation, the study makes it possible to move from a large and formally homogeneous institutional corpus to an empirically grounded thematic typology of the actions described in Improvement Plans. In this respect, the results confirm that PTOFs can be treated not only as administrative documents, but also as a rich textual source through which schools represent priorities, interventions, and educational orientations. Substantively, the analysis highlights a broad and articulated thematic landscape, in which issues such as digital innovation, inclusion, assessment, teacher training, civic education, sustainability, and family involvement occupy a central position. At the same time, the findings show that some themes are more concentrated and selective than others. This is particularly evident in the case of internationalization, which emerges as a recognizable topic but not as a fully transversal organizing principle of the educational offer. Rather, it appears mainly in relation to foreign-language teaching, certification pathways, mobility, and references to European competence frameworks. This suggests that, within the analysed PTOFs, internationalization is present more as a specific line of planning than as a pervasive dimension structuring school discourse as a whole. From a methodological point of view, the clustering results also reflect the heterogeneous and partly formulaic character of the corpus, confirming the suitability of a density-based approach for preserving only the most stable semantic groupings. Further developments may extend the present approach in several directions. First, selected clusters can be explored through co-occurrence-based network analysis to examine their internal semantic organization more closely and identify recurrent lexical communities. Second, the resulting outputs can be integrated into a multilayer classified textual database in which sections, paragraphs, lexical units, multiword expressions, and thematic labels are linked and made queryable at different levels of granularity. Such a resource could support comparative analyses across sections, school types, and territorial contexts, as well as the development of interactive dashboards for corpus exploration. In a further step, the same infrastructure could provide the basis for the construction of knowledge graphs connecting concepts, actors, and actions extracted from PTOF discourse. Overall, the study should therefore be seen not only as an application to a specific corpus section, but also as a first step toward a broader analytical environment for the systematic exploration of school planning texts.

References

- Bolasco, S. (2013). *L'analisi automatica dei testi: Fare ricerca con il text mining*. Carocci.
- Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172.
- Campello, R. J., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1), 1–51.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.
- DPR n.275 (Decreto Del Presidente Della Repubblica No. 275). (1999). Gazzetta Ufficiale.
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In *Corpus approaches to discourse* (pp. 225–258). Routledge.
- Kools, M., & Stoll, L. (2016). *What makes a school a learning organisation? OECD Education Working Papers, No. 137*, OECD Publishing, Paris.
- Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data*. Springer.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv Preprint arXiv:1802.03426*.
- Morzenti Pellegrini, R. (2006). *L'autonomia scolastica tra sussidiarietà, differenziazioni e pluralismi*. Giappichelli.
- Mulè, P., De Luca, C., & Notti, A. M. (2020). *L'insegnante e il dirigente scolastico nella scuola dell'autonomia tra didattica, governance e progetto culturale*.
- Pavone, P. (2018). Automatic Multiword Identification in a Specialist Corpus. In A. Tuzzi (Ed.), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences* (pp. 151–166). Springer International Publishing. https://doi.org/10.1007/978-3-319-97064-6_8
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv Preprint arXiv:1908.10084*.
- Scott, M., & Tribble, C. (2007). Textual patterns: Key words and corpus analysis in language education. *TESL-EJ*, 11(2).
- Tarozzi, M. (2021). Educating for global citizenship in diverse and unequal societies. In *Conversations on Global Citizenship Education* (pp. 89–102). Routledge.
- Tuzzi, A. (2024). *Fondamenti di analisi dei dati testuali*. Carocci editore.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

The Lexical Construction of Women’s ‘Image’: Party Specificity and Gender Asymmetry in Italy’s Constituent Assembly

Aulino Francesca¹, D’Agata Rosario², De Stefano Domenico³, Fabbrucci
Barbagli Amin Gino⁴

¹University of Catania – francesca.aulino@gmail.com

²University of Catania – rosario.dagata@unict.it

³University of Trieste – ddestefano@units.it

⁴University of Trieste – amingino.fabbruccibarbagli@phd.units.it

Abstract (in English)

This study is organized in two parts. In the first part, we applied traditional methodologies of text analysis to the records of the Italian Constituent Assembly (1946–1947). Specifically, we focused on the representation of women emerging from the Assembly debates that shaped the 1948 Italian Constitution. The *corpus* was constructed from the stenographic reports of the Assembly’s plenary sessions and the work of the Sub-committees, selecting all interventions in which women are explicitly mentioned. Texts were organized into *sub-corpora* according to the speakers’ political affiliation, allowing for a comparative analysis of the linguistic strategies employed by the different parliamentary forces. In the second part, after a first exploratory analysis of the data, we apply network analysis. This methodology allows to study the evolution of the co-occurrence network, to examine how different forces forged and constructed identity and roles, and to map the semantic distance between them and the main concepts in each party’s narrative. The study aims to bring to light lexical and semantic differences in the discursive construction of the female issue as adopted by the main political alignments active in the debate. The analysis reveals significant semantic divergences within an otherwise largely homogeneous stylistic framework. Overall, the study shows how the women’s issue constitutes a site of discursive polarization. While Christian Democracy tends to use a lexicon anchored in the family sphere and traditional roles, the socialist party appears more related to the issues of emancipation, labour and social rights. The Communist Party also appears sensitive to issues of work and social justice, but unlike the Socialist Party, it also introduces elements related to the traditional family into its lexicon.

Keywords: Italian Constituent Assembly; Women; Text Analysis; Network Analysis

1. Introduction

On 25 June 1946, the Constituent Assembly charged with drafting the Constitutional Charter of the Italian Republic officially convened. The Assembly’s proceedings continued until 31 January 1948. The Constituent Assembly (elected 1946) comprised 556 members in total: 535 (96.2%) men and 21 (3.8%) women. Moreover, within the 75-member Committee, tasked with drafting the Constitution’s working text, women were only five. Despite their small number, these women made a decisive contribution to the drafting of the Constitutional Charter and, consequently, to Italy’s modernization process Bozzoni (2025). The most significant contributions made by the “Founding Mothers” were not limited to areas traditionally associated with women, such as family, maternity and welfare, but also extended to domains like the labour market and, more broadly, the political provisions of the Constitutional Charter D’Amico (2023).

This study analyses the content of the debates of the Italian Constituent Assembly, seeking to bring out the image of women held by the constituent mothers and fathers, as well as the contribution women made to the drafting of the Constitution. From a methodological point of view, this initial exploratory phase relies on lexical specificity analysis and on the use of a semantic network based on co-occurrences.

2. Data

The Constituent Assembly officially began its proceedings on 25 June 1946. The political framework for its delegates was established by the Ministry of the Constituent Assembly and the National Consultative Assembly (Consulta Nazionale), both of which were founded during the brief government of Ferruccio Parri (June–December 1945). Under the leadership of important figures such as Pietro Nenni and Meuccio Ruini, these organizations were fundamental in drafting electoral laws and conducting essential preliminary studies for Italy's new Constitution. Their work was vital in shaping a democratic framework, ensuring that the key principles of governance were established to guide the country's future political system. The work of the Commission for the Constituent Assembly officially began on 23 July 1946 and ended on 1 February 1947, the actual date of cessation of the sessions. The Constitution of the Italian Republic was approved by 453 votes in favour and 62 against during the afternoon session of December 22, 1947. On 27 December, the new Italian Constitution was promulgated by the provisional Head of State, Enrico de Nicola, and on 1 January 1948, it officially came into force.

The shorthand reports of the work of the Constituent Assembly and the summary records of the Commission of the 75 and three Subcommissions constitute a historical and cultural heritage of great value. The legacy of the debate allows us to clearly grasp the spirit and intentions of the constituent fathers and mothers, as well as the values through which they intended to shape the new institutional framework and define the identity of the nascent Italian Republic. The richness and plurality of these documents lend themselves effectively to a textual analysis, capable of highlighting themes, meanings, and orientations that will be the subject of this study's investigation, with particular reference to the role of women. For the aim of this work the digital transcripts of each session were obtained through the open access platform *Come nasce la Costituzione*¹, which offers the possibility of carrying out full-text searches of keywords and viewing the list of all sessions relevant to the criteria set up.

In total, the website contains 621 documents, of which 183 contain at least one occurrence of the forms *donna* and/or *donne*. However, this initial lexical filtering was not sufficient to define the analytical *corpus*, since not all the selected sessions deal centrally with the issue of women. In several cases, the forms *donna* and *donne* occur only as isolated, marginal or incidental references within broader discussions that are not directly concerned with the role, condition or rights of women. For this reason, a two-step manual selection was carried out. First, the sessions in which the reference to women was embedded in a substantial and argumentatively relevant discussion were selected. Second, within these sessions, only the interventions that

¹ *Come nasce la Costituzione* (How the Constitution was born). Accessed March 17, 2026 by <https://comenascelacostituzione.it/?p=5082> "This work, entitled "How the Constitution is Born", was conceived by its editors in view of the 75° anniversary of the Italian Constitution, which came into force on January 1°, 1948. The work consists of an open access portal that reproduces those same texts in a digital format in which it is possible to perform full-text searches with keywords and a navigation tree, as well as copy and paste and print operations".

were actually pertinent to the topic were extracted. In particular, the selected interventions are those in which the figure of women is discussed in relation to issues such as family, marriage, motherhood, work, civil rights, legal equality and access to public offices. Conversely, cases in which the forms *donna* and/or *donne* appeared as isolated occurrences, without being developed in the debate, were excluded.

Through this procedure, a targeted *corpus* was obtained, consisting exclusively of the actual contributions of the Constituents concerning the role and condition of women. The final dataset is composed of 42 documents, from which only the interventions related to the topic were extracted. Since the selection was conducted by a single annotator, no inter-annotator agreement score was calculated. In order to reduce the margin of arbitrariness and to make the procedure more transparent and replicable, the inclusion and exclusion criteria were made explicit, and the reduction process from the initial archive to the final analytical corpus is documented in Figure 1.

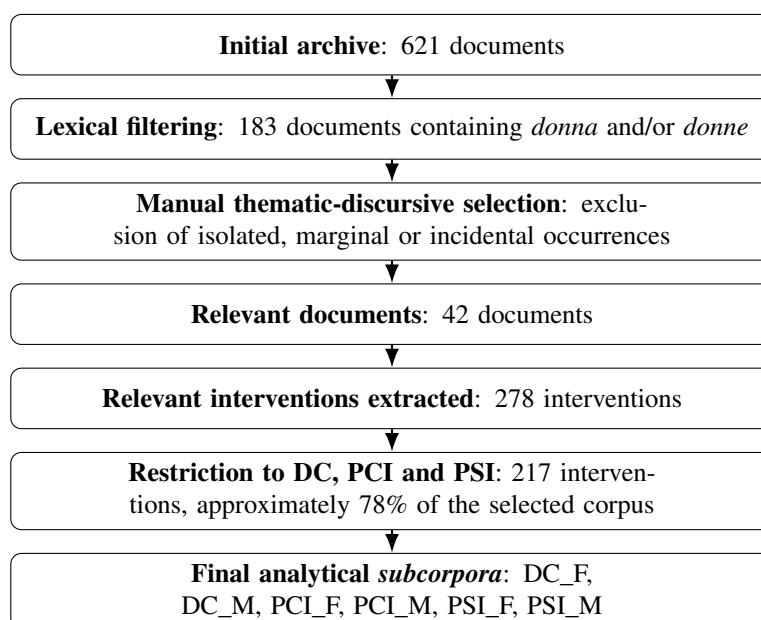


Figure 1: Flowchart of the corpus construction and reduction process.

Subsequently, a second phase of pre-treatment of the *corpus* was necessary in order to prepare it for the reading, segmentation and coding requirements of the software. In particular, asterisks were manually inserted as markers of the proper names of the Constituents, written in capital letters and placed at the end of the line, so as to delimit the speeches of each speaker. An example of the resulting syntactic structure is shown in Figure 2.

The final operation on the text consisted in building distinct *subcorpora* by aggregating the interventions of constituents belonging to the same party. The manually selected corpus consists of 278 interventions. For the purposes of the comparative analyses, this study focuses on the three main parties represented in the corpus, namely Democrazia Cristiana (Christian Democracy) (DC), Partito Comunista Italiano (Italian Communist Party) (PCI) and Partito Socialista Italiano (Italian Socialist Party) (PSI). Together, these three parties account for 217 interventions, corresponding to approximately 78% of the selected corpus (see Table 1). The remaining 61 interventions, corresponding to approximately 22%, are distributed across five minor parties and were excluded from the party-based comparison because their *subcorpora* would have been too small and fragmented to support a reliable textual analysis.

NOCE TERESA Correlatrice, rileva che la Sottocommissione è d'accordo sui concetti delle proposte; vi è solo divergenza sulla formula da adottare. Per esempio, il Presidente, nella sua proposta, ha usato il termine «gestanti» che si riferisce ad un periodo di tempo troppo limitato, contro l'intenzione stessa del proponente. Preferirebbe il termine «donna», più lato del termine «gestante».

TAVIANI propone l'adozione del termine «madre».

NOCE TERESA Correlatrice, preferirebbe la dizione «condizioni di procreazione», in quanto il termine «madre» è attributo della donna, anche quando ha una prole già maggiorenne. Pur tuttavia ritiene migliore il termine «donna», col quale si viene a specificare che i mezzi di assistenza non devono essere limitati al solo periodo della gestazione. Invece di «umane», desidererebbe una dizione più precisa, come, per esempio, «buone condizioni economiche, igieniche e sanitarie».

Figure 2: Extract of a session.

The resulting analytical corpus for the party- and gender-based comparisons therefore consists of the interventions belonging to DC, PCI and PSI. These interventions were subsequently divided according to the gender of the speaker, resulting in six subcorpora: DC_F, DC_M, PCI_F, PCI_M, PSI_F and PSI_M.

Table 1: Distribution of the 278 selected speeches by party.

Acronym	Party	Speeches	%
DC	Democrazia Cristiana	131	47%
PCI	Partito Comunista Italiano	51	18%
PSI	Partito Socialista Italiano	35	13%
PLI-UDN	Partito Liberale Italiano e Unione Democratica Nazionale	23	8%
UQ	Fronte dell'Uomo Qualunque	18	6%
PRI	Partito Repubblicano Italiano	11	4%
DL	Democrazia del Lavoro	5	2%
PdA	Partito d'Azione	4	1%
Total number of speeches		278	100%

3. Methodology

In this work, we adopted several steps to analyse the *corpus*. The *corpus* was divided into *subcorpora* based on the analytical objective. For gender comparison, the texts were aggregated into two sets: a female *subcorpus* and a male *subcorpus*, respectively obtained by concatenating the interventions of the constituent mothers and the constituent fathers who speak specifically about women. For cross-party comparison, three distinct *subcorpora* corresponding to DC, PCI and PSI were constructed.

For specificity analysis, pre-processing consisted of an elementary tokenization performed with *spaCy* and *itcorenewssm*, without *parser*, *NER* and *tagger*. Texts have been converted to lowercase and only alphabetic tokens have been retained, thus excluding digits, punctuation, and symbols. To avoid statistical dispersion, only lexical units with an overall minimum frequency of 4 were considered.

For semantic-reticular analysis, a more restrictive pre-processing pipeline was adopted. After preliminary cleaning of the text using regular expressions, the *subcorpora* were segmented into sentences and each sentence was treated as a documentary unit. The texts were then lemmatised with *spaCy* and filtered based on explicit linguistic criteria: only lemmas of at least 4 characters in length, belonging to the grammatical categories *NOUN*, *ADJ* and *VERB*, were retained; auxiliaries, proper names, personal entities, standard Italian stop words and a list of

custom stopwords built to remove very frequent but uninformative forms in the parliamentary domain. Sentences that were too short and those containing fewer than three valid tokens after filtering were also discarded. Finally, a shared global vocabulary was built on the set of segments covered, selecting lemmas that exceeded minimum thresholds for both total frequency and documentary frequency. This vocabulary was sorted according to a global *TF-IDF* score and used as a uniform basis for building and comparing the semantic networks of the different *subcorpora*. Once the filtered lexical inventory had been established, semantic relations between lemmas were modelled through co-occurrence analysis.

Usually, when analysing co-occurrence, raw frequency counts the number of times a couple of words appear together, showing the most common words, e.g., *donna* appears hundreds of times in the Constituent Assembly transcripts, and so does *diritto*. In order to capture more semantically related words. In order to measure it and to control the individual frequencies of two terms, we find how often two words appear together rather than alone, and we apply a measure of association, which is the Pointwise Mutual Information (PMI) Bos (2009); Fano (1961); Salle and Villavicencio (2023). It is a measure of how often two events x and y occur, compared with what we would expect if they were independent:

$$I(xy) = \log_2 \frac{P(xy)}{P(x)P(y)} \quad (1)$$

The pointwise mutual information between a target word w and a context word c (Church et al. (1989); Church and Hanks (1990)), measures the probability of capturing w and c together against the probability of observing them independently, allowing to find words that are strongly associated.

$$PMI(wc) = \log_2 \frac{P(wc)}{P(w)P(c)} \quad (2)$$

The main limitations of this measure are:

- high scores to rare pairs;
- it is particularly weak for small-medium corpora.

To overcome these issues, we applied the PMI^2 proposed by Evert (2005), where the co-occurrence frequency is squared in the numerator.

$$PMI^2(a, b) = \log_2 \frac{f_{ab}^2 \cdot N}{f_a \cdot f_b} \quad (3)$$

Squaring f_{ab} allows pairs with low co-occurrence counts to receive a proportionally smaller boost than pairs with high counts.

In Table 2, the pairs *indissolubilità* – *matrimonio* (indissolubility - marriage) and *potere* – *giudiziario* (power-judicial) have the same PMI^2 score, yet they differ in f_{ab} . This is because the PMI^2 measures the consistency of association, not volume. The high PMI^2 of *indissolubilità* – *matrimonio* indissolubility indicates that indissolubility was not treated as a general moral or philosophical principle; it was used almost exclusively in direct connection with marriage. In the discourse of the Assembly, the two words functioned as a single argumentative unit.

potere–*giudiziario* (power-judicial) and *judicial* - *organ* (organo giudiziario) both belong to the register of constitutional law: *potere*–*giudiziario* (power-judicial) designates the judiciary as

Table 2: The table reports the four highest PMI^2 scores for selected lexical pairs in the Constituent Assembly *corpus*.

Pair	f_{ab}	PMI^2
<i>indissolubilità – matrimonio</i>	49	11.11
<i>potere – giudiziario</i>	11	11.03
<i>progetto – costituzione</i>	45	10.01
<i>giudiziario – organo</i>	33	9.81

one of the three branches of state authority, while *judicial - organ* refers to its institutional embodiment. Unlike *indissolubilità – matrimonio* (indissolubility - marriage), these are not collocations born of political controversy — they are technical expressions whose coherence is codified in legal language itself. Their high PMI^2 scores reflect the fact that Assembly members used them with terminological precision: the two constituents of each pair were never separated because the expression only makes sense as a whole. Disaggregating *power* (*potere*) from *judicial* (*giudiziario*) would produce two ordinary words; keeping them together produces a constitutional concept. Finally, we built a co-occurrence network based on the 80 most characteristic lemmas that were selected using TF-IDF, which identifies terms that are frequent in the *corpus* but not uniformly distributed across all sentences. The edges between nodes were weighted using PMI^2 and filtered by the G log-likelihood ratio test at $p < 0.05$, retaining only statistically significant co-occurrences within a sliding window of 7 tokens. Thematic clusters were then identified by applying the Clauset-Newman-Moore greedy modularity maximisation Clauset et al. (2004); Reichardt and Bornholdt (2006) to the resulting network, with a resolution parameter of 0.8.

4. Results and final discussion

The analysis of lexical specificities reflects the different ideological matrices of the three parties examined. In Table 3, signed terms (+) indicate a statistical over-representation of graphic forms, i.e., lemmas used significantly more frequently than the *corpus* average; while signed terms (−) indicate an under-representation, identifying lemmas that *subcorpora* tend to use to a lesser extent than theoretically expected.

Table 3: The ten lexical specificities by political affiliation selected based on the p -value < 0.05 .

DC		PCI		PSI	
+	-	+	-	+	-
magistratura	italia	italia	donna	scuola	famiglia
matrimonio	lavoratori	lavoratori	scuola	legge	matrimonio
indissolubilità	paese	necessità	magistratura	uomo	indissolubilità
unità	vita	popolo	legge	donna	familiare
magistrati	donne	guerra	uomo	vita	magistrati
società	popolo	famiglia	giustizia	problema	coniugi
coniugi	scuola	giovani	coscienza	scuole	figli
familiare	necessità	genitori	libertà	sociale	divorzio
indipendenza	sovranità	diritti	funzione	carattere	genitori
amministrazione	diritti	figli	principio	sovranità	società

The extraction of graphic forms (of over- and under-use) was carried out using an association test based on the Chi-square test of independence (χ^2), which allows us to verify that the distribution of terms is not due to chance but reflects a specific communication strategy of the relevant parties. The order of presentation of the data in the tables follows a decreasing specificity gradient based on the strength of the gap between observed and expected ($\Delta = O - E$). To ensure robustness of the analysis, only terms with a p -value < 0.05 were considered relevant. The analysis in Table 2 highlights a distribution of lemmas characterised by an inverse correlation: the positive specificities of the DC terms almost systematically correspond to the negative specificities in the *subcorpus* of the left-wing parties (PC and PSI). A first difference concerns the relationship between DC and PSI: lemmas such as *matrimonio* (marriage), *indissolubilità* (indissolubility), *coniugi* (spouses) and *familiare* (family) are over-used in DC and under-used in PSI and lemmas such as *scuola* (school), *vita* (life) and *sovranità* (sovereignty) are under-used in DC and over-used in PSI. Christian Democrats favour the centrality of the family as a normative institution and the legal-moral constraints that affect women, while socialists instead use a vocabulary more oriented towards public reform and the definition of collective issues. A second contrast is that between the DC and the PCI: terms like *italia* (italy), *lavoratori* (workers), *necessità* (needs), *popolo* (people) and *diritti* (rights) are over-represented in the PCI and under-represented in the DC, highlighting the opposition between a discourse centred on collective subjects, social needs and the demands dimension and a discourse less inclined to organise itself around the lexicon of popular mobilisation and social conflict. From this perspective, lexical specificity signals not only what each party says most, but also what it chooses to say least, and it is precisely in this tension between over- and under-use that the ideological boundaries between the political cultures become visible.

Even with reference to lexical specificities by gender, two different discursive perspectives emerge that are clearly opposed (Tab. 4). Given the imbalance between the male and female subcorpora, due to the lower number of interventions by female constituents, the analysis of lexical specificities by gender was supplemented with a permutation test. The test was applied to the lexical forms selected for the interpretation of the gender contrast, in order to verify whether the observed differences between the two *subcorpora* were stronger than those obtainable by randomly reassigning gender labels to the interventions.

Table 4: The top ten lexical specificity by gender selected on the basis of the lexical specificity test ($p < 0.05$).

Note: *indicates statistical significance at permutation test with $p < 0.05$

Females	Males
donna	problema*
donne*	giudiziario
famiglia	divorzio*
maternità*	magistratura
diritto	diritto
lavoro	giudice
diritti	matrimonio
condizioni*	magistrati
vita	legge
bambini*	indissolubilità*

The permutation test does not confirm all the lexical specificities with the same degree of robustness. However, it reinforces the central interpretative contrast between the two *subcorpora*.

In the female *subcorpus*, the forms *donne* (women), *maternità* (maternity), *condizioni* (conditions), and *bambini* (children) are statistically significant. These forms point to a discourse oriented towards the protection of the concrete conditions of women, motherhood and childhood. In the male *subcorpus*, by contrast, the statistically significant forms are *problema* (problem), *divorzio* (divorce) and *indissolubilità* (indissolubility), which place the issue of women within a problematising and juridical-family framework, particularly evident in the debate on the marital bond. The remaining forms in Table 4., although relevant at a descriptive level, do not exceed the conventional threshold of the permutation test ($p_{\text{value}} < 0.05$) and are therefore interpreted as lexical tendencies rather than as statistically robust evidence.

The construction of the semantic network (Fig. 3) of the *corpus* outlines a tripartite structure of discourse. The first cluster is positioned on the institutional and judicial system; the second on that of social entities and citizenship; the third on that of the family and the regulation of private relationships. The three clusters are connected by several bridging nodes: in particular, terms such as woman, law, order, social and legal build a strong interconnection between the different clusters, signalling that family, civic and institutional issues are not watertight compartments but semantically interdependent areas.

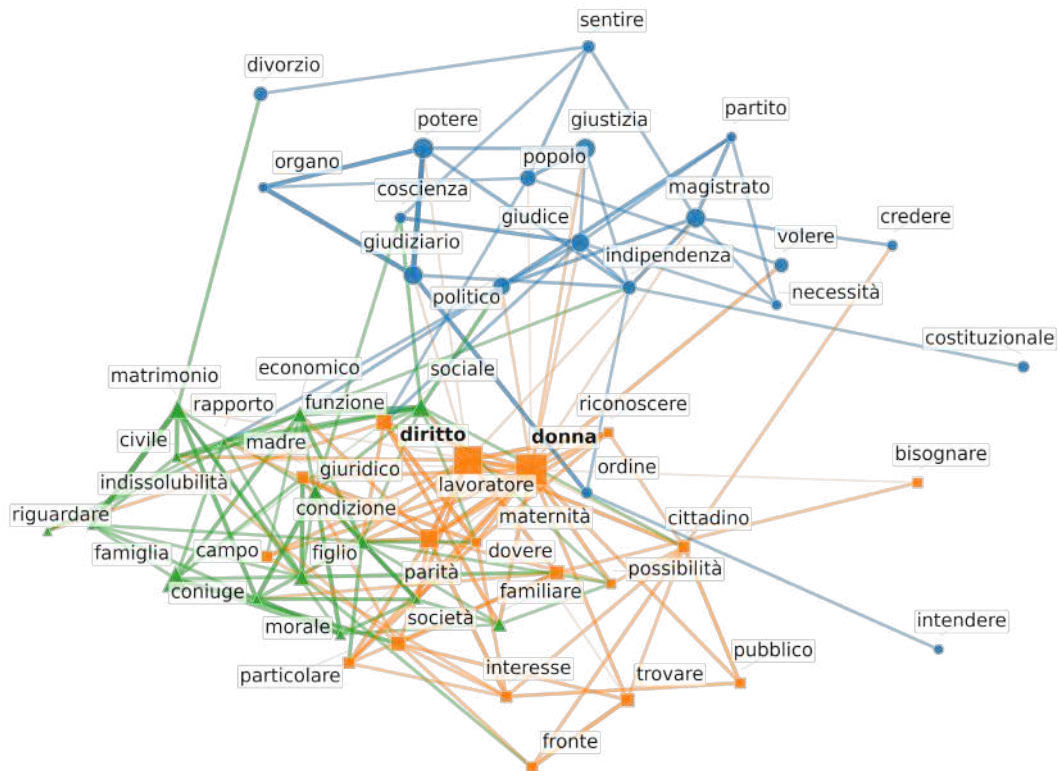


Figure 3: Co-occurrence network based on the 80 most characteristic lemmas that were selected using TF-IDF. Edges between nodes were weighted using PMI^2 . *Blue rounded nodes*: judiciary cluster; *orange squared nodes*: social rights cluster; *green triangular nodes*: family cluster.

Cluster 1 focused on terms such as *indipendenza* (independence), *magistrato* (magistrate), *giudice* (judge), *giudiziario* (judiciary), *potere* (power), *partito* (party), *giustizia* (justice), *popolo* (people), *organo* (organ), *politico* (political), *ordine* (order). It presents values that are relatively close among the three parties, with a slight prevalence of PSI (37%), followed by DC (34%) and PCI (32%). This data suggests that the institutional and judicial lexicon runs throughout the entire *corpus*, with a slightly higher incidence in the socialist *subcorpus*. Cluster 2 built around

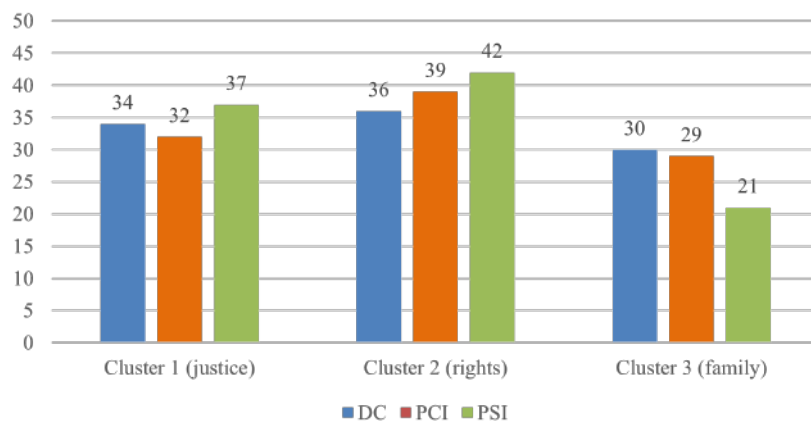


Figure 4: Distribution of semantic weight of parties (DC, PCI, PSI) in the three clusters (%).

terms such as *donna* (woman), *lavoratore* (worker), *diritto* (right), *cittadino* (citizen), *interesse* (interest), *maternità* (motherhood), *riconoscere* (recognise), *società* (society), *madre* (mother), *dovere* (duty). It shows the highest concentration in the PSI (42%), followed by the PCI (39%) and the DC (36%): it is therefore the semantically strongest and most transversally shared nucleus, but also the one in which the Socialist Party seems to invest most intensely. Cluster 3 is characterised by terms such as *giuridico* (legal), *coniuge* (spouse), *matrimonio* (marriage), *economico* (economic), *condizioni* (condition), *morale* (moral), *indissolubilità* (indissolubility), *rapporto* (relationship), *famiglia* (family), *sociale* (social), *parità* (equality), *figlio* (child). It distinguishes the parties much more clearly: it is stronger in the DC (30%) and in the PCI (29%), while it is significantly less present in the PSI (21%). Overall, the analysis not only returns three thematic nuclei but suggests that the global semantic network is composed of three competing and complementary discursive fields: the field of institution, the field of rights and subjects and the field of family.

5. Conclusion remarks and future work

This study shows that the debate on women in the Constituent Assembly does not develop within a unitary discursive structure but rather through heterogeneous lexical and semantic matrices. The analysis of the women's issue discussed by the constituents places the figure of women in relation to other institutions, groups and socialising agencies such as the family, marriage and children, effectively lending itself to bringing out cross-party ideological divergences. The differences between DC, PCI, and PSI concern not only the content of their respective positions but also the use of certain linguistic and stylistic forms through which the themes of family, rights, citizenship, and institutional order are shaped. The gender variable serves as a discriminating element between a feminine lexicon aimed at affirming women in the country's upcoming political and social scene and a masculine lexicon focused on the maintenance of traditional institutions, due, above all, to the conservative pressures of the right.

The analysis of the semantic network refers to a discursive structure organised around three main nuclei: the family, social rights and subjectivity, and the judicial-institutional order. From this perspective, the Constituent Assembly appears not only as a tireless ground of the new constitutional architecture but also as a space for discussion and symbolic reworking, in which different political cultures contribute to defining a renewed vision of society, citizenship, and gender relations. Using only *PMI*² alone would risk retaining spuriously high-scoring rare

pairs, and G^2 alone would favour frequent pairs regardless of actual attraction. In this work, G^2 is used as a data-reduction filter rather than a confirmatory test, this is considered a reasonable approximation, though future work should assess the sensitivity of the network topology to alternative significance thresholds and multiple-testing correction strategies. Other possible extensions will involve replacing PMI^2 with other measures that do not rely on a baseline random distribution to capture the asymmetrical nature of lexical attraction. Furthermore, the current fixed window could be replaced with dependency-parsed co-occurrence, in which edges are drawn only between words that stand in a direct grammatical relation, thereby solving the problem of positional proximity. The results confirm the usefulness of a textual analysis approach to the study of constituent debates: identifying specificities and co-occurrence structures allows us to grasp the intrinsic discursive nuances that outline the image of women.

Finally, recalling the importance of the contributions women made to drafting the Constitutional Charter, it is worth noting that one of them was responsible for the extension of Article 3, which establishes the equality of all citizens before the law. Initially, Article 3 affirmed the equality of citizens without distinction of race, language, religion, political opinion, or social condition. Constituent mother Lina Merlin observed that difference of sex was not mentioned. When it was suggested to her (calling “Signorina”) that this was unnecessary because the phrase “all citizens” (in the masculine) already included women, Merlin replied that in 1789 France had proclaimed the rights of “man and citizen,” yet “citizen” was understood to mean only “the man in trousers.” Women remained excluded. This seemingly small amendment would become the foundation for a series of subsequent gains by Italian women.

References

- Bos J. (2009). Towards a large-scale formal semantic lexicon for text processing. In *From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference*, pp. 3–14.
- Bozzoni M. (2025). *Madri Costituenti: The Construction of Memory and Gendered Representation in the Public Space of the “Founding Mothers” of the Italian Post-WWII Constitution*. Univerzita Karlova, Fakulta humanitních studií.
- Church K., Gale W. A., Hanks P., and Hindle D. (1989). Parsing, word associations and typical predicate-argument relations. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*.
- Church K. and Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- Clauset A., Newman M. E. J., and Moore C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066–111.
- D’Amico M. (2023). *The Ambivalent Parity: The Italian Constitution and Women’s Rights*, pp. 109–135. Springer International Publishing, Cham.
- Evert S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. PhD thesis, Stuttgart, Univ., Diss., 2004.
- Fano R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press.
- Reichardt J. and Bornholdt S. (2006). Statistical mechanics of community detection. *Phys. Rev. E*, 74(1):016–110.
- Salle A. and Villavicencio A. (2023). Understanding the effects of negative (and positive) pointwise mutual information on word vectors. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(8):1161–1199.

Perceptions of Emilia-Romagna: a cross-linguistic analysis of keywords in English and French travel literature¹

Rita Gramellini¹, Daniel Henkel²

¹University of Bologna – rita.gramellini2@unibo.it

²University of Paris 8 Vincennes-Saint-Denis – daniel.henkel@univ-paris8.fr

Abstract

This paper reports a cross-linguistic keyword analysis of travel literature in English and French describing the Italian region of Emilia-Romagna. Using a multi-corpus design, this study identifies the lexical items that distinguish representations of Emilia-Romagna from those of other Italian regions and examines convergences and divergences across languages. Keywords were extracted using the ratio of relative frequencies, log-likelihood and a dispersion criterion, and were ranked by frequency to identify the most specific and commonly used items. Qualitative categorization was then undertaken based on the analysis of their distributional profiles. Results show a broadly convergent cross-linguistic representation of the region defined by its artistic heritage, religious traditions, and historical figures, with some differences in lexical salience. Comparison with travel writing on other Italian regions confirms Emilia-Romagna's distinctive profile and cultural influence, particularly with respect to its pictorial tradition.

Keywords: keyword analysis, corpus linguistics, comparable corpora, travel literature, Italy.

Résumé

Cet article présente une analyse contrastive de mots-clés dans les récits de voyage en anglais et en français consacrés à la région italienne d'Émilie-Romagne. S'appuyant sur un modèle multi-corpus, cette étude identifie les unités lexicales qui distinguent les représentations de l'Émilie-Romagne de celles d'autres régions italiennes et examine les convergences et divergences entre les deux langues. Les mots-clés ont été extraits sur la base du ratio des fréquences relatives, du calcul log-likelihood (fonction log-vraisemblance), et d'un critère de dispersion, et ont été classés par fréquence afin d'identifier les mots-clés les plus spécifiques et les plus utilisés. Une catégorisation qualitative a ensuite été effectuée à travers l'analyse de leurs profils distributionnels. Les résultats révèlent une représentation interlinguistique largement convergente de la région, centrée sur son patrimoine artistique, ses traditions religieuses et ses figures historiques, avec quelques variations de saillance lexicale. La comparaison avec les récits de voyage consacrés à d'autres régions italiennes confirme le profil distinctif de l'Émilie-Romagne et son influence culturelle, notamment en ce qui concerne sa tradition picturale.

Mots clés : analyse de mots-clés, linguistique de corpus, corpus comparable, littérature de voyage, Italie.

1. Introduction

Since the mid-1990s, keyword analysis has become a central method in corpus linguistics for identifying thematic content and stylistic distinctiveness. In statistical terms, keywords are defined as unusually frequent lexical items in a target text or corpus by comparison with a reference corpus (Scott, 1997; Baker, 2004; Scott and Tribble, 2006 *inter alia*), typically

¹ Rita Gramellini was responsible for writing the manuscript. Daniel Henkel contributed to data collection and analysis and revised the manuscript.

extracted using significance tests such as log-likelihood (LL) (Dunning, 1993). The resulting statistic, associated with a p -value, indicates the reliability of differences in word relative frequencies, though it does not convey information about its magnitude (Gabrielatos, 2018). Effect-size measures like the ratio of relative frequencies (RRF) (Damerau, 1993) are therefore recommended alongside significance tests to provide a more meaningful assessment of the degree to which a lexical item is associated with a given text or corpus (Gries, 2010; Gabrielatos, 2018; Pojanapunya and Watson Todd, 2018).

Frequency-based keyword lists, however, can be skewed by items that are statistically key yet appear in only a small number of texts, making them unrepresentative of the corpus as a whole. To address this, dispersion criteria are commonly applied to ensure that selected keywords occur across multiple texts rather than being concentrated in a limited subset (or single text), so that findings reflect generalizable trends rather than idiosyncratic authorial choices (Baker, 2004; Scott and Tribble, 2006; Egbert and Biber, 2019; Gries, 2021).

Building on this methodological framework, the present study investigates the lexical representation of the Italian region of Emilia-Romagna in travel literature. While keyword-based approaches have proven effective for analyzing the lexical and stylistic features of this genre (see Gerbig, 2010), few corpus-based studies have systematically examined how Italy has been represented across languages. Even less is known about individual regions, despite the strongly localized nature of Italy's cultural, historical and artistic identities. To address this gap, this paper reports a cross-linguistic keyword analysis of English and French travel literature describing Emilia-Romagna. A multi-corpus design is adopted to distinguish region-specific lexical features from broader genre patterns. For each language, a target corpus on Emilia-Romagna is compared with a contrastive corpus covering other Italian regions, both of which are evaluated with respect to a reference corpus of general travel literature. The analysis combines quantitative keyword extraction with a qualitative examination of collocations and concordances to identify shared patterns across languages and corpora.

The study is guided by the following research questions:

- Which lexical items characterize representations of Emilia-Romagna in English and French travel literature, as determined by statistically significant disparities in frequency?
- To what extent do these keywords converge across the two languages?
- How does the lexical representation of Emilia-Romagna differ from that of other Italian regions?

By focusing on a single region and integrating a cross-linguistic perspective, multi-corpus design and combined quantitative-qualitative approach, the study offers an innovative contribution to corpus-based travel literature research and sheds light on the interaction between language, place, and cultural representation in historical travel discourse.

The remainder of the paper is structured as follows: Section 2 outlines the methodology, including the corpus design, keyword extraction procedures and qualitative categorization. Section 3 presents the results, and Section 4 discusses the findings.

2. Methodology

2.1. Corpus design

The analysis is based on comparable corpora of travel literature in English and French. For each language, three corpora were compiled: a corpus of interest focusing on Emilia-Romagna

(§2.1.1.), a corpus describing other Italian regions for purposes of comparison (§2.1.2.), and a reference corpus representative of the overall genre (§2.1.3.).

To ensure reliable results, all texts underwent lemmatization and part-of-speech tagging using *TreeTagger* (Schmid, 1994). Moreover, spelling variations and archaic forms typical of historical corpora were normalized prior to analysis.

2.1.1. BER-TL corpus

The Bologna-Emilia-Romagna-Travel-Literature (BER-TL) corpus includes travel writings on the Italian region of Emilia-Romagna spanning the 17th to the 20th centuries, with one 21st-century French text. It builds on the multilingual UniVOCItà corpus² compiled for heritage tourism purposes by the LBC-CeSLiC unit at the University of Bologna, comprising English, French, Spanish, and Russian literary testimonies on the area of Bologna (Zotti, 2026). The English and French subcorpora were subsequently restricted to travel writing and expanded to cover the broader Emilia-Romagna region.

2.1.2. ITA-TL corpus

The Italy-Travel-Literature (ITA-TL) corpus comprises travel writings on Italy outside Emilia-Romagna. The English subcorpus draws on a freely available historical corpus of 30 non-fictional travel narratives³ (Sprugnoli, 2018). Since no pre-existing equivalent French subcorpus was available, a comparable dataset was compiled collecting text from open-access repositories such as the Project Gutenberg⁴. To avoid overlap with the BER-TL corpus, all chapters or sections describing Emilia-Romagna were identified and removed.

2.1.3. REF-TL corpus

The Reference-Travel-Literature (REF-TL) corpus is representative of the travel writing discourse type and serves as a common baseline for both BER-TL and ITA-TL in keyword extraction. It consists of English and French travel writings describing destinations worldwide, compiled using the *Project Gutenberg Corpus Builder* (Moncomble, 2025) through searches constrained by genre (“Travel Writing”) and language (“English” or “French”). The resulting texts were reduced to a random sample of 100 documents per language, ensuring that all were originally written in the target language and published between the 17th and 21st centuries.

Table 1 provides an overview of all comparable corpora described above, reporting the number of files, token count, Type/Token Ratio (TTR), Standardized Type/Token Ratio (STTR) and time frame⁵.

Table 1. Metadata of the comparable corpora

Corpus	Lang.	Files	Tokens	TTR	STTR	Time frame (median)
BER-TL	English	175	1,354,787	0.034	0.468	17 th -20 th c. (1847)
	French	147	1,226,309	0.043	0.472	17 th -21 st c. (1857)
ITA-TL	English	30	2,395,623	0.023	0.464	19 th -20 th c. (1909)

² <https://univocitta.github.io/UniVOCItta/#/corpus>

³ <https://sites.google.com/view/travelwritingsonitaly/corpus>

⁴ <https://www.gutenberg.org/>

⁵ Tokens are defined as sequences of alphanumeric characters. STTR = mean TTR over 1,000-token windows.

REF-TL	French	27	1,709,854	0.036	0.500	18 th -20 th c. (1866)
	English	100	8,047,253	0.013	0.458	18 th -20 th c. (1892)
	French	100	5,712,416	0.019	0.472	18 th -20 th c. (1887)

2.2. Keyword extraction and categorization

Keywords were extracted from the BER-TL and ITA-TL corpora using *AntConc* 3.5.9 (Anthony, 2020), with the REF-TL corpus serving as the reference corpus to filter out lexical units typical of travel writing and obtain keyword lists primarily capturing regional distinctiveness rather than genre-specific lexical patterns⁶.

The extraction procedure combined LL ($p < 0.001$), as a significance filter retaining only items whose frequency differences are unlikely to be due to chance, and RRF (≥ 10), which provides a more intuitive and interpretable measure of the overrepresentation magnitude, ensuring that extracted items are statistically significant and substantially overrepresented. To improve the results' generalizability, a dispersion criterion was applied. Since no consensus exists in the literature regarding the optimal dispersion cut-off, a minimum of 5% of the texts was adopted here as a practical, albeit necessarily arbitrary, threshold, while remaining consistent with dispersion criteria commonly used in corpus-based studies (Pojanapunya and Watson Todd, 2021). Finally, as these measures do not in themselves capture the relative prominence of a keyword within the corpus, insofar as certain technical terms may be highly specific yet relatively uncommon, frequency was used as a final criterion for prioritizing the “core vocabulary” (Henkel, 2020), namely keywords that are both highly specific and widely used in the discourse type under investigation.

As Table 2 shows, the $\text{RRF} \geq 10$ threshold eliminated between 41 and 51% of the initial keywords, while the subsequent dispersion criterion further reduced them by approximately 20-34%, suggesting that a non-negligible proportion of statistically significant items were concentrated in a small number of texts and thus unevenly distributed across the corpus.

Table 2. Number of keywords retained at each extraction step for each subcorpus

Subcorpus	LL	LL + RRF ≥ 10	LL+RRF ≥ 10 +dispersion $\geq 5\%$
English BER-TL	1,488	871	614
French BER-TL	1,176	621	410
English ITA-TL	1,780	959	771
French ITA-TL	1,203	592	418

Following quantitative extraction, the top 50 keywords were qualitatively categorized through a close examination of their collocational profiles and concordance patterns. Collocates were identified within a span of ± 3 words left and right, with a minimum frequency of 3 occurrences (range ≥ 2 texts), using LL ($p < 0.05$) and mutual information (MI) as association measures. Thematic categories were assigned based on concordance analysis, in line with the Firthian principle that meaning is determined by context (Firth, 1957: 11). This applies also to items that may function as components of larger n-grams rather than as independent units, such as Italian function words (e.g. *di*, *della*). In cases where a keyword demonstrated multiple distinct uses across the corpus⁷, the item was assigned to all relevant categories.

⁶ Candidate keywords include all lemmas attested in either corpus, with zero-frequency items not excluded a priori.

⁷ For instance, the French keyword *parmesan* refers both to the painter Parmigianino and to the cheese.

3. Results

Table 3 presents the top 50 keywords extracted from each BER-TL and ITA-TL subcorpus against the REF-TL corpus, ranked by normalized frequency/10k words (f), to prioritize widely used highly specific items, and followed by their RRF values (r) indicating the degree of overrepresentation. While only the top 50 keyword are reported here for reasons of space, the discussion draws on the full ranked lists for a more comprehensive interpretation of the results.

Table 3. Top 50 keywords for each subcorpus

	English BER-TL	French BER-TL	English ITA-TL	French ITA-TL
1	<i>bologna</i> [f:17.9, r:349]	<i>bologne</i> [f:19.7, r:257.5]	<i>rome</i> [f:10.4, r:11]	<i>rome</i> [f:8.2, r:12.3]
2	<i>italy</i> [f:13, r:20.7]	<i>tableau</i> [f:14.4, r:10.9]	<i>italian</i> [f:9.2, r:11.7]	<i>italie</i> [f:8.2, r:11.3]
3	<i>ravenna</i> [f:9, r:225.7]	<i>italie</i> [f:10, r:17.9]	<i>italy</i> [f:8.1, r:12.9]	<i>corinne</i> [f:7.2, r:4044.5]
4	<i>painting</i> [f:7.5, r:15.1]	<i>pape</i> [f:8.3, r:15.4]	<i>venice</i> [f:6.5, r:23.6]	<i>di</i> [f:7.1, r:22.2]
5	<i>ferrara</i> [f:6.2, r:709.3]	<i>rome</i> [f:6.9, r:11.3]	<i>florence</i> [f:5.1, r:16.2]	<i>lord</i> [f:3.3, r:34.6]
6	<i>pope</i> [f:6.1, r:23.8]	<i>parme</i> [f:6.8, r:240]	<i>di</i> [f:4.3, r:19.6]	<i>che</i> [f:3.1, r:61.8]
7	<i>virgin</i> [f:5.9, r:14.9]	<i>ferrare</i> [f:6.5, r:394.9]	<i>piazza</i> [f:4.1, r:21.7]	<i>dante</i> [f:2.7, r:22.5]
8	<i>parma</i> [f:5.8, r:354.6]	<i>carrache</i> [f:6.4, r:875.9]	<i>venetian</i> [f:4, r:28.7]	<i>questo</i> [f:1.9, r:156.7]
9	<i>caracci</i> [f:5.3, r:2124.9]	<i>ravenne</i> [f:5.9, r:697.1]	<i>naples</i> [f:3.7, r:10.2]	<i>per</i> [f:1.8, r:25.8]
10	<i>venice</i> [f:5.3, r:19.4]	<i>autel</i> [f:5.8, r:10.2]	<i>pope</i> [f:3.5, r:13.3]	<i>napolitain</i> [f:1.6, r:13.1]
11	<i>madonna</i> [f:5.2, r:59.4]	<i>peintre</i> [f:5.7, r:11.2]	<i>madonna</i> [f:3, r:34.7]	<i>come</i> [f:1.5, r: 46.3]
12	<i>di</i> [f:5.2, r:23.6]	<i>modène</i> [f:5.7, r:5050.2]	<i>fresco</i> [f:2.6, r:18.6]	<i>pise</i> [f:1.4, r:21.2]
13	<i>guido</i> [f:4.9, r:216.4]	<i>bolonais</i> [f:5.4, r:2155.8]	<i>via</i> [f:2.5, r:12.3]	<i>una</i> [f:1.4, r:44.9]
14	<i>florence</i> [f:4.7, r:14.8]	<i>cardinal</i> [f:5.2, r:13.7]	<i>etruscan</i> [f:2.2, r:61.4]	<i>al</i> [f:1.4, r:14.7]
15	<i>fresco</i> [f:4.6, r:33]	<i>venise</i> [f:5, r:11.2]	<i>palazzo</i> [f:2.2, r:17.5]	<i>montaigne</i> [f:1.2, r:39.1]
16	<i>rimini</i> [f:4.2, r:481.3]	<i>di</i> [f:4.6, r:11.7]	<i>della</i> [f:2.1, r:35.5]	<i>sicilien</i> [f:1.2, r:11.1]
17	<i>modena</i> [f:3.8, r:340.8]	<i>guerchin</i> [f:4.2, r:612.4]	<i>maria</i> [f:2.1, r:12.1]	<i>essere</i> [f:1.1, r:42.7]
18	<i>correggio</i> [f:3.7, r:600.5]	<i>rimini</i> [f:4.1, r:812.5]	<i>giovanni</i> [f:2, r:31.5]	<i>più</i> [f:1, r:65.4]
19	<i>angel</i> [f:3.7, r:10.3]	<i>corrège</i> [f:4.1, r:173.9]	<i>medici</i> [f:2, r:98]	<i>altro</i> [f:1, r:80.2]
20	<i>mosaic</i> [f:3.3, r:21.3]	<i>plaisance</i> [f:3.7, r:22.3]	<i>lorenzo</i> [f:2, r:59.9]	<i>vésuve</i> [f:1, r:15]
21	<i>milan</i> [f:3.3, r:30]	<i>madone</i> [f:3.6, r:15.5]	<i>pisa</i> [f:1.8, r:23]	<i>messin</i> [f:1, r:12.9]
22	<i>raphael</i> [f:3.1, r:38.7]	<i>raphaël</i> [f:3.4, r:16.9]	<i>cardinal</i> [f:1.7, r:11.4]	<i>catane</i> [f:0.9, r:20.5]
23	<i>piazza</i> [f:2.9, r:15.7]	<i>catherine</i> [f:3.2, r:16.4]	<i>genoa</i> [f:1.6, r:17.9]	<i>con</i> [f:0.9, r:24.2]
24	<i>po</i> [f:2.9, r:96.3]	<i>pétrone</i> [f:3.2, r:2539]	<i>cosimo</i> [f:1.4, r:369.7]	<i>syracuse</i> [f:0.9, r:11.8]
25	<i>cardinal</i> [f:2.8, r:19.3]	<i>arioste</i> [f:3.1, r:141.1]	<i>capri</i> [f:1.3, r:45.6]	<i>lady</i> [f:0.8, r:21.6]
26	<i>piacenza</i> [f:2.8, r:742.8]	<i>césar</i> [f:3.1, r:10.1]	<i>façade</i> [f:1.3, r:19.4]	<i>padoue</i> [f:0.8, r:16.3]
27	<i>guercino</i> [f:2.8, r:554.1]	<i>pó</i> [f:3.1, r:34.7]	<i>duomo</i> [f:1.3, r:46.6]	<i>arno</i> [f:0.7, r:18.4]
28	<i>palazzo</i> [f:2.7, r:21.8]	<i>apennin</i> [f:3, r:42.7]	<i>doge</i> [f:1.3, r:23.6]	<i>léonard</i> [f:0.7, r:11.6]
29	<i>bolognese</i> [f:2.7, r:715.2]	<i>dante</i> [f:3, r:17.6]	<i>florentine</i> [f:1.3, r:34.5]	<i>sienne</i> [f:0.7, r:11.1]
30	<i>da</i> [f:2.5, r:16.9]	<i>jérôme</i> [f:2.8, r:23.5]	<i>donatello</i> [f:1.2, r:497.9]	<i>tutto</i> [f:0.7, r:30.8]
31	<i>dante</i> [f:2.3, r:22]	<i>académie</i> [f:2.5, r:10.2]	<i>renaissance</i> [f:1.2, r:15.1]	<i>quale</i> [f:0.6, r:183.4]

32	<i>apennine</i> [f:2.3, r:55.2]	<i>dominique</i> [f:2.5, r:17.8]	<i>micelangelo</i> [f:1.2, r:922.5]	<i>gondole</i> [f:0.6, r:10.8]
33	<i>ariosto</i> [f:2.2, r:889.6]	<i>annibal</i> [f:2.5, r:84.5]	<i>gondola</i> [f:1.2, r:12.6]	<i>bianca</i> [f:0.6, r:17.2]
34	<i>tasso</i> [f:2.2, r:123.3]	<i>antoine</i> [f:2.4, r:10.5]	<i>andrea</i> [f:1.1, r:48.2]	<i>pisan</i> [f:0.6, r:22.3]
35	<i>il</i> [f:2.1, r:13.1]	<i>albane</i> [f:2.3, r:99.2]	<i>titian</i> [f:1.1, r:28.3]	<i>home</i> [f:0.6, r:10.5]
36	<i>padua</i> [f:2.1, r:65.2]	<i>faenza</i> [f:2.3, r:195.7]	<i>neapolitan</i> [f:1.1, r:13.2]	<i>agata</i> [f:0.6, r:667.5]
37	<i>arcade</i> [f:2.1, r:18.5]	<i>forli</i> [f:2.3, r:927.9]	<i>porta</i> [f:1.1, r:14.3]	<i>milanais</i> [f:0.6, r:55.6]
38	<i>francesco</i> [f:2, r:108.8]	<i>légat</i> [f:2.3, r:228.5]	<i>forum</i> [f:1, r:13]	<i>tibre</i> [f:0.6, r:20.7]
39	<i>della</i> [f:2, r:33.7]	<i>alphonse</i> [f:2.2, r:30.8]	<i>giotto</i> [f:1, r:82.3]	<i>véronne</i> [f:0.6, r:19.4]
40	<i>giovanni</i> [f:2, r:30.7]	<i>imola</i> [f:2.2, r:1698.8]	<i>campagna</i> [f:1, r:20.1]	<i>vatican</i> [f:0.6, r:10.7]
41	<i>venetian</i> [f:1.9, r:13.8]	<i>augustin</i> [f:2.1, r:15.7]	<i>siren</i> [f:1, r:136.5]	<i>era</i> [f:0.6, r:161.9]
42	<i>maria</i> [f:1.9, r:11.3]	<i>parmesan</i> [f:2.1, r:273.9]	<i>siena</i> [f:1, r:54]	<i>francesco</i> [f:0.6, r:13.5]
43	<i>adriatic</i> [f:1.8, r:43.9]	<i>farnèse</i> [f:2, r:136.2]	<i>gondolier</i> [f:1, r:30.5]	<i>beaux</i> [f:0.6, r:10.8]
44	<i>petronio</i> [f:1.8, r:1442.2]	<i>reggio</i> [f:2, r:260.1]	<i>vatican</i> [f:1, r:29.3]	<i>arts</i> [f:0.6, r:11.4]
45	<i>theodoric</i> [f:1.8, r:75.6]	<i>ancône</i> [f:2, r:110.1]	<i>fra</i> [f:0.9, r:48]	<i>antonia</i> [f:0.6, r:26.4]
46	<i>domenico</i> [f:1.8, r:109.1]	<i>benoît</i> [f:2, r:27.5]	<i>angela</i> [f:0.9, r:716]	<i>matteo</i> [f:0.5, r:38.4]
47	<i>este</i> [f:1.8, r:176.6]	<i>basilique</i> [f:1.9, r:15.7]	<i>perugia</i> [f:0.9, r:39.8]	<i>dell</i> [f:0.5, r:23.6]
48	<i>saviour</i> [f:1.7, r:11.5]	<i>padoue</i> [f:1.9, r:27.7]	<i>pompeii</i> [f:0.9, r:14.3]	<i>corso</i> [f:0.5, r:20.5]
49	<i>papal</i> [f:1.6, r:32.4]	<i>che</i> [f:1.8, r:26.7]	<i>francesco</i> [f:0.9, r:47.5]	<i>io</i> [f:0.5, r:50.7]
50	<i>cupola</i> [f:1.6, r:19.5]	<i>romagne</i> [f:1.8, r:1431]	<i>campanile</i> [f:0.9, r:24.2]	<i>vasari</i> [f:0.5, r:11.6]

4. Discussion

Keywords of the English and French BER-TL subcorpora show a marked degree of cross-linguistic convergence in the lexical representation of Emilia-Romagna. Many items have direct correspondences or close semantic equivalents in the other language, whether among the top-ranking keywords or lower in the lists. Moreover, non-shared keywords tend cluster around the same domains. While the discussion below focuses primarily on the top 50 items, reference to the full ranked lists is necessary to capture the subcorpora's lexical specificity.

Geographical references form the most salient shared domain. The highest-ranking keywords include major cities such as *Bologna* (f:17.9, r:349)/*Bologne* (f:19.7, r:257.5), *Ravenna* (f:9, r:225.7)/*Ravenne* (f:5.9, r:697.1), and *Ferrara* (f:6.2, r:709.3)/*Ferrare* (f:6.5, r:394.9), alongside natural features, namely the *Po* (f:2.9, r:96.3)/*Pô* (f:3.1, r:34.7) and the *Apennine* (f:2.3, r:55.2)/*Apennin* (f:3, r:42.7). Together, these items suggest that Emilia-Romagna is represented in both traditions as a region defined by historically prominent urban centers embedded within a recognizable landscape.

Moreover, the occurrence of extra-regional cities such as *Venice* (f:5.3, r:19.4)/*Venise* (f:5, r:11.2) and *Padua* (f:2.1, r:65.2)/*Padoue* (f:1.9, r:27.7) indicate that the region is frequently framed within the context of broader Italian itineraries and through inter-city comparisons, a pattern further supported by collocation and concordance evidence⁸. By the same logic, it is not surprising that the ITA-TL subcorpora, dominated by Italy's cultural capitals like *Rome* (En:

⁸ “The railroad from Venice to Bologna passes through Ferrara”; “[In] Modena (...) the houses are built of brick, and most of them have portici, as at Padua and Bologna”.

f:10.4, r:11; Fr: f:8.2, r:12.3), *Pisa* (f:1.8, r:23)/*Pise* (f:1.4, r:21.2), *Venice* (En: f:6.5, r:23.6), identity markers (e.g. *florentine*, *venetian*, *neapolitan/napolitain*) and distinctive local features (e.g. *gondola*, *gondolier*, and *doge* in relation to Venice), still contain references to some Emilian and Romagnol cities, albeit lower in the ranking and with considerably reduced salience. Examples include *Ferrara* (f:0.4, r:46.2)/*Ferrare* (f:0.3, r:15.8), *Rimini* (En f:0.1, r:16.7, Fr: f:0.2, r:43), and *Modène* (f:0.2, r:178.4).

The French BER-TL subcorpus seems to lexicalize the region at a finer level of geographical coverage. In fact, in addition to the major shared toponyms, the French top 50 includes *Faenza* (f:2.3, r:195.7), *Forlì* (f:2.3, r:927.9), *Imola* (f:2.2, r:1698.8), and *Reggio* (f:2, r:260.1). It must be noted, however, that all four also occur within the English BER-TL top 100. Particularly notable is the keyword *Romagne* (f:1.8, r:1431) in the French BER-TL top 50, whose high RRF and collocational profile (*duché*, *capitale*, *pays*) suggest an explicit awareness of Romagna as a distinct historical and political entity. English shows a comparable, though less salient, pattern for *Romagna* (f:0.7, r:1064).

Artistic heritage constitutes a second major area of convergence. In both BER-TL subcorpora, Emilia-Romagna emerges as a region strongly associated with the visual arts, and painting in particular. Artists linked to the Emilian pictorial tradition, such as *Caracci* (f:5.3, r:2124.9)/*Carrache* (f:6.4, r:875.9), *Correggio* (f:3.7, r:600.5)/*Corrège* (f:4.1, r:173.9), and *Guercino* (f:2.8, r:554.1)/*Guerchin* (f:4.2, r:612.4), are highly salient in both languages, alongside more canonical extra-regional references such as *Raphael* (f:3.1, r:38.7)/*Raphaël* (f:3.4, r:16.9), whose works are housed in collections across the region. In addition, both lists clearly highlight the Bolognese school of painting: *bolognese* (f:2.7, r:715.2)/*bolonais* (f:5.4, r:2155.8) show strong collocational attraction with *school* (MI:8.8) and *école* (MI:8.3) respectively, suggesting that Bologna is framed as a center of a recognizable artistic tradition.

The two subcorpora differ, however, in how salience is distributed within this domain. The English top 50 includes *Caracci* as a collective label, alongside *Guido* (f:4.9, r:216.4), a keyword with multiple possible referents but strongly associated with Guido Reni (MI:10.9). By contrast, the French top 50 gives prominence to *Carrache*, *Annibal* (f:2.5, r:84.5) and *Augustin* (f:2.1, r:15.7), and also includes additional artist names such as *Albane* (f:2.3, r:99.2)⁹.

Unlike the BER-TL corpus, both ITA-TL subcorpora associate the visual arts with canonical artists, including *Donatello* (En: f:1.2, r:497.9), *Veronese* (f:0.4, r:29.6; *Véronese* (Fr: f:0.4, r:19), *Tintoretto* (f:0.8, r:60.9)/*Tintoret*; (f:0.4, r:12.9), *Leonardo* (f:0.7, r:33.3)/*Léonard* (f:0.7, r:11.6), the latter strongly associated with *Vinci* (En: MI:13.8 ; Fr: MI:13.2). However, some of the Emilian painters mentioned in the BER-TL do appear among the ITA-TL keywords: in the French list, *Corrège* (f:0.3; r:17.3), *Carrache* (f:0.1; r:20.1), and *Guerchin* (f:0.1; r:23.1); in the English one, *Hannibal* (f:0.3; r:15.1) and *Guido* (f:0.3; r:11.3), which occur in collocation with *Caracci* (MI:13) and *Reni* (MI:15.2) respectively. Concordance analysis shows that these items typically appear in contexts where the artists are cited for artworks located in other Italian cities¹⁰, discussed in broader reflections on Italian art¹¹, or mentioned in comparison with other

⁹ These keywords also occur in the English BER-TL list at lower levels of ranking.

¹⁰ “On your right is the Church of S. Ambrogio [of Genoa], built by Pallavicini, with three pictures, a Guido Reni, the Assumption of the Virgin, and two Rubens”

¹¹ “For the last three hundred years, ever since the Carracci opened their academy at Bologna, there has been no lack of art education in Italy.”

artists elsewhere in Italy¹². The comparison with ITA-TL thus highlights the cultural significance of Emilia-Romagna's artistic tradition more broadly in Italy.

A further contrast concerns the lexical realization of artistic discourse. The English BER-TL subcorpus emphasizes media and techniques through items like *painting* (f:7.5, r:15.1), *fresco* (f:4.6, r:33), and *mosaic* (f:3.3, r:21.3), whereas the French BER-TL top 50 includes *tableau* (f:14.4, r:10.9), *peintre* (f:5.7, r:11.2), and *autel* (f:5.8, r:10.2), the latter frequently occurring in collocation with *tableau* (MI:6.1). The difference between the two languages appears again to be one of relative salience. In fact, art-related vocabulary is present further down the ranked lists in both BER-TL subcorpora, suggesting that pictorial heritage is a distinctive feature that characterizes framing of Emilia-Romagna. This regional specificity is further reinforced by the comparison with the ITA-TL corpus, where none of these terms reach the $RRF \geq 10$ threshold in the respective languages. The only exception is *fresco* (f:2.6, r: 18.6), which is however considerably less salient than in the English BER-TL.

Literary and historical figures emphasize the region's cultural profile. *Ariosto* (f:2.2, r:889.6)/*Arioste* (f:3.1, r:141.1) and *Dante* (En: f:2.2, r:22; Fr: f:3, r:17.6) appear in connection with specific regional cities, most notably Ferrara and Ravenna, anchoring literary prestige to local urban memory. Unlike *Ariosto*, whose salience is concentrated in the BER-TL corpus, *Dante* also appears prominently in the French ITA-TL top 50 (f:2.7, r:22.5), which is unsurprising given his broader association with Italian literary culture and cities such as Florence. This salience in the French ITA-TL subcorpus, however, should be interpreted with caution: although this lexical item is mentioned in 20 out of 27 of the texts, roughly 58% of its occurrences are concentrated in a single work, pointing to an uneven distribution that was not eliminated by the dispersion criterion, and which nonetheless distorts the statistical results.

Religious heritage is also prominent in both BER-TL subcorpora and is lexicalized through a broad and overlapping set of lexical items that includes ecclesiastical authorities, sacred figures, iconographic subjects, and religious sites. Examples include *pope* (f:6.1, r:23.8)/*pape* (f:8.3, r:15.4), *cardinal* (En: f:2.8, r:19.3; Fr: f:5.2, r:13.7), *madonna* (f:5.2, r:59.4)/*madone* (f:3.6, r:15.5) and *Petronio* (f:1.8, r:1442.2)/*Pétrone* (f:3.2, r:2539). Moreover, the English top 50 further evokes iconographic imagery by the presence of *virgin* (f:5.9, r:14.9), *angel* (f:3.7, r:10.3), and *saviour* (f:1.7, r:11.5), whereas the French list places greater emphasis on sacred sites, with items such as *autel* and *basilique* (f:1.9, r:15.7), and on saints and ecclesiastical figures (*Jérôme*, *Antoine*, *Benoît*, *Catherine*). Concordance analysis reveals that most of these items occur also in descriptions of sacred buildings and artworks (e.g. *church of St. Petronio*), functioning simultaneously as religious, artistic and cultural heritage landmarks.

More broadly, several personal names occur across multiple thematic domains, linking political history, artistic traditions, and architectural heritage. For instance, the keyword *Theodoric* (f:1.8, r:75.6) in the English BER-TL occurs in contexts referring both to historical narrative and to monuments such as the mausoleum at Ravenna. Dynastic memory is likewise salient, with the Este family evoked directly in English through *Este* (f:1.8, r:176.6) and indirectly in French through *Alphonse* (f:2.2, r:30.8), which is strongly associated with *Este*¹³ (MI:8.9). Together, these items suggest that travel writing frequently constructs an image of Emilia-Romagna through the cultural and political actors who shaped its monumental landscape.

¹² “Il y avait alors à Gênes deux peintres en vogue, appelés les frères Carlone (...) Ils prétendaient ressusciter la peinture dans leur pays, comme les Carraches à Bologne.”

¹³ *Este* itself appears at rank 55 in the French keyword list (f:1.6, r:70.2).

A further point of convergence between the BER-TL subcorpora is the presence of Italian function words such as *di*, *della*, *che*. In both subcorpora, such items typically occur in proper names of artists, monuments or places (e.g. *Arca di San Domenico*, *Monte della Guardia*), and can therefore be read as markers of local specificity and interlingual contact. More generally, the English list appears somewhat more receptive to Italianized urban and architectural vocabulary, as shown by items such as *piazza*, *palazzo* and *cupola* in the top 50, and by terms like *duomo*, *campanile*, *porta* and *torre* in the top 250. This does not imply that architecture is absent from the French corpus: *piazza* and *palazzo* appear further down the French ranked list, and *basilique*, together with items such as *dôme* and *cathédrale*, points in the opposite direction. Unlike pictorial terms, however, architecture-related lexical items also occur in ITA-TL across both languages. Their presence should therefore be understood primarily as a broader feature of Italy, rather than as a specifically Emilian-Romagnol marker.

Collocational and concordance examination also reveals a minor but consistent gastronomic thread in the specific vocabulary of Emilia-Romagna, with lexical items that are absent from the ITA-TL subcorpora. In the English BER-TL, the keyword *Bologna* functions not only as a place name but also as a gastronomic reference (*bologna sausage*). The French BER-TL subcorpus offers a parallel case in *parmesan*, which refers both to the painter Parmigianino and, more marginally, to Parmesan cheese. The region's culinary identity is further attested by the presence of *sausage* (f:0.6, r:11.4) and *parmesan* (f:0.4, r:28) in the English BER-TL, and *mortadelle* (f:0.2, r:129.3) in the French BER-TL¹⁴.

5. Conclusion

This study has examined the lexical representation of Emilia-Romagna in English and French travel literature through a multi-corpus keyword analysis. Overall, the findings reveal a broadly stable representation of the region as a culturally prestigious destination defined by pictorial traditions, religious heritage, historical figures, and culinary culture, with cross-linguistic differences concerning lexical salience rather than fundamentally different perceptions. The comparison with a corpus of travel writing on other Italian regions confirms the distinctiveness of Emilia-Romagna, particularly in its association with its painting tradition.

By focusing on a single Italian region from a cross-linguistic perspective, the study offers a novel contribution to corpus-based travel literature research, providing a more nuanced understanding on how regional identity is constructed and mediated in historical travel discourse as well as insights relevant to cultural heritage valorization and tourism promotion.

Some limitations should nonetheless be acknowledged. The qualitative categorization was largely restricted to the top 50 keywords per subcorpus due to space limitations. Furthermore, the extraction procedure focuses primarily on a single combination of keyness measures. Future work will extend the analysis to a larger portion of the ranked lists and explore alternative metrics, with the aim of gaining additional insights into the data and providing a more comprehensive picture of the lexical representation of the region.

¹⁴ The association between the region and its culinary identity is further reinforced linguistically: *bologna sausage* and *parmesan* derive metonymically from the place names Bologna and Parma.

References

- Anthony, L. (2020). *AntConc* (Version 3.5.9). Computer Software. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software/AntConc>
- Baker, P. (2004). Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis: Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*, 32(4): 346-359.
- Damerau, F. J. (1993). Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts. *Information Processing & Management*, 29(4): 433-447.
- Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74.
- Egbert J. and Bieber D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14(1): 77-104.
- Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930–1955. In *Studies in Linguistic Analysis*. Blackwell.
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In *Corpus approaches to discourse*. Routledge.
- Gerbig, A. (2010). Key words and key phrases in a corpus of travel writing: From Early modern English literature to contemporary “blooks”. In Bondi, M. and Scott, M., editors, *Keyness in texts*. John Benjamins Publishing.
- Gries, S. T. (2010). Useful Statistics for Corpus Linguistics. In Sánchez Pérez, A. and Almela Sánchez, M., editors, *A Mosaic of Corpus Linguistics: Selected Approaches*. Peter Lang.
- Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2): 1-33.
- Henkel, D. (2020). Lexical Association as an Indicator of Conceptual Convergence and Divergence in English, French and Italian. In Aláman, A. and Zotti, V., editors, *The Language of Art and Cultural Heritage. A plurilingual and digital perspective*. Cambridge Scholar Publishing.
- Moncomble, F. (2025) 2026. *Project Gutenberg Corpus Builder*. Web application. Arras, France: Université d'Artois. <https://corpustools.prendrelangue.fr/pgcorpusbuilder/>
- Pojanapunya, P. and Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1): 133-167.
- Pojanapunya, P. and Watson Todd, R. (2021). The Influence of the Benchmark Corpus on Keyword Analysis. *Register studies*, 3(1): 88-114.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester.
- Scott, M. (1997). PC analysis of key words – And key key words. *System*, 25(2): 233-245.
- Scott, M. and Tribble, C. (2006). *Textual patterns: key words and corpus analysis in language education* (Studies in corpus linguistics 22). John Benjamins Publishing.
- Sprugnoli, R. “Two days we have passed with the ancients...”: a Digital Resource of Historical Travel Writings on Italy. In Spampinato D., editor, *The Book of Abstract of AIUCD 2018*. Bologna.
- Zotti, V. (2026), Lessico artistico plurilingue e valorizzazione del patrimonio culturale italiano: i progetti UNICittà e UniVOCittà. In Altmanova, J., Centrella, M. and Grimaldi, C., editors, *Terminologia del patrimonio culturale materiale e immateriale: analisi e approcci di studio*. Peter Lang.

C13

Health, clinical and accessibility text analytics

09-lug
17:00 - 18:30
Aula Multimediale C

Living with osteoporosis across contexts: a lexicometric comparison of Italian and Spanish narratives

Chiara Tedesco¹, Ercole Vellone¹, Rosaria Alvaro¹, Gianluca Pucciarelli¹,
Vicente Bernalte-Martí², Mariachiara Figura³

¹ RN, Ph.Ds in Nursing Science and Public Health. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: chiara.tedesco.17@students.uniroma2.eu

² Full Professor in Nursing Science. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: ercole.vellone@uniroma2.it

³ Full Professor in Nursing Science. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: rosaria.alvaro@uniroma2.it

⁴ Associate Professor in Nursing Science. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: gianluca.pucciarelli@uniroma2.it

⁵ Assistant Professor in Nursing Science. Department of Nursing, Faculty of Health Sciences, University of Jaume I, Castelló, Spain

e-mail: bernalte@uji.es

⁶ Assistant Professor in Nursing Science. Department of Maternal and Child Health Promotion, Internal Medicine & Specialties of Excellence “G. D’Alessandro” (PROMISE), Palermo (PA). Italy

e-mail: mariachiara.figura@unipa.it

Abstract

Osteoporosis is a chronic condition that requires continuous self-care and adaptation in everyday life, giving rise to complex narrative constructions of bodily experience, management practices and interaction with healthcare systems. This study investigates cross-national differences in the narratives of people living with osteoporosis through an Automatic Analysis of Textual Data approach within the framework of Exploratory Multidimensional Data Analysis.

A corpus of forty in-depth semi-structured interviews (Italy $n = 20$; Spain $n = 20$) was analysed using correspondence analysis and Labbé's intertextual distance in order to examine both the semantic positioning of national sub-corpora and their degree of lexical divergence. The interview corpus was analysed using Labbé's intertextual distance and correspondence analysis to assess whether the two national sub-corpora exhibit measurable lexical divergence and characterise the semantic organisation underlying this divergence.

Labbé's intertextual distance revealed a marked lexical separation between the Italian and Spanish interviews (0.44), **suggesting partially differentiated semantic configurations in the ways osteoporosis experience is articulated across the two national contexts**. In line with its use in comparative text analysis, this measure is interpreted as a descriptive indicator of lexical divergence rather than as a threshold-based classificatory criterion. The correspondence analysis of lexical forms highlights a differentiated semantic positioning of Italian and Spanish narratives. In particular, Italian narratives tend to construct osteoporosis as a condition to be understood, managed and negotiated through clinical knowledge and therapeutic practices, whereas Spanish narratives more often frame the illness through embodied experience, everyday routines and situated vulnerability.

Overall, the combined use of intertextual distance and correspondence analysis supports a cross-cultural reading of osteoporosis narratives. These findings illustrate how lexicometric tools can inform both linguistic analysis of illness narratives and, more broadly, the understanding of how cultural settings shape the discursive articulation of everyday life with a chronic illness.

Keywords: Automatic textual analysis, Self-care practices, Osteoporosis, EMDA, Nursing

Abstract in Italiano

L'osteoporosi è una patologia cronica che richiede una cura di sé costante e un continuo adattamento nella vita quotidiana, dando origine a complesse costruzioni narrative dell'esperienza corporea, delle pratiche di gestione della malattia e dell'interazione con i sistemi sanitari. Questo studio indaga le differenze transnazionali nelle narrazioni delle persone con l'osteoporosi attraverso un approccio di Analisi Automatica dei Dati Testuali nel quadro dell'Analisi Esplorativa Multidimensionale dei Dati.

Un corpus di quaranta interviste approfondite semi-strutturate (Italia $n = 20$; Spagna $n = 20$) è stato analizzato utilizzando l'analisi delle corrispondenze e la distanza intertestuale di Labbé. Il corpus delle interviste è stato analizzato utilizzando la distanza intertestuale di Labbé e l'analisi delle corrispondenze per valutare se i due subcorpora nazionali presentassero una divergenza lessicale misurabile e per caratterizzare l'organizzazione semantica alla base di tale divergenza. La distanza di Labbé ha rivelato una separazione lessicale tra le interviste italiane e quelle spagnole (0,44), **suggerendo configurazioni semantiche parzialmente differenziate nelle modalità attraverso cui l'esperienza dell'osteoporosi viene articolata nei due contesti nazionali**. L'analisi delle corrispondenze evidenzia un posizionamento semantico differenziato delle narrazioni italiane e spagnole. In particolare, le narrazioni italiane tendono a costruire l'osteoporosi come una condizione da comprendere, gestire e negoziare attraverso conoscenze cliniche e pratiche terapeutiche, mentre le narrazioni spagnole inquadrano più spesso la malattia attraverso l'esperienza incarnata, le routine quotidiane e la vulnerabilità situata. Nel complesso, l'uso combinato della distanza intertestuale e dell'analisi delle corrispondenze favorisce una lettura interculturale delle narrazioni sull'osteoporosi. Questi risultati dimostrano come gli strumenti lessicometrici possano fornire spunti sia per l'analisi linguistica delle narrazioni sulla malattia sia, più in generale, per comprendere in che modo i contesti culturali influenzano l'articolazione discorsiva della vita quotidiana con una malattia cronica.

Parole chiave: Analisi testuale automatica, Pratiche di auto-cura, Osteoporosi, EMDA, Assistenza infermieristica

1. Introduction

The growing volume of textual data available in health research has fostered increasing interest in methods capable of treating discourse as a structured linguistic system, facilitating the examination of large corpora and the detection of underlying semantic patterns.

Among the domains most suited to such approaches, patient narratives hold particular relevance. People managing chronic conditions are constantly engaged in interpreting and negotiating medical knowledge alongside bodily experience and daily routines, and their accounts provide a privileged window into illness understanding and self-care organisation.

Osteoporosis offers a compelling illustration of this complexity. A chronic skeletal disorder characterised by elevated fracture risk, it demands sustained behavioural engagement across multiple domains (including medication adherence, fall prevention, physical activity, and nutritional adjustment) (Hemmati et al., 2021; Johnston & Dagar, 2020; Kanis et al., 2021; Tarantino et al., 2017). Despite extensive biomedical investigation, how patients discursively construct and communicate their illness experience has received comparatively little attention. Lexicometric methods provide a systematic approach to uncovering the semantic architecture of such narratives and the linguistic configurations through which health experiences are expressed (Figura et al., 2024). **Rather than approaching interviews as isolated subjective accounts, EMDA conceptualises the corpus as a structured semantic field in which latent oppositions, lexical regularities, and discursive configurations emerge through the statistical modelling of textual distributions. Within this perspective, textual analysis does not simply describe recurring themes but aims to reconstruct the semantic organisation underlying discursive representations of illness experience.**

Italy and Spain offer two comparable yet culturally distinct contexts for this investigation. Despite sharing universal healthcare systems and similar epidemiological profiles, both countries differ in health behaviours and cultural attitudes toward food, family roles, mental health, and stress responses (Bersia et al., 2024; Giannotti et al., 2022; Meñaca et al., 2012; Scarpato et al., 2022). Grounded in Exploratory Multidimensional Data Analysis (EMDA), this study therefore aims to explore and compare self-care narratives among Italian and Spanish adults living with osteoporosis.

2. Methods

A multicentric qualitative study was conducted with 40 adults living with osteoporosis (Italy $n = 20$; Spain $n = 20$), interviewed through in-depth semi-structured interviews informed by the Middle-Range Theory of Self-Care of Chronic Illness (Riegel et al., 2012, 2025).

Corpus processing and analysis were performed in IRaMuTeQ, an open-source R-based software for multidimensional textual analysis (Ratinaud, 2024). Transcripts were standardised, segmented into text segments (TS), and lemmatised to normalise morphological variation. Descriptive textometric indicators were computed (including tokens, types, hapax legomena, mean segment length, and TTR) alongside lexical frequency distribution examined against Zipf's law to assess the linguistic adequacy of the corpus. The corpus was processed through Automatic Analysis of Textual Data (AATD) within the EMDA framework (Bolasco, 2021; Fraire, 2009).

Two complementary analysis were performed: Labbé's intertextual distance to quantify lexical divergence between the two national sub-corpora (Cortelazzo et al., 2013; Labbé & Labbé, 2003), and correspondence analysis to explore semantic organisation within a multidimensional factorial space, projecting Italian and Spanish sub-corpora onto shared factorial dimensions.

The sequential integration of intertextual distance and correspondence analysis was intended to provide a multi-level exploration of the corpus. While Labbé's distance allowed the quantification of lexical divergence between national sub-corpora, correspondence analysis enabled this divergence to be structurally interpreted through the reconstruction of latent semantic oppositions within a shared factorial space.

Because the aim of the study was not to compare Italian and Spanish linguistic systems per se, but rather to explore the semantic organisation of illness narratives across cultural contexts, all interviews were translated into English before corpus construction (Chen & Boore, 2010; Cuoco et al., 2022). This choice was intended to allow all textual units to be projected within a shared factorial space and to avoid fragmentation linked to language-specific lexical systems. Nevertheless, translation cannot be considered a neutral preprocessing procedure in lexicometric analysis. Lexical equivalences, semantic nuances, and distributional regularities may be partially reshaped through translators' interpretive decisions, potentially influencing both intertextual distance and factorial structures. For this reason, translation was conceptualised not as a purely technical step, but as an interpretive stage of corpus construction requiring methodological transparency and reflexive consideration.

3. Results

3.1 Sample characteristics

The textual corpus consisted of 40 semi-structured interviews, evenly distributed between Italy ($n = 20$) and Spain ($n = 20$). Participants were predominantly women (38 out of 40), with a mean age of 68.6 years ($SD = 5.74$). All participants had received a diagnosis of primary osteoporosis, most commonly postmenopausal osteoporosis, and were undergoing clinical monitoring and pharmacological treatment at the time of the interview. These interviews constituted the corpus submitted to lexicometric analysis.

3.2 Corpus description

The corpus comprised 120 interview texts, segmented into 1,536 TS, with 53,937 tokens, 2,677 lexical forms, and a mean segment length of 35.1 words. Hapax legomena accounted for 993 forms (37.09%), reflecting a balance between individual lexical variability and shared thematic coherence. The TTR (0.05), though low in absolute terms, is consistent with interview corpora characterised by thematic concentration. Lexical frequency distribution conformed to Zipf's law (Bolasco, 2021), confirming the coexistence of a stable high-frequency nucleus and a long tail of individual narrative variation, and supporting the linguistic validity of the corpus for multidimensional analysis. The DHC retained 1,428 TS (92.97%), indicating high internal coherence (Table 1).

Table 1: General characteristics of the textual corpus.

Indicator	Value
Number of texts	120
Text segments	1,536
Word occurrences (tokens)	53,937
Lexical forms (types)	2,677
Hapax legomena	993
Hapax proportion	37.09%

Average segment length	35.1 words
Type–Token Ratio (TTR)	0.05

3.3 Intertextual distance

The comparison between the Italian and Spanish interview sets first revealed a measurable lexical divergence. The intertextual distance computed according to Labbé’s metric reached a value of 0.438, suggesting an observable degree of lexical differentiation between the Italian and Spanish sub-corpora (Table 2).

Because Labbé’s distance does not rely on predefined interpretative thresholds, the observed value should be interpreted comparatively rather than classificatorily. In the present study, the measure is therefore considered a descriptive indicator of lexical divergence whose analytical relevance emerges primarily through its integration with correspondence analysis and the reconstruction of the underlying semantic structure (Cortelazzo et al., 2013; Labbé & Labbé, 2003). Within this perspective, the observed distance suggests that the experience of living with osteoporosis is articulated through partially differentiated semantic and lexical configurations across the Italian and Spanish interview sets.

Table 2. Intertextual distance

Comparison	X1	X2	Average Distance	Interpretation
Between Italian and Spanish sets	ITA	ESP	0.438	Measurable lexical divergence between sub-corpora

3.4 Correspondence analysis

The correspondence analysis generated a factorial representation of the corpus in which lexical forms and contextual modalities are positioned according to their distribution across text segments. The projection of national modalities onto the factorial plane (Figure 1) reveals that the first factorial axis, accounting for 42.9% of the total variance, primarily structures the corpus according to national origin, separating the Italian and Spanish sub-corpora along a horizontal dimension. The Italian modality is positioned on the right side of the factorial plane, the Spanish on the left.

The lexical map associated with this axis suggests two partially differentiated semantic configurations. Within the Italian sub-corpus, high-frequency terms related to clinical management — *osteoporosis* (285), *medication* (118), *treatment* (100), *condition* (91), *manage* (75) — provide the lexical backdrop against which three thematic dimensions emerge. The illness experience dimension is most strongly associated with *osteoporosis* ($\chi^2=12.88$), *fracture* ($\chi^2=6.92$), *condition* ($\chi^2=7.80$), and *hope* ($\chi^2=5.73$). The self-care dimension is structured around *experience* ($\chi^2=18.64$), *side effects* ($\chi^2=10.62$), and *medication* ($\chi^2=7.74$), suggesting that self-care is narrated primarily through engagement with treatment and its effects. Contextual factors include *difficulty* ($\chi^2=7.28$), *lifestyle* ($\chi^2=6.73$), *challenge* ($\chi^2=6.39$), and *specialize* ($\chi^2=7.85$), though whether these reflect facilitating or constraining elements cannot be determined from distributional analysis alone.

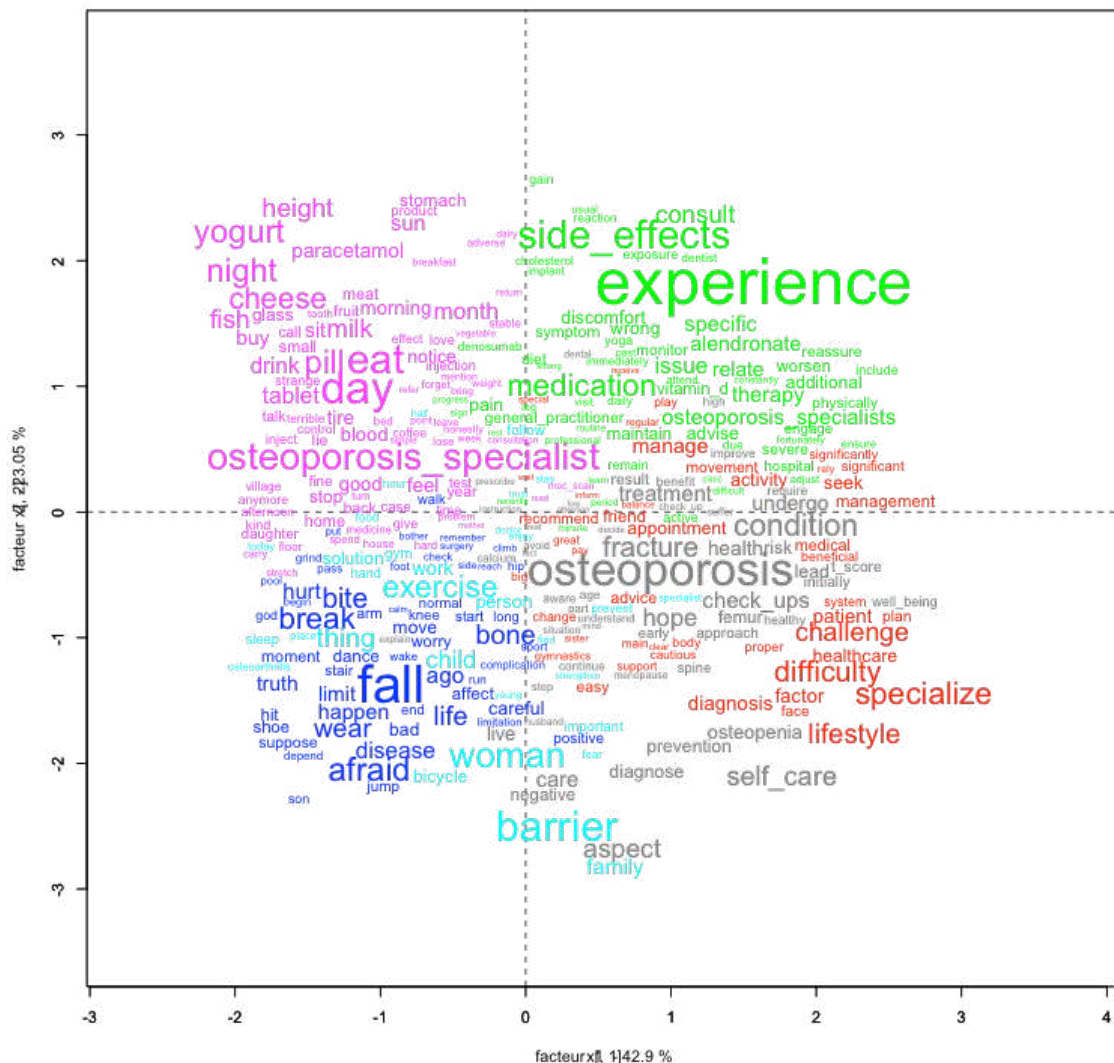


Figure 1. Correspondence analysis: Correspondence analysis of lexical forms, illustrating the factorial space defined by lexical co-occurrences and the latent semantic dimensions structuring narratives of living with osteoporosis.

Within the Spanish sub-corpus, the dominant high-frequency terms — *feel* (150), *pain* (120), *walk* (125), *fall* (81), *hurt* (62) — suggest a semantic organisation oriented toward embodied experience and everyday life. The illness experience dimension is most strongly associated with *fall* ($\chi^2=14.77$), *afraid* ($\chi^2=8.41$), *bone* ($\chi^2=6.56$), and *break* ($\chi^2=7.71$), pointing to a narrative centred on bodily vulnerability and fear. The self-care dimension is structured around everyday habits and dietary practices: *eat* ($\chi^2=10.95$), *day* ($\chi^2=12.35$), *pill* ($\chi^2=8.38$), *osteoporosis_specialist* ($\chi^2=9.97$), and *cheese* ($\chi^2=7.61$). Contextual factors include *barrier* ($\chi^2=11.41$), *woman* ($\chi^2=9.34$), and *exercise* ($\chi^2=7.80$), though again their facilitative or constraining directionality remains beyond the reach of the present analysis.

The second factorial axis, explaining 23.05% of the variance, introduces a further semantic opposition cutting across both national groups, contrasting active illness management strategies

in the upper part of the factorial space with terms referring to constraints, concerns, and experiential difficulties in the lower part. Together, the two axes account for 65.95% of the total variance and show that narratives of living with osteoporosis are organised around the interplay between agency and limitation, while Italian and Spanish participants occupy distinct positions within this semantic structure.

4. Discussion

The present study illustrates how lexicometric tools and EMDA may offer productive resources for the systematic comparison of qualitative corpora across cultural contexts. The sequential application of two complementary analytical procedures (Labbé's intertextual distance and correspondence analysis) proved particularly suited to this comparative purpose, allowing lexical divergence to be first quantified and then semantically explored. Labbé's intertextual distance (0.44) provided an initial, descriptive indication that the Italian and Spanish narratives draw on partially distinct lexical repertoires to articulate the experience of living with osteoporosis. While this metric does not operate according to predefined classificatory thresholds, the value observed suggests a meaningful degree of lexical separation between the two sub-corpora, warranting further structural investigation. **For this reason, the analytical relevance of the observed value should not be interpreted in absolute terms, but rather in relation to the broader semantic structure reconstructed through correspondence analysis. Future studies may further contextualise intertextual distance through internal comparative procedures, such as within-corpus splits or sensitivity analyses, to strengthen interpretative benchmarking.**

The correspondence analysis then allowed this divergence to be mapped onto a shared factorial space, revealing the semantic oppositions potentially underlying it. **The findings should therefore not be interpreted as direct linguistic differences between Italian and Spanish languages, but rather as differentiated semantic configurations emerging from translated and standardised illness narratives constructed within a shared analytical space.**

Beyond lexical differentiation, the factorial structure suggests that chronic illness is discursively organised through culturally situated semantic configurations. Italian narratives tended to frame osteoporosis through clinical negotiation, therapeutic management, and interaction with specialised care, whereas Spanish narratives more frequently articulated illness through embodied vulnerability, everyday routines, and experiential constraints. These oppositions suggest that self-care practices are not constructed as universal behavioural categories, but rather emerge through culturally mediated relationships between body, agency, and healthcare systems. The separation of national modalities along the first factorial axis (42.9% of variance) suggested that the two groups tend to organise their illness narratives around structurally distinct discursive configurations, while the second axis (23.05%) pointed to a tension between active management and experiential limitation that appears to cut across both national groups. **While the prominence of nationality along the first factorial axis is partly expected given its role as a contextual variable structuring the corpus, the analytical relevance of the correspondence analysis lies less in the separation itself than in the semantic organisation associated with each pole of the factorial space. In this sense, the factorial structure does not merely differentiate national groups, but highlights the distinct discursive configurations through which illness experience and self-care are articulated across contexts.** The chi-square specificity measures further allowed lexical associations to be quantified within each thematic dimension, providing statistical grounding for interpretive claims.

More broadly, this approach treats interview data not as a collection of individual statements but as a structured linguistic system whose distributional properties can be modelled statistically. **The relatively high proportion of hapax legomena further suggests the coexistence of shared semantic regularities and substantial individual lexical variation within the corpus.**

While this variability is consistent with the heterogeneous and experiential nature of illness narratives, it also indicates that the observed semantic structures may partially reflect speaker-level differences that cannot be fully disentangled within the present corpus size.

The regularities identified through automatic textual analysis may serve as heuristic devices guiding interpretation, though their relationship to deeper discursive structures remains a matter for further investigation. Lexical distributions should therefore not be interpreted as direct representations of discourse, but rather as distributional traces through which broader semantic and interpretative configurations may be analytically inferred. Accordingly, terms such as “discursive configurations” or “cultural framing” are used in the present study as interpretative constructs grounded in statistical regularities rather than as direct reconstructions of discourse in a qualitative-hermeneutic sense.

The present analysis was primarily oriented toward the reconstruction of corpus-level semantic structures rather than toward the idiographic interpretation of individual narratives or excerpts. At the same time, the factorial organisation observed in the corpus should be interpreted with caution, as additional contextual factors, including interviewer interaction, socio-demographic variation, and contextual differences in interview settings, may also contribute to shaping lexical distributions and semantic positioning within the factorial space.

These findings also resonate with broader literature on chronic illness narratives and self-care, suggesting that the experience of chronicity is not articulated exclusively through biomedical management but through culturally mediated relationships between bodily vulnerability, therapeutic responsibility, and everyday life. In this sense, the semantic oppositions identified through correspondence analysis may reflect different modes of negotiating chronic illness experience within distinct healthcare and socio-cultural contexts.

Although the corpus size was adequate for exploratory lexicometric analysis, the relatively limited number of interviews within each national sub-corpus may reduce the stability and generalisability of the observed lexical configurations. The relatively high proportion of hapax legomena further suggests the coexistence of shared semantic regularities and substantial individual lexical variation within the corpus. While this variability is consistent with the heterogeneous and experiential nature of illness narratives, it also indicates that the observed semantic structures may partially reflect speaker-level differences that cannot be fully disentangled within the present corpus size.

Future studies based on larger and more heterogeneous corpora may further assess the reproducibility of the semantic structures identified in the present analysis. Future studies may also strengthen the robustness of factorial interpretation through stability-oriented procedures, such as resampling strategies or split-corpus analyses aimed at assessing the reproducibility of semantic structures across subsamples.

5. Conclusion

Overall, the combined use of intertextual distance and correspondence analysis supports a cross-cultural reading of osteoporosis narratives. These findings illustrate how lexicometric tools can inform both linguistic analysis of illness narratives and, more broadly, the understanding of how cultural settings shape the discursive articulation of everyday life with a chronic illness.

References

- Bersia, M., Charrier, L., Zanaga, G., Gaspar, T., Moreno-Maldonado, C., Grimaldi, P., Koumantakis, E., Dalmasso, P., & Comoretto, R. I. (2024). Well-being among university students in the post-COVID-19 era: a cross-country survey. *Scientific Reports*, 14(1), 18296. <https://doi.org/10.1038/s41598-024-69141-9>
- Bolasco, S. (2021). *L'analisi automatica dei testi. Fare ricerca con il Text Mining*. Carrocci Editore.
- Chen, H., & Boore, J. R. (2010). Translation and back-translation in qualitative nursing research: methodological review. *Journal of Clinical Nursing*, 19(1–2), 234–239. <https://doi.org/10.1111/j.1365-2702.2009.02896.x>
- Cortelazzo, M. A., Nadalutti, P., & Tuzzi, A. (2013). Improving Labbé's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature. *Journal of Quantitative Linguistics*, 20(2), 125–152. <https://doi.org/10.1080/09296174.2013.773138>
- Cuoco, A., Arcadi, P., Figura, M. C., Piervisani, L., Alvaro, R., Vellone, E., & Durante, A. (2022). Designing and conducting qualitative research across countries and cultures: challenges for inclusiveness and rigour. *European Journal of Cardiovascular Nursing*, 21(8), 873–879. <https://doi.org/10.1093/eurjcn/zvac067>
- Figura, M., Arcadi, P., Vellone, E., Pucciarelli, G., Simeone, S., Piervisani, L., & Alvaro, R. (2024). Living in a multicultural context: Health and integration from the perspective of undocumented Mediterranean migrants, residents, and stakeholders in Italy. A qualitative-multimethod study. *Journal of Advanced Nursing*, 80(8), 3382–3394. <https://doi.org/10.1111/jan.16036>
- Fraire, M. (2009). Statistical methods for exploratory multidimensional data analysis on time use. *Statistica*, 69(4), 317–341.
- Giannotti, M., Mazzoni, N., Facchini, M., de Falco, S., Venuti, P., & Iandolo, G. (2022). Determinants of maternal stress during COVID-19 outbreak in Italy and Spain: A cross-cultural investigation. *Journal of Family Psychology*, 36(6), 827–838. <https://doi.org/10.1037/fam0000993>
- Hemmati, E., Mirghafourvand, M., Mobasser, M., Shakouri, S. K., Mikaeli, P., & Farshbaf-Khalili, A. (2021). Prevalence of primary osteoporosis and low bone mass in postmenopausal women and related risk factors. *Journal of Education and Health Promotion*, 10, 204. https://doi.org/10.4103/jehp.jehp_945_20
- Johnston, C. B., & Dagar, M. (2020). Osteoporosis in Older Adults. *The Medical Clinics of North America*, 104(5), 873–884. <https://doi.org/10.1016/j.mcna.2020.06.004>
- Kanis, J. A., Norton, N., Harvey, N. C., Jacobson, T., Johansson, H., Lorentzon, M., McCloskey, E. V., Willers, C., & Borgström, F. (2021). SCOPE 2021: a new scorecard for osteoporosis in Europe. *Archives of Osteoporosis*, 16(1), 82. <https://doi.org/10.1007/s11657-020-00871-9>
- Labbé, C., & Labbé, D. (2003). La distance intertextuelle. *Corpus*, (2). <https://doi.org/10.4000/corpus.31>
- Meñaca, A., Evans, N., Andrew, E. V. W., Toscani, F., Finetti, S., Gómez-Batiste, X., Higginson, I. J., Harding, R., Pool, R., & Gysels, M. (2012). End-of-life care across Southern Europe: A critical review of cultural similarities and differences between Italy, Spain and Portugal. *Critical Reviews in Oncology/Hematology*, 82(3), 387–401. <https://doi.org/10.1016/j.critrevonc.2011.06.002>
- Ratinaud, P. (2024). *IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [software]* (Version 0.7 alpha 2). (2009) Available from: <http://www.iramuteq.org>. In <http://www.iramuteq.org/>.

- Riegel, B., Jaarsma, T., & Strömberg, A. (2012). A middle-range theory of self-care of chronic illness. *ANS. Advances in Nursing Science*, 35(3), 194–204.
<https://doi.org/10.1097/ANS.0b013e318261b1ba>
- Riegel, B., Jaarsma, T., & Strömberg, A. (2025). An Update to the Middle-Range Theory of Self-Care of Chronic Illness. *Advances in Nursing Science*.
<https://doi.org/10.1097/ANS.0000000000000594>
- Scarpato, D., Rotondo, G., & Simeone, M. (2022). Healthy and Sustainable Food: A Cross-Cultural Study among Spanish and Italian Consumers. *Foods*, 11(13), 1941.
<https://doi.org/10.3390/foods11131941>
- Tarantino, U., Iolascon, G., Cianferotti, L., Masi, L., Marcucci, G., Giusti, F., Marini, F., Parri, S., Feola, M., Rao, C., Piccirilli, E., Zanetti, E. B., Cittadini, N., Alvaro, R., Moretti, A., Calafiore, D., Toro, G., Gimigliano, F., Resmini, G., & Brandi, M. L. (2017). Clinical guidelines for the prevention and treatment of osteoporosis: summary statements and recommendations from the Italian Society for Orthopaedics and Traumatology. *Journal of Orthopaedics and Traumatology*, 18(S1), 3–36.
<https://doi.org/10.1007/s10195-017-0474-7>

Self-care behaviours in people with osteoporosis: a multidimensional analysis of textual data

Chiara Tedesco¹, Ercole Vellone², Rosaria Alvaro³, Gianluca Pucciarelli⁴,
Vicente Bernalte-Martí⁵, Mariachiara Figura⁶

¹ RN, Ph.Ds in Nursing Science and Public Health. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: chiara.tedesco.17@students.uniroma2.eu

² Full Professor in Nursing Science. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: ercole.vellone@uniroma2.it

³ Full Professor in Nursing Science. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: rosaria.alvaro@uniroma2.it

⁴ Associate Professor in Nursing Science. Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

e-mail: gianluca.pucciarelli@uniroma2.it

⁵ Assistant Professor in Nursing Science. Department of Nursing, Faculty of Health Sciences, University of Jaume I, Castelló, Spain

e-mail: bernalte@iji.es

⁶ Assistant Professor in Nursing Science. Department of Maternal and Child Health Promotion, Internal Medicine & Specialties of Excellence “G. D’Alessandro” (PROMISE), Palermo (PA). Italy

e-mail: mariachiara.figura@unipa.it

Abstract 1 (in English)

Osteoporosis is a chronic skeletal condition that significantly affects everyday life and requires ongoing self-care (Johnston & Dagar, 2020; Kanis et al., 2021). Self-care can be understood as a naturalistic process comprising multiple behaviours shaped by contextual and cultural factors, as well as perceived barriers and facilitators (Riegel et al., 2012, 2026). This study explores the lived experiences of people with osteoporosis to understand how the condition is discursively constructed in everyday life.

A cross-national qualitative study was conducted using Automatic Analysis of Textual Data within the framework of Exploratory Multidimensional Data Analysis (Bolasco, 2021; Fraire, 2009). Forty people with osteoporosis (38 women and 2 men; mean age 68.6 years, SD = 5.74) were recruited in Italy (n = 20) and Spain (n = 20). Participants took part in in-depth semi-structured interviews focusing on illness experiences, self-care practices, and perceived barriers and facilitators. A corpus of 40 interviews was analysed using IRaMuTeQ. The analytical procedure combined descending hierarchical classification, correspondence analysis, and similitude analysis to identify the main discursive structures emerging from the corpus. The analysis revealed seven lexical classes organised into three broader thematic domains: clinical management of osteoporosis, embodied illness experience, and everyday self-care practices. The factorial structure highlighted the semantic oppositions between biomedical discourse and subjective bodily experience, as well as between medically supervised treatment and self-directed health practices. Similitude analysis further showed how these dimensions are interconnected within the lexical network of the narratives. Overall, the findings demonstrate the potential of multidimensional textual analysis to uncover the semantic organisation of patient discourse and to provide new insights into how individuals interpret and manage chronic illness in everyday life.

Keywords: Self-care behaviours, Osteoporosis narratives, Exploratory Multidimensional Data Analysis, Lived experience

1. Introduction

The increasing availability of textual data in health research has stimulated interest in analytical approaches capable of examining discourse as a structured linguistic system, enabling exploration of large corpora and identification of latent semantic structures.

Patient narratives constitute an important domain for such analyses. Individuals living with chronic illnesses continuously interpret and negotiate biomedical knowledge, bodily sensations, and everyday practices, offering insights into how they understand their conditions and organise self-care behaviours.

Osteoporosis represents a particularly relevant case. As a chronic skeletal disease associated with increased fracture risk, it requires long-term management and sustained engagement in preventive behaviours (Hemmati et al., 2021; Johnston & Dagar, 2020; Kanis et al., 2021; Tarantino et al., 2017), integrating pharmacological treatment, fall-prevention strategies, physical activity, and dietary adjustments (Tedesco et al., 2025). *These behaviours can be associated to self-care, a naturalistic process made of three dimensions shaped by contextual and cultural factors, as well as perceived barriers and facilitators (Riegel et al., 2012, 2026).* While biomedical aspects have been widely investigated, the discursive structures through which patients narratively construct their illness experience remain relatively unexplored. Lexicometric approaches enable researchers to examine the semantic organisation of narratives and identify linguistic patterns through which health experiences are articulated (Figura et al., 2024).

This study, therefore, explored the semantic structure of patient narratives about osteoporosis and self-care behaviours through multidimensional statistical analysis of an interview corpus.

2. Methods

2.1 Data Collection

Participants were recruited through purposive sampling from outpatient clinics specialising in osteoporosis management in Italy and Spain. Eligibility criteria included a confirmed diagnosis of primary osteoporosis, full cognitive orientation, and absence of linguistic barriers. Individuals with secondary osteoporosis, osteopenia, or cognitive impairment were excluded. Data collection took place between March and June 2023 in Italy and between March and May 2024 in Spain. Ethical approval was obtained in both national contexts prior to recruitment. Semi-structured interviews were conducted remotely via video call by trained native-speaking researchers in participants' native languages. The interview guide was developed in accordance with the Middle-Range Theory of Self-Care of Chronic Illness (Riegel et al., 2012, 2026) and comprised open-ended questions exploring self-care maintenance, monitoring, and management. Each interview lasted approximately 30 minutes and was audio-recorded with participants' informed consent. All recordings were transcribed verbatim and anonymised.

2.2 Corpus Construction and Characteristics

To ensure cross-national comparability while preserving narrative meaning, all interviews were translated into English by bilingual experts following forward and backward translation procedures (Chen & Boore, 2010; Cuoco et al., 2022). Translated transcripts were merged into a single corpus and formatted according to IRaMuTeQ standards (Ratinaud, 2024). Metadata variables were embedded at the beginning of each interview section to enable software-assisted grouping during analysis. Transcripts were then segmented prior to analysis, and corpus reliability was assessed against Bolasco's (2021) criteria for lexical balance, which require a minimum of 25,000 occurrences, a TTR below 20%, and a hapax proportion below 50%.

The corpus consisted of 40 interview texts segmented into 1,536 text segments, totalling 53,937 word occurrences and 2,677 distinct lexical forms, with an average segment length of 35.1 words.

Among lexical forms, 993 were hapax legomena (37.09% of vocabulary), a proportion consistent with semi-structured interview corpora, reflecting the balance between individual lexical variability and shared thematic coherence. The type–token ratio (TTR) was 0.05, which, although relatively low in absolute terms, is typical of interview-based corpora characterised by thematic concentration and recurrent reference to core concepts. The descending hierarchical classification successfully retained 1,428 text segments (92.97% of the corpus), indicating high lexical structuration and internal coherence (Table 1).

Table 1: General characteristics of the textual corpus.

Indicator	Value
Number of texts	40
Text segments	1,536
Word occurrences (tokens)	53,937
Lexical forms (types)	2,677
Hapax legomena	993
Hapax proportion	37.09%
Average segment length	35.1 words
Type–Token Ratio (TTR)	0.05

Lexical frequency distribution followed Zipf’s law (Bolasco, 2021), displaying the typical pattern of natural language corpora: a limited set of highly frequent forms and a progressively larger set of low-frequency items. This reflects the coexistence of a stable lexical nucleus associated with core interview themes and a long tail of low-frequency forms representing individual narrative variation. This distribution confirms the linguistic validity of the corpus and its suitability for multidimensional statistical analysis.

2.3 Data Analysis

Textual data were analysed using IRaMuTeQ (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires), an open-source software built on the R statistical environment for multidimensional statistical analysis of textual corpora (Ratinaud, 2024).

Prior to analysis, transcripts were standardised and pre-processed. Text segmentation was performed automatically by IRaMuTeQ according to the software’s default segmentation parameters, generating text segments of approximately 40 words based on punctuation and internal lexical boundaries.

Descriptive textometric indicators were calculated to characterise the corpus, including tokens, types, hapax legomena, average segment length, and the type–token ratio (TTR) as a measure of lexical diversity. Lexical frequency distribution was examined according to Zipf’s law, providing an indication of the linguistic validity of the dataset.

To identify the main discursive structures, a descending hierarchical classification (DHC) based on the Reinert (1990) method was performed. This procedure groups text segments by lexical profile similarity, using chi-square statistics to identify forms most associated with each class and producing a hierarchical dendrogram of the principal discursive configurations. A correspondence analysis (CA) was then conducted on the contingency table derived from the DHC, projecting lexical classes into a multidimensional factorial space to identify the principal semantic oppositions structuring the corpus.

Finally, a similitude analysis based on graph theory examined lexical co-occurrence patterns, reconstructing the network structure of the discourse and highlighting central nodes and clusters of semantically related terms.

Together, these procedures provide a multidimensional representation of the corpus, enabling identification of thematic structures, semantic oppositions, and relational lexical networks within participants' narratives.

3. Results

3.1 Hierarchical Descending Classification

The descending hierarchical classification produced seven lexical classes, representing distinct discursive configurations structuring the corpus and participants' narratives about osteoporosis and its management. The classification successfully assigned 1428 out of 1536 text segments (92.97%), indicating strong internal coherence of the lexical partitions (Figure 1, Table 2).

Table 2: Indicators of HDC

Indicator	Value	Description
Total text segments	1,536	Segments generated during corpus segmentation
Classified segments	1,428	Segments assigned to lexical classes
Classification rate	92.97%	Percentage of segments included in the classification
Number of lexical classes	7	Classes identified through hierarchical clustering

The dendrogram reveals a two-level hierarchical organisation of the corpus. At the first level, the corpus separates into two macro-discursive domains. At the first level, the corpus separates into two macro-discursive domains. The first macro-domain includes Classes 6, 7, 2, and 3, encompassing narratives related to healthcare engagement, lifestyle regulation,

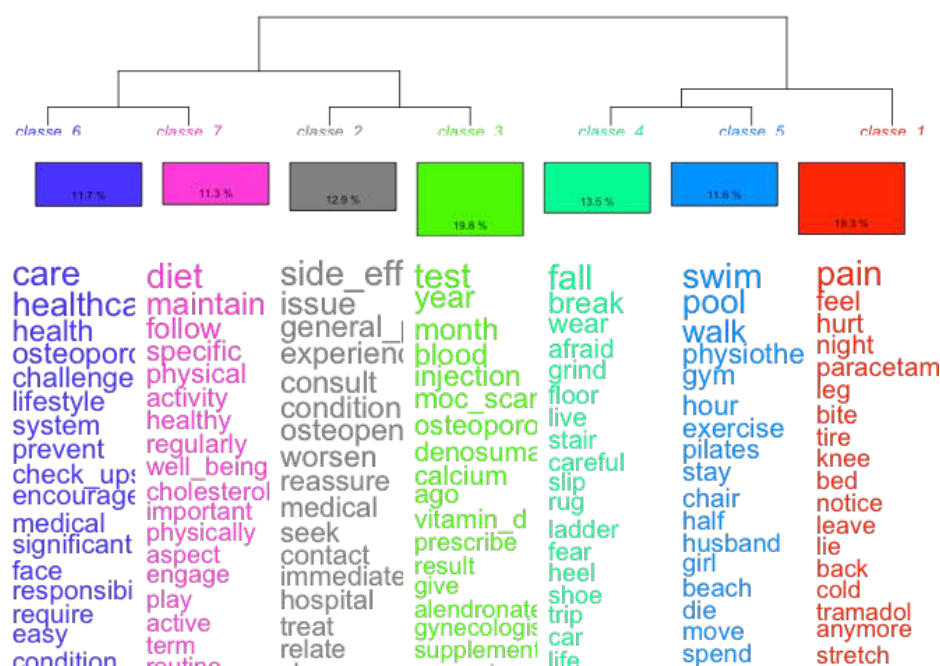


Figure 1: Hierarchical descending classification

participants' experience of living with the condition, and diagnostic and pharmacological pathways. The second macro-domain includes Classes 4, 5, and 1, grouping narratives centred

on fall prevention, physical activity practices, and the embodied experience of pain and discomfort.

Class 3 (19.8%) was characterised by lexical forms such as *test*, *year*, *blood*, *injection*, *moc_scan*, *calcium*, and *denosumab*, reflecting narratives centred on diagnostic and therapeutic pathways. Class 2 (12.9%) included *side_effects*, *issue*, *experience*, and *general_practitioner*, reflecting participants' accounts of living with the condition and negotiating medical care. Class 6 (11.7%) was structured around *care*, *healthcare*, *check_ups*, and *prevent*, reflecting engagement with the healthcare system. Class 7 (11.3%) included *diet*, *maintain*, *follow*, and *physical*, emphasising dietary regulation and lifestyle adjustments. Class 4 (13.5%) was structured around *fall*, *break*, *afraid*, and *live*, highlighting participants' perception of vulnerability and strategies aimed at preventing accidents. Class 5 (11.6%) included *swim*, *pool*, *walk*, and *physiotherapist*, reflecting physical activity practices adopted to maintain mobility and bone health. Class 1 (19.3%) was characterised by *pain*, *feel*, *hurt*, and *back*, capturing the embodied experience of the disease and the management of physical discomfort.

Taken together, the seven classes reveal a structured discursive system in which three broader thematic domains can be identified: the clinical management of osteoporosis, the embodied experience of illness, and everyday self-care practices (Figure 1).

Table 3: Lexical classes identified through HDC

Class	% of text segments	Most characteristic lexical forms (χ^2)	Interpretation
Class 1	19.3%	pain, feel, hurt, back	Embodied experience of pain
Class 2	12.9%	side_effects, issue, experience, general_practitioner	Living with the condition and negotiating care
Class 3	19.8%	test, year, blood, injection, moc_scan, calcium, denosumab	Diagnostic and therapeutic pathway
Class 4	13.5%	fall, break, afraid, live	Fracture risk perception and fall prevention
Class 5	11.6%	swim, pool, walk, exercise	Physical activity practices
Class 6	11.7%	care, healthcare, check_ups, prevent	Healthcare engagement and medical follow-up
Class 7	11.3%	diet, maintain, follow, physical, activity	Lifestyle and self-care routines

3.2 Correspondence Analysis (CA)

To explore the multidimensional semantic organisation of the corpus, the lexical classes identified through descending hierarchical classification were projected into a factorial correspondence space through correspondence analysis (CA). The analysis generated six factorial axes, of which the first two were retained for interpretation because they accounted for the largest proportion of total inertia and provided the most interpretable semantic configuration of the corpus. The first factorial axis accounted for 25.33% of the total inertia, while the second axis accounted for 19.90%, together representing 45.23% of the total inertia (Figure 2).

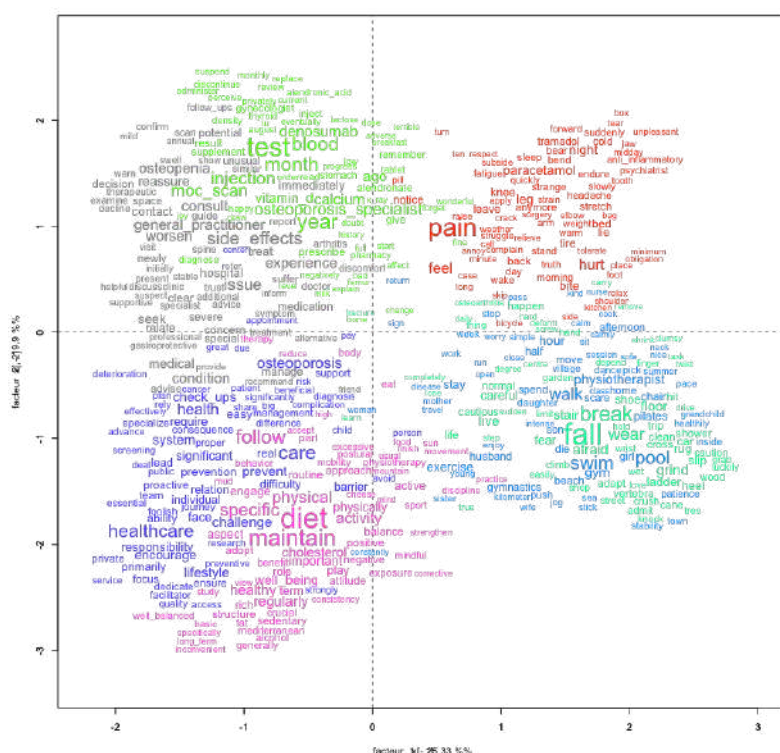


Figure 2: Correspondence analysis factorial plane of lexical classes

3.2.1 Axis 1: Biomedical management vs embodied illness experience

The first factorial axis (25.33%) differentiates discourse centred on the clinical and biomedical management of osteoporosis from discourse focused on the embodied experience of illness. On the left side of the factorial space, lexical forms such as *test*, *blood*, *injection*, *moc_scan*, *calcium*, and *denosumab* characterise classes associated with diagnostic procedures, pharmacological treatments, and medical monitoring, alongside narratives describing interactions with healthcare professionals and lifestyle regulation. In contrast, the right side of the axis is dominated by lexical forms such as *pain*, *feel*, *hurt*, *fall*, *break*, *swim*, and *walk*, reflecting participants' descriptions of bodily sensations, physical discomfort, fall prevention strategies, and physical activity practices.

3.2.2 Axis 2: Reactive illness management vs proactive health maintenance

The second factorial axis (19.90%) differentiates discourse centred on the reactive management of illness symptoms and clinical follow-up from discourse oriented towards proactive health maintenance and preventive practices. In the upper region of the factorial plane, lexical forms such as *pain*, *hurt*, *side_effects*, *experience*, *test*, *blood*, and *injection* reflect narratives centred on the lived experience of symptoms, diagnostic pathways, and therapeutic monitoring. In the lower region, lexical forms such as *diet*, *maintain*, *healthcare*, *prevent*, *check_ups*, *swim*, *walk*, and *fall* highlight narratives focused on lifestyle regulation, physical activity, and strategies aimed at preventing deterioration and maintaining health.

A closer examination of the factorial plane reveals four semantically coherent quadrants, each corresponding to a specific discursive configuration. The upper-left quadrant is dominated by lexical forms associated with diagnostic procedures and pharmacological treatment, including *test*, *blood*, *injection*, *moc_scan*, and *calcium*, reflecting narratives focused on medical monitoring and therapeutic pathways. The upper-right quadrant is characterised by lexical forms such as *pain*, *hurt*, *knee*, and *back*, representing the embodied dimension of illness, where participants describe physical discomfort and symptomatic experiences. The lower-right

quadrant includes lexical forms such as *fall*, *afraid*, *break*, *walk*, and *swim*, highlighting narratives centred on risk awareness and behavioural adaptation, particularly strategies aimed at preventing fractures and maintaining mobility. Finally, the lower-left quadrant is characterised by lexical forms including *diet*, *maintain*, *physical*, and *healthcare*, reflecting everyday health maintenance practices and lifestyle regulation (Figure 3).

Taken together, the factorial structure shows that participants' narratives are organised around three interrelated thematic domains: clinical management of osteoporosis, embodied illness experience, and everyday self-care practices (Figure 3).

When nationality was introduced as an illustrative modality, Italian participants (*naz_ita*) were positioned on the left side of the factorial plane, in proximity to classes characterised by biomedical management and lifestyle regulation. Spanish participants (*naz_esp*) were located on the right side, closer to classes associated with embodied illness experience, fall prevention, and physical activity practices.

3.3 Similitude Analysis

Similitude analysis reveals the relational structure of lexical co-occurrences within the corpus. The network is organised around the central node osteoporosis, which acts as a lexical hub connecting several semantic substructures within the corpus. Four main lexical clusters emerge from this central node.

The first cluster relates to clinical monitoring and treatment, including lexical forms referring to medical tests, injections, and pharmacological therapies.

The second cluster concerns embodied illness experience, structured around lexical items associated with pain and bodily discomfort.

A third cluster highlights physical activity and mobility practices, including walking, swimming, and exercise routines that participants adopt as part of their self-care strategies.

A fourth cluster relates to dietary regulation and lifestyle management, reflecting participants' efforts to maintain health through nutritional choices and behavioural adjustments.

The structure of the network indicates that these domains are not isolated but interconnected through the central concept of osteoporosis, illustrating how participants integrate biomedical knowledge with everyday health practices in their narratives (Figure 3).

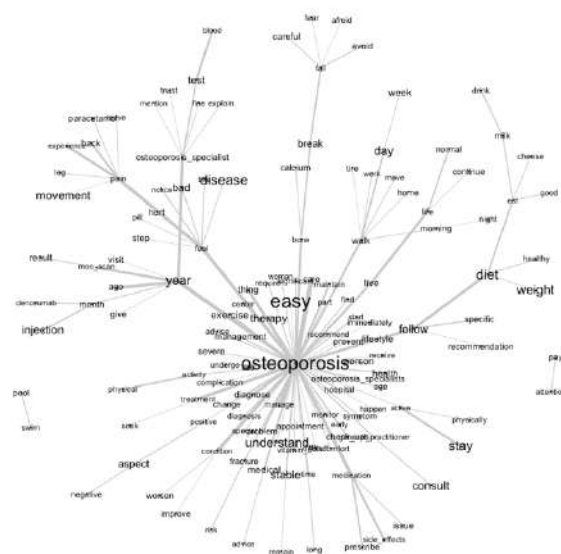


Figure 3: Similitude analysis

4. Discussion

The findings reveal that patient narratives about osteoporosis are organised around a structured semantic system in which biomedical discourse, embodied illness experience, and everyday self-care practices are dynamically interconnected. The similitude analysis corroborates and enriches this picture: the lexical network is organised around the central node *osteoporosis*, from which four main clusters emerge, clinical monitoring and treatment, embodied illness experience, physical activity, and dietary regulation, suggesting that participants do not experience these dimensions as separate domains but as an integrated whole in which clinical knowledge, bodily sensations, and everyday practices mutually inform one another. These discursive configurations can be interpreted in light of the self-care framework proposed by Riegel et al. (2012, 2026): narratives centred on follow up pathways, pharmacological management, and symptom interpretation reflect dimensions of self-care monitoring and management, while narratives oriented towards physical activity, dietary regulation, and fall prevention align with self-care maintenance. This suggests that people with osteoporosis simultaneously navigate medically supervised care and autonomous health practices, integrating them within a unified illness narrative.

The correspondence analysis further highlights two fundamental semantic oppositions structuring this narrative space, biomedical management vs embodied illness experience, and reactive illness management vs proactive health maintenance, with direct clinical relevance. Healthcare professionals working with people with osteoporosis should be aware that patients construct their illness experience across both dimensions simultaneously, and that interventions focused exclusively on pharmacological compliance may fail to address the broader self-care landscape patients navigate in everyday life.

The cross-national dimension adds a further layer of interpretive depth. Italian participants were discursively oriented towards biomedical management and clinical monitoring, while Spanish participants showed a stronger emphasis on physical activity, fall prevention, and proactive lifestyle practices. Consistent with Riegel et al. (2026), this suggests that self-care behaviours are shaped by contextual and cultural factors, warranting further investigation in future cross-national studies.

From a methodological perspective, the combination of descending hierarchical classification, correspondence analysis, and similitude analysis demonstrated the potential of multidimensional textual analysis to uncover latent semantic structures in patient narratives, offering a valuable complement to conventional qualitative approaches in chronic illness research.

5. Conclusion

This study illustrates the potential of multidimensional statistical analysis of textual data for exploring the semantic structures underlying patient narratives in chronic illness. The integration of descending hierarchical classification, correspondence analysis, and similitude analysis allowed the identification of distinct discursive configurations in which clinical management, embodied illness experience, and everyday self-care practices appear dynamically interconnected. The cross-national sampling further suggested that contextual and cultural factors may shape the discursive organisation of self-care, with Italian and Spanish participants showing distinct narrative orientations that can be interpreted in light of the role of context in self-care behaviours.

Some limitations should be acknowledged. The translation of interviews into a common language for IRaMuTeQ analysis, although conducted following rigorous procedures, may have introduced lexical losses affecting the textometric results. Additionally, participants were recruited from public hospital outpatient clinics in two specific national contexts, which may limit the transferability of findings to broader populations. Future research should explore self-

care narratives across wider national and cultural contexts, considering multilingual corpus analysis approaches to preserve the richness of participants' original language.

References

- Bolasco, S. (2021). L'analisi automatica dei testi. Fare ricerca con il Text Mining. Carrocci Editore.
- Chen, H., & Boore, J. R. (2010). Translation and back-translation in qualitative nursing research: methodological review. *Journal of Clinical Nursing*, 19(1–2), 234–239. <https://doi.org/10.1111/j.1365-2702.2009.02896.x>
- Cuoco, A., Arcadi, P., Figura, M. C., Piervisani, L., Alvaro, R., Vellone, E., & Durante, A. (2022). Designing and conducting qualitative research across countries and cultures: challenges for inclusiveness and rigour. *European Journal of Cardiovascular Nursing*, 21(8), 873–879. <https://doi.org/10.1093/eurjcn/zvac067>
- Figura, M., Arcadi, P., Vellone, E., Pucciarelli, G., Simeone, S., Piervisani, L., & Alvaro, R. (2024). Living in a multicultural context: Health and integration from the perspective of undocumented Mediterranean migrants, residents, and stakeholders in Italy. A qualitative-multimethod study. *Journal of Advanced Nursing*, 80(8), 3382–3394. <https://doi.org/10.1111/jan.16036>
- Fraire, M. (2009). Statistical methods for exploratory multidimensional data analysis on time use. *Statistica*, 69(4), 317–341
- Hemmati, E., Mirghafourvand, M., Mobasser, M., Shakouri, S. K., Mikaeli, P., & Farshbaf-Khalili, A. (2021). Prevalence of primary osteoporosis and low bone mass in postmenopausal women and related risk factors. *Journal of Education and Health Promotion*, 10, 204. https://doi.org/10.4103/jehp.jehp_945_20
- Johnston, C. B., & Dagar, M. (2020). Osteoporosis in Older Adults. *The Medical Clinics of North America*, 104(5), 873–884. <https://doi.org/10.1016/j.mcna.2020.06.004>
- Kanis, J. A., Norton, N., Harvey, N. C., Jacobson, T., Johansson, H., Lorentzon, M., McCloskey, E. V., Willers, C., & Borgström, F. (2021). SCOPE 2021: a new scorecard for osteoporosis in Europe. *Archives of Osteoporosis*, 16(1), 82. <https://doi.org/10.1007/s11657-020-00871-9>
- Ratinaud, P. (2024). IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [software] (Version 0.7 alpha 2). (2009) Available from: <http://www.iramuteq.org>. In <http://www.iramuteq.org/>
- Reinert, M. (1990). Alceste, une méthodologie d'analyse des données textuelles et une application: Aurelia de Gerard de Nerval. *Bulletin de Méthodologie Sociologique*, 26(1), 24–54.
- Riegel, B., Jaarsma, T., & Strömberg, A. (2012). A middle-range theory of self-care of chronic illness. *ANS. Advances in Nursing Science*, 35(3), 194–204. <https://doi.org/10.1097/ANS.0b013e318261b1ba>
- Riegel, B., Jaarsma, T., & Strömberg, A. (2026). An update to the middle-range theory of self-care of chronic illness. *ANS. Advances in Nursing Science*, 49(1), 14–25. <https://doi.org/10.1097/ANS.0000000000000594>
- Tarantino, U., Iolascon, G., Cianferotti, L., Masi, L., Marcucci, G., Giusti, F., Marini, F., Parri, S., Feola, M., Rao, C., Piccirilli, E., Zanetti, E. B., Cittadini, N., Alvaro, R., Moretti, A., Calafiore, D., Toro, G., Gimigliano, F., Resmini, G., & Brandi, M. L. (2017). Clinical guidelines for the prevention and treatment of osteoporosis: summary statements and recommendations from the Italian Society for Orthopaedics and Traumatology. *Journal of Orthopaedics and Traumatology*, 18(S1), 3–36. <https://doi.org/10.1007/s10195-017-0474-7>
- Tedesco, C., Bernalte-Martí, V., Pucciarelli, G., Vellone, E., Basilici Zannetti, E., Cittadini, N., Pennini, A., Tarantino, U., & Alvaro, R. (2025). Self-care experiences and behaviors in people with osteoporosis: A meta-synthesis. *Maturitas*, 195, 108213. <https://doi.org/10.1016/j.maturitas.2025.108213>

Text Mining Applications in the Evaluation of AIFA Drug Innovativeness Reports in Italy

Carlotta Galeone¹, Paolo Mariani², Andrea Marletta³, and Daniele Pirotta⁴

¹ Biostatistics & Outcome Research, Statinfo – carlotta.galeone@statinfo.org

² University of Milano-Bicocca – paolo.mariani@unimib.it

³ University of Milano-Bicocca – andrea.marletta@unimib.it

⁴ University of Milano-Bicocca – daniele.pirotta@unimib.it

Abstract

The evaluation of drug innovativeness is a critical regulatory process that determines patient access to therapies and provides financial incentives for pharmaceutical companies. Since 2017, the Italian Medicines Agency (AIFA) has managed this process through structured assessment reports based on three fundamental domains: therapeutic need, added clinical value, and quality of evidence. This study applies textual analysis to reports published between 2017 and 2025 to explore linguistic patterns and thematic shifts, while identifying the primary predictors of regulatory outcomes. The research adopts a multi-method text mining approach consisting of several phases: a pre-processing stage for text extraction and cleaning; an exploratory corpus analysis; and a keyword coding phase utilizing a dedicated dictionary. The analytical framework includes descriptive statistics, frequency analysis, word clouds, Semantic Network Analysis (SNA), Latent Dirichlet Allocation (LDA) for topic modeling, and Correspondence Analysis (CA) to map latent associations between keywords, domains, and regulatory decisions. Each method was employed to systematically compare the three assessment domains and final outcomes (Approved vs. Not Approved Innovation). Preliminary findings reveal a fundamentally similar linguistic structure between approved and rejected drugs. However, distinctive features emerge suggesting that Quality of Evidence and Added Therapeutic Value serve as the primary drivers of the institutional discourse. Furthermore, the longitudinal analysis tracked the evolution of the most influential terms and topics over the years.

Keywords: AIFA, Drug Innovativeness, Text Mining, Topic Modeling, Correspondence Analysis.

1. Introduction

In 2017, the Italian Medicines Agency (AIFA) began systematically publishing innovation assessment reports for newly licensed medicinal products seeking specialised status. This evaluative process is governed by a tripartite framework encompassing therapeutic necessity, incremental clinical benefit and the robustness of scientific evidence. While these assessments primarily serve a regulatory purpose, influencing reimbursement negotiations, expedited market entry and the allocation of dedicated national funds, they also represent a substantial yet underutilised repository of institutional knowledge. These documents provide an unparalleled insight into the linguistic and conceptual construction of Innovation within a highly specialised regulatory environment. Furthermore, between 2023 and 2024, the responsibility for these evaluations transitioned from the Technical-Scientific Commission (CTS) to the new Scientific and Economic Commission (CSE). Beyond their regulatory role, these reports constitute a rich information source on how pharmaceutical innovation is conceptualized and communicated. However, to date, these documents have not been systematically analyzed using text mining

techniques. For these reasons, the strategic value of this body of work has remained largely unexplored using computational linguistics or automated text mining methodologies. The present study addresses this gap by systematically analysing the entire AIFA archive spanning 2017–2025. Transitioning from a granular review of individual clinical dossiers to comprehensive, large-scale textual analysis, this research identifies persistent linguistic trends and latent thematic structures. Ultimately, this work clarifies the institutional logic and semantic nuances that distinguish Innovative from Not Innovative status, providing valuable insights into the criteria that define pharmaceutical excellence within the Italian healthcare landscape. The analyses carried out in this study were performed using the R statistical programming language (R Core Team, 2025). Section 2 explains the methodologies used, Section 3 presents the obtained results, and Section 4 summarises the conclusions and outlines future work. This project is a collaboration between B-ASC (Bicocca Applied Statistics Center) of the University of Milano-Bicocca and the ISPOR Italy Rome Chapter, which provides methodological and scientific support.

2. Methodology

The research architecture follows a multi-stage computational linguistics pipeline, specifically designed to extract and quantify institutional logic from a corpus of 294 AIFA innovation reports (AIFA, 2026). The workflow transitions from fundamental corpus characterisation to advanced multivariate mapping, as detailed below.

The initial phase involved constructing a structured database of all reports published between 2017 and 2025. To ensure textual standardization, a rigorous pre-processing protocol was implemented (Lebart and Salem, 1994). Specifically, the corpus was pre-processed by converting the text to lowercase, removing punctuation and numbers, and applying stemming and lemmatization.

The texts were then tokenised and transformed into a document–term matrix (DTM) using the *tm* package (Feinerer and Hornik, 2024), where the rows represent documents and the columns represent terms, with the entries corresponding to term frequencies. Rare terms were optionally removed to reduce sparsity. This matrix forms the basis for subsequent analytical stages, enabling the application of multivariate statistical techniques to textual data.

A Frequency Analysis was conducted to capture the 'surface' vocabulary of the regulatory body. This step provides a quantitative inventory of the most prevalent terms within the three evaluation domains: therapeutic need, clinical value and evidence quality. These frequencies were visualized through Word Clouds using the *wordcloud* package (Fellows, 2018), where font size is proportional to term recurrence, offering an immediate and intuitive synthesis of prevailing institutional priorities (Heimerl et al., 2014), and also in the form of bar charts.

In order to move beyond isolated words and examine the 'architecture of meaning', the study employed Semantic Network Analysis (SNA). This technique treats terms as nodes and their co-occurrences as edges. Specifically, we analysed bigrams and trigrams (sequences of two or three adjacent words) via the *tidytext* framework (Silge and Robinson, 2016) to identify stable conceptual formulas and technical idioms. By mapping these linguistic strings, SNA reveals the syntactic links between specific clinical outcomes and evaluative judgements (Wiedemann, 2013).

For unsupervised theme extraction, Latent Dirichlet Allocation (LDA) was applied using the `topicmodels` package (Grün and Hornik, 2011) only considering unigrams. This generative probabilistic model identifies the primary themes within the corpus by statistically detecting sets of terms that frequently co-occur, grouping them into coherent topics. Within this framework, each assessment report is treated as a multi-topic mixture, while each topic is defined by a distribution of words with associated probabilities. The analysis focused on characterising topics through per-topic-per-word probabilities (β) and evaluating thematic prevalence over the 2017–2025 period using per-document-per-topic probabilities. This approach objectively uncovers "latent" thematic clusters — such as orphan drug designation or statistical robustness — characterising the AIFA discourse without the bias of a priori coding while providing a quantitative basis for longitudinal analysis.

Finally, Correspondence Analysis (CA) was used to examine the relationship between linguistic markers and regulatory outcomes. When applied to the Document-Term Matrix (DTM), utilizing the `FactoMineR` package (Lê et al., 2008) CA acts as a multivariate dimensionality reduction technique. By projecting terms and metadata (e.g. Approved vs. Not Approved status) onto a low-dimensional Cartesian plane, it provides a visual semantic map. This synthesis reveals which keywords are statistically overrepresented in successful innovativeness claims compared to rejected ones, effectively mapping the 'language of success' in the Italian regulatory landscape.

3. Results

The final dataset comprised 294 assessment reports. Descriptive analysis indicates a relatively balanced distribution of regulatory outcomes, with 166 (56% of claims) achieving innovativeness status, compared to 128 (44%) that were not approved. On average, each report contained 854 words, with a marginal difference in documentation density between the Approved (860 words) and Not Approved (845 words) cohorts. As illustrated in Figure 1, longitudinal analysis reveals a significant increase in the depth of institutional documentation. The average word count per record almost doubled during the study, increasing from 490 in 2017 to 995 in 2025, but with no significant differences among approved and not approved reports. This growth was particularly pronounced between 2019 and 2021, marking a shift towards more exhaustive and detailed regulatory discourse. This is likely to reflect the increasing complexity of clinical evidence and the maturation of AIFA's evaluative framework.

A rigorous pre-processing phase, including text cleaning, tokenization, stop-word removal, and term normalization has been applied on the final dataset to ensure textual standardization. Moreover, a keyword coding phase utilizing a dedicated dictionary has been proposed. This keyword normalization is a technique used in the text mining analysis to recognise and unify different terms explaining the same concept. It is of help to reduce the linguistic variability and to improve the semantic coherence of the corpus. In this study, 47 keywords have been selected among the most significant terms influencing the innovativeness of the drug. Once data cleaning and keyword coding were complete, the frequency analysis revealed a high degree of structural similarity in the distribution of key terms across different evaluative categories.

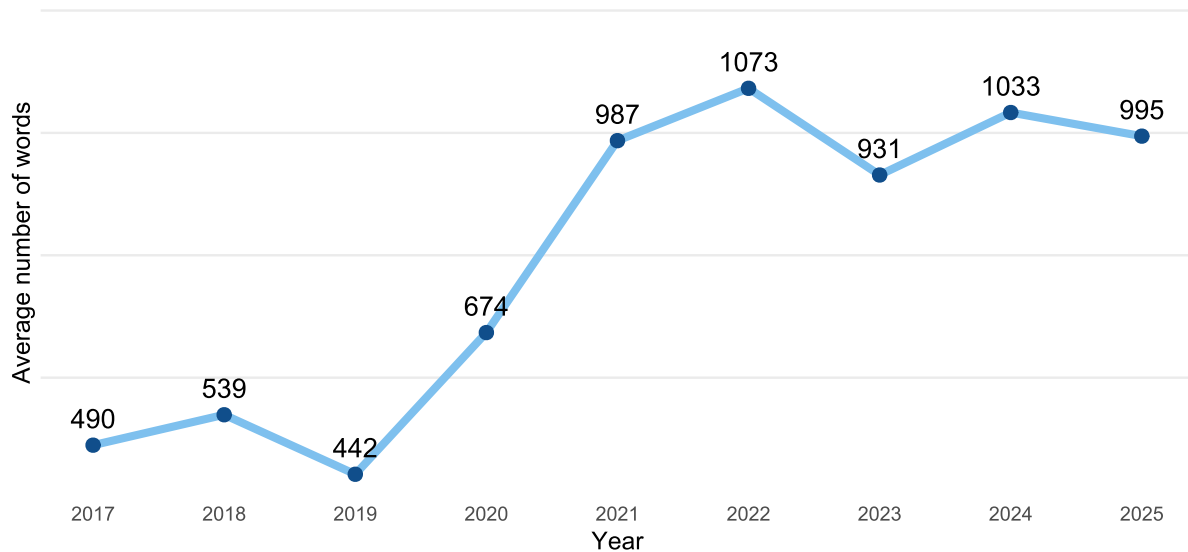


Figure 1: Average number of words in an AIFA innovation report from 2017 to 2025 (n=294).

Figure 2 illustrates this, displaying the twenty most frequent keywords stratified by innovativeness status. The overarching linguistic architecture remains largely consistent between groups. Notably, reports associated with a Not approved status frequently feature the keyword 'indirect comparison', which is a fundamental methodology in pharmacological studies. This statistical prominence suggests that, although the core vocabulary of regulatory discourse is shared, specific methodological markers differentiate unsuccessful innovation dossiers from successful ones.

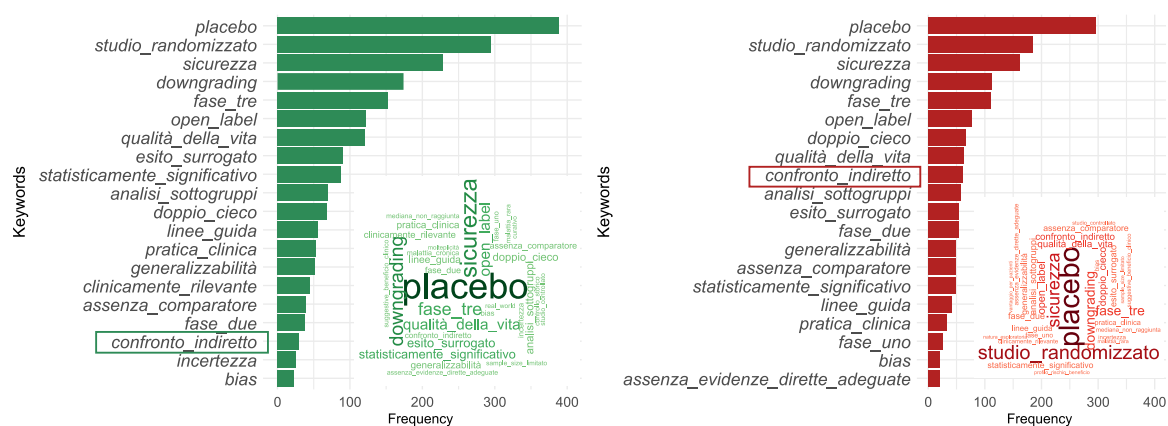


Figure 2: Top Keyword frequency by Innovation approved (n=166; green), vs Innovation Not Approved (n=128; red) from 2017 to 2025.

Semantic Network Analysis (SNA) reveals distinct conceptual clusters and an insightful divergence in n-gram structures between the two regulatory categories. Given the structural homogeneity between the approved and non-approved report corpora, the most compelling insights within the SNA framework emerge from nodes that occur exclusively under one specific status. As shown in Figure 3, the network topology of the Approved Innovation sub-corpus reveals a close integration of terms centred on 'significant therapeutic benefit' and 'robust clinical evidence'. The discourse in these reports is often anchored by key phrases such as 'currently available' and 'overall survival', suggesting a strong linguistic association between favourable outcomes and specific clinical endpoints. In contrast, the semantic network for the Not Approved group, also shown in Figure 3, reveals comparatively weaker connections between clinical results and therapeutic requirements. Here, the institutional language shifts towards more cautious or neutral terminology, primarily utilising expressions such as 'clinical benefit' and 'assessable impact'. This divergence in the relational architecture of the text highlights that the connectivity between evaluative domains, rather than the frequency of individual words, is a primary factor in the AIFA decision-making process.

Figure 3: Network of the most frequent bigrams by innovation status Approved (n=166; green) and Not approved (n=128; red), 2017–2025 AIFA report corpus

The application of Latent Dirichlet Allocation (LDA) successfully classified the specialised regulatory jargon into three primary thematic categories: 'Study Design', 'Clinical Outcomes' and 'Therapeutic Treatment'. The use of Coherence Score analysis or Perplexity plots suggested the number of topics equal to 3 or 5, results for three categories are presented in Figure 4. These clusters closely align with the tripartite evaluative framework used by AIFA. As shown in Figure 4, the consistent presence of key terms such as 'patients', 'study' and 'treatment' across the entire corpus of reports confirms a stable structural baseline. A granular analysis of the individual topics reveals distinct institutional focuses. The 'Study Design' topic focuses heavily on the experimental and methodological rigour of clinical trials, emphasising the robustness of the submitted evidence. The 'Clinical Outcomes' cluster prioritises efficacy endpoints and the statistical significance of the results obtained. Finally, the 'Therapeutic Treatment' theme addresses the practical aspects of pharmacological intervention, including administration protocols and real-world clinical application. This thematic stratification, visualised through the distribution of the most important keywords, shows how the LDA model effectively captures the multifaceted nature of the innovation assessment process.

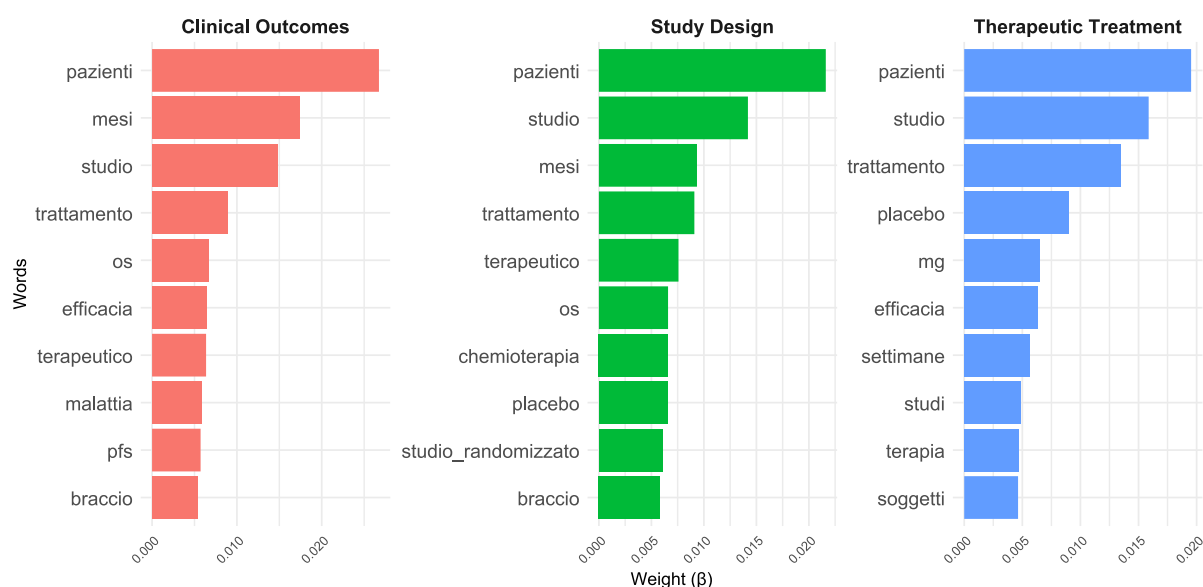


Figure 4: Latent Dirichlet Allocation (LDA) topic clusters ($k=3$) within the 2017–2025 AIFA report corpus ($n=294$).

Applying Correspondence Analysis (CA) enabled the creation of a comprehensive semantic map, projecting the most frequent keywords onto a two-dimensional Cartesian plane. This spatial representation is defined by two primary axes, which are derived from extracting the first two principal components of a factor analysis based on keyword frequency distributions within the document-term matrix.

As the current research is still in its preliminary phase, we hypothesise that these orthogonal dimensions reflect 'study solidity' and the 'uncertainty level' inherent in the regulatory evidence. As can be seen in the resulting map (Figure 5), the positioning of specific terms in relation to these axes provides a visual summary of how institutional discourse balances methodological rigour with evidentiary gaps. This multivariate approach effectively clarifies the latent associations between terminological clusters and the overall evaluative landscape of pharmaceutical innovation in Italy. As a limitation of the study, the first two principal components explain a very low percentage of the variance, but this is a very common situation when correspondence analysis involves terms because of sparsity of the distribution of the words. Nevertheless, the choice of representing only the first two components is useful for a better visualization of the results.

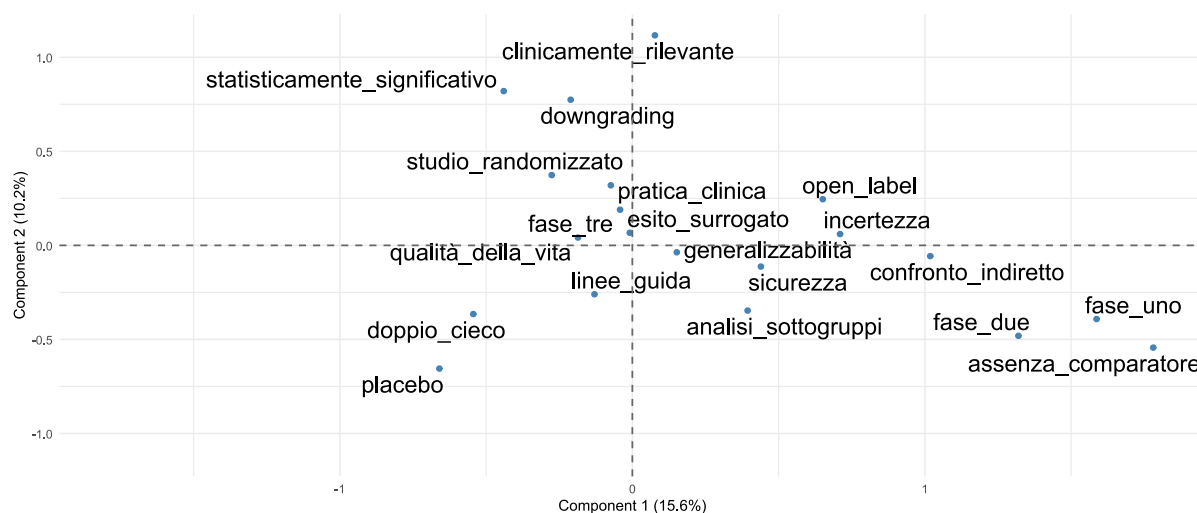


Figure 5: Correspondence Analysis (CA) semantic map of high-frequency keywords, 2017–2025 AIFA report corpus ($n=294$).

4. Conclusions

This study demonstrates that text mining is a powerful and innovative analytical framework for interpreting the complex language of pharmaceutical regulation. By analysing a corpus of 294 AIFA reports published between 2017 and 2025, the research reveals how the concept of institutional 'innovation' has changed over the past decade. Although preliminary findings suggest a broadly similar semantic structure between approved and rejected drugs, specific keyword coding and frequency analysis reveal distinctive linguistic markers that differentiate the two groups, like the highest presence of the keyword 'indirect comparisons' in not approved reports.

Integrating Network and Correspondence Analysis further clarifies these distinctions, revealing word clusters that reflect underlying levels of evidence robustness and regulatory uncertainty. Specifically, the prominence of terms related to 'quality of evidence' and 'added therapeutic value' confirms these domains as the primary drivers of institutional discourse. Moving forward, the research will expand to include more granular topic modelling and focused analyses of specific pathological areas, paying particular attention to methodological indicators such as 'indirect comparisons'. This methodological approach ultimately paves the way for the development of predictive tools in drug evaluation, fostering greater transparency and strategic alignment for all stakeholders within the Italian healthcare landscape.

References

- Agenzia Italiana del Farmaco (AIFA). (2026.). Farmaci innovativi. 2026, <http://aifa.gov.it/farmaci-innovativi>.
- Feinerer, I. and Hornik, K. (2024). tm: Text Mining Package. R package version 0.7-11
- Fellows, I. (2018). wordcloud: Word Clouds. R package version 2.6.
- Grün, B. and Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), 1–30.
- Heimerl, F., Lohmann, S., Lange, S. and Ertl, T. (2014): Word Cloud Explorations: Text Analysis Based on Word Clouds. In: *Proceedings of the 47th Hawaii International Conference on System Sciences*, pp. 1833-1842. *IEEE Computer Society, Washington*
- Kesselheim, A. S., Wang, B. and Avorn, J. (2013). Defining “innovativeness” in drug development: a systematic review. *Clinical Pharmacology & Therapeutics*, 94(3), 336-348.
- Lebart L. and Salem A. (1994). *Statistique textuelle*; In: Dunod Paris.
- Lê, S., Josse, J. and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1–18.
- Punukollu, M. (2023). NLP Applications in Mining Pharmaceutical Text Data for Competitive Intelligence and Market Insights. *Essex Journal of AI Ethics and Responsible Innovation*, 3, 594-635.
- R Core Team (2025). *_R: A Language and Environment for Statistical Computing_*. R Foundation for Statistical Computing, Vienna, Austria. <<https://www.R-project.org/>>.
- Silge, J. and Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software*, 1(3), 37.
- Wiedemann, G. (2013): Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences. *Historical Social Research*. 38, 332-357

Characterization of texts written by deaf people through syntactic markers

Mireille Esther Gettler Summa¹, Caroline Bogliotti², Anne Lacheret Dujour³

¹²³MODYCO, UMR 7114, University of Paris Nanterre France

Abstract

We aim to conduct a syntactic analysis excluding morphosyntax of a pioneering corpus of its kind in France, produced by deaf secondary school pupils aged between fourteen and sixteen. Using a data-driven approach, we aim to shed light on certain preconceptions regarding the writing of deaf students in French. The data comprises 19,458 words spread across 216 texts by 66 writers in four genres: narrative, descriptive, explanatory and argumentative. The pages, originally handwritten, are accompanied by metadata on the scriptors as well as syntactic annotations of the texts in relation to the standard of written work expected for the French Brevet des Collèges, selected following inter-annotator agreement. We supplement the syntactic annotations with new variables: to do this, we must first segment the texts appropriately before parsing them and feeding them into the relevant algorithms. We then have initial text datasets accompanied by their metadata and annotations, enriched with more general syntactic features such as their complexity scores. Analysis of this data reveals, through unsupervised classification, four scriptor profiles characterised by their syntactic specificities as well as their metadata (Sign Language at birth or not, deaf parents or not, etc.). We also observe that when we calculate the average syntactic dependence complexity measure for each class, the four profiles are ordered from least close to the French norm to closest. This result could be explained by a neurolinguistics perspective on the learning of written French by deaf people. All of these analyses confirm the results of the automatic classification: there is no homogeneous written syntax for deaf people despite their common disability; on the contrary, the study data show a diversification of their grammatical usage.

Keywords: Clustering, Corpus, Deaf, Syntactic Pattern, French Literacy

Résumé

Nous nous proposons de faire un inventaire de la syntaxe, morphosyntaxe exclue, d'un corpus premier du genre en France produit par des scripteurs sourds collégiens de quatorze à seize ans. Par une approche data driven nous cherchons à apporter un éclairage sur certains a priori de l'écrit des sourds en langue française. Les données comportent 19458 mots répartis en 216 textes sur 66 scripteurs dans quatre genres, narratif, descriptif, explicatif, argumentatif. Les pages initialement manuscrites sont accompagnés de métadonnées sur les scripteurs ainsi que d'annotations syntaxiques des textes en regard à la norme de l'écrit attendu pour le brevet des collèges et retenues après accord inter-annotateurs. Nous complétons les annotations syntaxiques par de nouvelles variables : il faut pour cela segmenter de façon appropriée les textes avant de les parser et de les soumettre aux algorithmes dédiés. Nous disposons alors des ensembles de données textuelles initiales accompagnées de leurs métadonnées et annotations, enrichis de caractéristiques syntaxiques plus générales comme leurs scores de complexités. L'analyse de ces données révèle par classification automatique quatre profils de scripteurs

caractérisés par leurs spécificités syntaxiques ainsi que par leurs métadonnées (Langue des Signes à la naissance ou pas, parents sourds ou pas, etc.). On observe également que lorsque l'on calcule la moyenne de la mesure de complexité syntaxique du flux de dépendance pour chaque classe les quatre profils sont ordonnés du moins proche à la norme du français au plus proche. Ce résultat trouverait une explication par un éclairage neurolinguistique de l'apprentissage de l'écrit par les sourds. L'ensemble de ces traitements confirme les résultats de la classification: il n'y a pas de syntaxe de l'écrit homogène pour les sourds malgré leur handicap commun, dans les données de l'étude il y a bien diversification de leurs usages grammaticaux.

Mots clés : Classification automatique, Corpus, Sourd, Syntaxe, Français Ecrit

1. Introduction

Nous présentons une analyse de la syntaxe, morphosyntaxe exclue des scripteurs sourds dans des textes écrits en langue française. La littératie des locuteurs sourds en français écrit est un sujet encore très peu abordé car les corpus de scripteurs sourds restent quasiment inexistant en France (Périni, 2013 ; Balosetti, 2011). C'est pourquoi il existe quelques a priori sur la syntaxe des sourds à l'écrit en français. Le français écrit est pour les sourds signeurs à double titre une L2 : la Langue des Signes Française (LSF) est une langue visuo-gestuelle et de plus sans écrit associé. Leurs productions sont donc un sujet intéressant pour les linguistes (Durand, 2003 ; Aleksandrova et al., 2024). L'approche utilisée de type 'corpus driven', guidée par les données, a été appliquée sur un corpus de littératie produit par 66 adolescents sourds, scolarisés dans le secondaire, dit corpus « Alpha » recueilli en 2019 (Prénovec, 2025). Le corpus Alpha comporte 216 textes produits par 66 adolescents sourds profonds, congénitaux et pré-linguaux. Il a fait l'objet d'une annotation collaborative experte des écarts produits par les scripteurs aux compétences attendues en syntaxe pour le brevet des collèges. Nous disposons également de 59 métadonnées comme l'âge ou la scolarisation des scripteurs. Nous allons dans un premier temps rassembler trois bases de travail : les textes, un tableau d'énoncés ciblés 'annotations en syntaxe' et un tableau de scripteurs décrits par plusieurs groupes de variables : des nouvelles variables construites en comptabilisant, pour chaque jeune, diverses caractéristiques d'annotations en syntaxe, en adjoignant des descripteurs linguistiques hors cadre normatif de ses écrits comme la parataxe, et enfin les métadonnées. On mettra alors en œuvre sur ces trois ensembles de données, les textes, leurs annotations expertes en terme d'écart à la norme attendue au brevet des collèges, et enfin les scripteurs, une approche typologique des scripteurs par classification automatique (Balbi, 2010 ; Summa, 2015) complétée par des calculs de scores de complexité syntaxique. On tirera enfin des premières hypothèses des résultats obtenus pour les développements futurs de l'étude.

2. LES VARIABLES CONSTRUITES A PARTIR DU CORPUS ALPHA

2.1. L'annotation des écarts à la syntaxe attendue

Le corpus Alpha a été annoté de façon experte manuellement et a fait l'objet d'accords inter annotateurs. Cette étape est décrite de façon détaillée dans (Lacheret-Dujour et al. 2024).

Dans ces annotations initiales nous disposons de 9 variables d'intérêt pour notre approche :

4 variables catégorisées

Part Of Speech (POS) qui se décline sur 16 catégories effectives (on exclut par exemple 'non catégorisable')

Type d'Unité Impliquée qui se décline sur 7 catégories

Module de traitement qui se décline sur 12 catégories

Opération Formelle qui se décline sur 6 catégories

5 variables textuelles

les énoncés annotés et leurs reformulations

les POS in extenso

des commentaires in extenso des annotateurs

exemple pour la catégorie « Syntaxe » de la variable 'Module de Traitement' pour le texte dans le genre descriptif du scripteur 3-B-3GB1-25/02/2020

« Emploi de la conjonction de coordination "et" au lieu de l'adverbe "puis" qui est plus approprié pour décrire la succession des actions.»

L'une de ces variables, labellisée 'Module de Traitement', comporte 12 catégories dont « Syntaxe ». On compte ainsi 3197 annotations épinglées « Syntaxe » soit 37,6% du total global des annotations d'écarts à la syntaxe attendue.

On obtient à la fin de l'étape d'annotations des textes un tableau global d'annotations d'écarts à la syntaxe attendue de 8499 lignes et l'on peut repérer par des tests statistiques ce qui est caractéristique des écarts à la « Syntaxe » par rapport aux autres 11 catégories de 'Module de Traitement'

- en terme des autres écarts à la syntaxe attendue : POS -Partie du Discours-, Opération Formelle, Type d'Unité,
- et en terme de métadonnées sur les scripteurs

2.2. L'annotation des écarts à la syntaxe attendue

Nous présentons dans le Tableau 1 le résultat des tests de liaisons statistiquement significatives pour la modalité 'syntaxe'

Voici en exemple la lecture de la première ligne de résultats:

- dans le 'Type d'Unité Impliquée', seul 'Syntagme' est significativement lié à « Syntaxe » en 'Module de Traitement'
- 'Syntagme' représente d'ailleurs 100% des caractérisations en 'Type d'Unité Impliquée' des écarts à la « Syntaxe », donc aucun annotateur n'a reporté d' autre 'Type d'Unité Impliquée' dans « Syntaxe » que 'Syntagme'.
- 'Syntagme' est cité dans 52% des annotations d'écarts à la syntaxe attendue au total et ne couvre que 23% des écarts à la « Syntaxe » en 'Module de Traitement', et c'est bien l'écart entre 100% et 23% qui contribue à la significativité de la relation entre 'Syntagme' et « Syntaxe »
- enfin les écarts à la syntaxe attendue accompagnent à hauteur de 72,4% les annotations 'Syntagme'
- On n'obtient donc de liens statistiquement significatifs aux écarts à la syntaxe attendue, plongée dans l'ensemble de toutes les autres annotations, que pour les éléments suivants :

'Syntagme', 'Conjonctions de Coordination', 'Déterminants', 'Préposition », 'Pronom', 'Construction Verbale', 'Adverbe', 'Insertion', 'Omission', 'Permutation', 'Parents, Grands Parents sourds', 'LSF découverte et pratiquée en langue native en contexte familiale', 'LSF pratiquée en famille, à l'école, avec les camarades', 'Choix de la LSF pour l'entretien', 'LSF privilégiée spontanément en 1ère langue à l'oral' et 'Autoévaluation en LSF noté en Très Bien'

Libellé de la variable	Modalité caractéristique	% de la modalité dans la classe	% de la modalité dans l'échantillon	% de la modalité dans le complément aire de la classe	% de la classe dans la modalité	Probabilité
TYPE D'UNITÉ IMPLIQUÉE	Syntaxme	100,0	52,0	23,0	72,4	0,000
PARTIE DU DISCOURS (POS)	CCONJ	5,2	2,1	0,2	93,8	0,000
PARTIE DU DISCOURS (POS)	DET	21,8	10,1	2,9	81,7	0,000
PARTIE DU DISCOURS (POS)	PREP	19,0	9,1	3,2	78,3	0,000
PARTIE DU DISCOURS (POS)	PRON	12,5	6,2	2,4	76,2	0,000
OPERATION FORMELLE	Insertion	12,5	7,8	5,0	60,2	0,000
OPERATION FORMELLE	Omission	46,9	38,6	33,6	45,7	0,000
OPERATION FORMELLE	Permutation	5,0	2,4	0,8	78,9	0,000
PARTIE DU DISCOURS (POS)	CONST VERB	5,4	2,8	1,2	73,1	0,000
PARTIE DU DISCOURS (POS)	ADV	7,5	5,0	3,5	56,2	0,000
Pratique de la LSF	PratLSFnat	43,2	38,3	35,4	42,4	0,000
Age de découverte de la LSF	ÂdecLSF0a	42,0	37,4	34,6	42,3	0,000
Pratique la LSF	NIVfamLSFnatif	42,0	37,4	34,6	42,3	0,000
Contexte de pratique de la LSF	CTXsco&cam&famAUTRIs	12,3	9,6	7,9	48,4	0,000
Père	PèS	38,8	34,6	32,0	42,3	0,000
Modalité choisie pour l'entretien	CHXoLSF	54,2	49,8	47,1	41,0	0,000
Autoévaluation du niv. en LSF	AutoNIVlsfTB	10,3	8,0	6,6	48,7	0,000
Langue 1 (privilegiée)	LV1oralSPONTIsf	41,7	37,7	35,3	41,6	0,000
Mère	MèS	39,6	35,7	33,3	41,8	0,000
Grand-père paternel	GpèPS	13,6	11,0	9,5	46,4	0,000
Grand-mère paternelle	GmèPS	13,6	11,0	9,5	46,4	0,000

Tableau 1 : Catégories caractéristiques de l'annotation « Syntaxe » dans les 8499 annotations

Nous filtrons par la catégorie « Syntaxe » les énoncés ou autres éléments d'organisation textuelle (listes etc.) annotés initialement en « Syntaxe » pour former de nouveaux tableaux visant à étudier les scripteurs au travers de leur seules annotations en « Syntaxe ». Le tableau filtré à partir du tableau initial ne comporte alors que 3473 lignes parmi les 8499 du départ, Nous y adjoignons également la variable textuelle 'Commentaire des annotateurs' qui fera l'objet d'un traitement spécifique. Nous avons donc en colonnes 10 variables, les 8 variables décrites au paragraphe 2.1 à l'exclusion de « Module de Traitement », l'identifiant du scripteur répété en ligne autant de fois qu'il a été annoté « Syntaxe ». et la variable textuelle.

Nous notons ce tableau T_1 , où 'I' est la première lettre de 'Initial', il y aura d'autres lettres pour indiquer les autres tableaux.

2.3. Variables dérivées des annotations expertes

A partir des éléments de T_1 annotés non conformes aux compétences attendues du Brevet, nous construisons des variables supplémentaires que nous a suggérées en partie la grille de Boivin et Pinsonneault (Boivin and Pinsonneault, 2020) et qui complètent l'examen des textes. Pour chaque scripteur:

nombre d'annotations au total ,

nombre de phrases

longueur Max des phrases sur l'ensemble de ses textes

3 variables : nombre de mots, nombre d'annotations « Syntaxe », NM100 -pourcentage sur les mots

16 variables : nombre d'annotations POS pour chaque catégorie de POS

1 variable : nombre d'annotations Module de Traitement cumul 'Syntaxe' ou 'Ponctuation'

5 variables : nombre d'annotations Opération Formelle pour chaque catégorie uniquement pour les annotations Syntaxe ou Ponctuation

6 variables : nombre d'annotations Type d'Unité Impliqué pour chaque catégorie uniquement pour les annotations Syntaxe ou Ponctuation

1 variable : calcul de la complexité syntaxique moyenne de ses textes

Nous formons ainsi un deuxième tableau qui comporte en lignes non pas les annotations mais les scripteurs repérés par leurs identifiants et en colonnes les variables dérivées des

Nous avons obtenu à la fin de ces diverses étapes de prétraitement deux types de tableaux de données.

Le premier est le tableau T_1 qui filtre sur les éléments initiaux les éléments fléchés « Syntaxe », de 3473 lignes et 10 colonnes.

Le deuxième est un tableau de 66 lignes que nous noterons T_J , où 'J' est l'initiale de 'Jointés' qui est la jointure sur les identifiants des scripteurs des trois tableaux T_D , T_N et T_M et qui comporte donc 66 lignes et 111 colonnes.

Ces deux tableaux sont résumés sur la figure 2

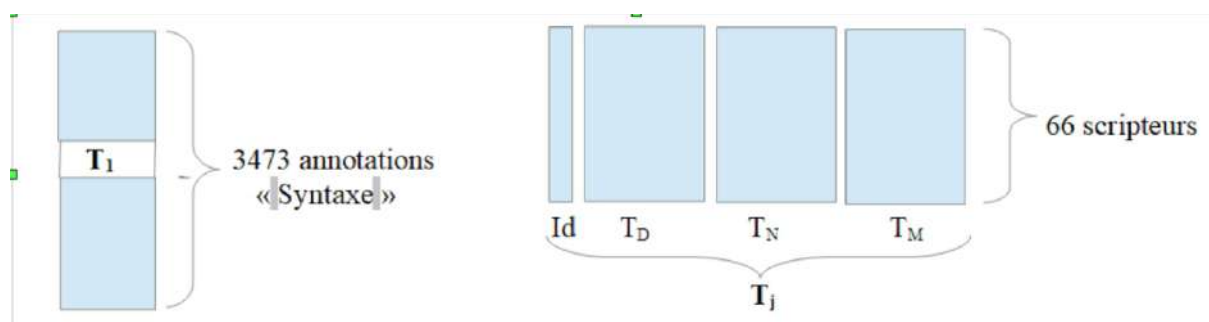


Figure 2: Les deux tableaux de l'analyse de données textuelles

3. Analyse des tableaux de données filtrés par la modalité « Syntaxe »

Notre démarche de recherche consiste dans son ensemble :

1. dans la détermination des paramètres les plus exhaustifs possibles pour une approche multidimensionnelle de notre hypothèse ; en output de l'étape l'ensemble des variables de l'étude, données, reconstruites ou ajoutées ex novo
2. en la construction d'une synthèse de l'information portée par les textes parsés et annotés en écarts à la syntaxe attendue au brevet des collèges; en output les caractéristiques en syntaxe de chaque scripteur, comme des moyennes, des syntagmes textuels spécifiques ou encore les modalités associées de variables catégorisées syntaxiques ciblées
3. en une étude des scripteurs sourds renseignés par tout résumé, continu, catégorisé ou textuel, issus de leurs textes pré-traités à l'étape 2 ou des leurs métadonnées disponibles ; en output les classes résultantes d'une classification automatique non supervisée des élèves sourds
4. en la caractérisation des classes issues de l'étape 3 par les variables déjà en jeu mais aussi par des indicateurs nouveaux propres au Traitement Automatique du Langage ; en output des descriptions explicites regroupant et discriminant les scripteurs sourds source de réflexion pour un éclairage purement linguistique

Les données que nous soumettons à l'étude sont après les étapes de prétraitement l'ensemble T des textes résultats de la numérisation des 226 pages manuscrites initiales , et les deux tableaux de données T_1 et T_J

3.1. Résultats de la classification automatique sur les scripteurs

3.1.1. Résultat du partitionnement

Nous cherchons parmi les 66 scripteurs sourds s'il existe des types aisément définissables et qui se ressemblent du point de vue de leurs syntaxe. La dissimilarité utilisée dans la classification mixte sera fondée sur les 109 descripteurs syntaxiques de T_J à l'exclusion de la variable textuelle.

Cette classification est l'enchaînement d'une réduction de dimension par analyse factorielle puis d'une classification mixte où après des étapes d'aggrégation autour de centres mobiles nombreux, de quantités variables avec les essais et avec plus de centres initiaux que le résultat attendu, on ne garde que les agrégations robustes c'est-à-dire les sous groupes stables au fil des déroulés. On hiérarchise enfin de façon ascendante les sous groupes retenus. Le choix de la coupure finalement retenue se fait suivant les gradients comparés des variances dans le suivi des agrégations, et suivant un point de vue expert fondé sur l'interprétabilité des classes au regard du problème que l'on se pose. Nous retenons la coupure en quatre classes.

3.1.2. Interprétation statistique des profils retenus par les variables actives de l'annotation

Nous avons en résultats pour chaque classe des tableaux semblable au Tableau 1. Il y en a donc quatre et les variables en jeu sont toutes les variables de (T_D U T_N) à l'exclusion de la variable textuelle.

Voici en exemple sur le tableau 2 les résultats pour la classe 1

Libellé de la variable	Modalité caractéristique	% de la modalité dans la classe	% de la modalité dans l'échantillon	% de la modalité dans le complémentaire de la classe	% de la classe dans la modalité	Probabilité
OPERATION FORMELLE	Omission	60,3	46,8	45,8	9,3	0,000
PARTIE DU DISCOURS (POS)	DET	30,6	21,8	21,1	10,2	0,000
PARTIE DU DISCOURS (POS)	VERB	9,1	5,1	4,8	12,8	0,000

Tableau 2 : Caractérisation du profil 1 par les variables actives d'annotation

Voici les principales informations qui émergent pour chacune des classes que nous nommerons classe 1, 2, 3 et 4

Dans la classe 1 les annotations de POS sur les verbes, les constructions verbales et les déterminants sont significativement plus présentes que chez les autres scripteurs, de même que l'annotation Omission et l'absence de subordonnées, le nombre moyen d'écarts à la norme en pourcentage rapporté au nombre de mots utilisés dans les textes de ces scripteurs ; la syntaxe est significativement remarquée comme dérivant de la syntaxe LSF et des connecteurs inusités sont employés de façon caractéristique avec une parataxe importante.

Dans la classe 2 on remarque un défaut statistiquement significatif d'annotations en POS Verbe ou d'annotation Omission, de même que d'annotations qualifiées de 'syntaxe de l'oral vers l'écrit', par contre une bonne linéarité, et une abondance significative de subordonnées.

Dans la classe 3 on remarque un défaut statistiquement significatif d'annotations de POS Déterminant ou d'annotation Préposition, de même que d'écarts qualifiées de 'syntaxe de l'oral vers l'écrit' et de nombre de subordonnées, par contre une abondance significative de parataxes.

Dans la classe 4, on remarque une présence significative d'annotations en POS sur les verbes et les pronoms et d'annotation 'permutations' ainsi que d'annotations qualifiées de 'syntaxe de l'oral vers l'écrit', un défaut statistiquement significatif d'annotations en POS Déterminant et Prépositions.

Nous nommerons dorénavant la classe 1 'Syntaxe non conforme aux compétences attendues du Brevet', la classe 2 'Syntaxe conforme aux compétences attendues du Brevet', la classe 3 'Syntaxe plutôt conforme aux compétences attendues du Brevet', la classe 4 'Syntaxe plutôt non conforme aux compétences attendues du Brevet'.

3.1.3. Interprétation statistique des profils retenus par les métadonnées

Chaque profil de scripteur produit un tableau comme le Tableau 1 mais qui met en jeu les métadonnées qui le caractérisent.

Voici en figure 4 un extrait en exemple pour le profil des scripteurs 'Syntaxe conforme aux compétences attendues du Brevet' les métadonnées surprésentées dans la classe:

Libellé de la variable	Modalité caractéristique	Probabilité
Utilisation de la LSF J éva	jÉVALutilLSFnon	0,000
Contexte de pratique de la LSF	CTXsco&cam	0,000
avec Mère	ENVmèLSFN	0,000
Type de collège	COLLincluIND	0,000
Type de lycée	LYCincluIND	0,000

Figure 4 : Caractérisation par les métadonnées du profil 2

Sur l'ensemble des résultats on peut retenir en tendance que :

1. le profil des scripteurs à 'Syntaxe conforme aux compétences attendues du Brevet' est une classe plutôt appareillée, peu immergée dans la LSF, auto évaluée mauvaise en LSF et bonne en français, échangeant spontanément en français, de famille élargie toute entendant avec un défaut d'échanges en LSF avec la mère, en inclusion scolairement.
2. le profil des scripteurs à 'Syntaxe non conforme aux compétences attendues du Brevet' est une classe de famille globalement sourde, privilégiant la LSF dans les échanges, qui est une langue première native, plutôt non appareillée au moment de l'entretien et avec un nombre de locuteurs échangeant en LSF avec la mère statistiquement significatif
3. le profil des scripteurs à 'Syntaxe plutôt conforme aux compétences attendues du Brevet' est une classe qui a découvert la LSF seulement à partir du primaire, appareillée, privilégiant globalement le français dans les échanges, de famille globalement entendant et parlant français ou une autre langue
4. le profil des scripteurs à 'Syntaxe plutôt non conforme aux compétences attendues du Brevet' est une classe plutôt de surdité congénitale, connue dans la première année de vie ; les parents sont tous deux sourds mais les grands parents entendants . L'immersion dans la LSF est dès la naissance mais les échanges se font aussi bien en LSF qu'en français qui est la langue choisie pour l'entretien. L'autoévaluation en LSF est très bonne, moyenne pour le français écrit.

3.2 Caractérisation des classes par leur complexité syntaxique

La notion de complexité syntaxique que nous envisageons ici est ancrée dans les Tree banks Universal Dependencies (Kahane et al., 2017). La complexité syntaxique est en relation avec l'ordre des mots ou encore des éléments psycholinguistiques comme la mémorisation linéaire de gauche à droite en français. Il existe plusieurs métriques de complexité syntaxique comme la longueur de dépendance qui est le nombre de mots entre le dépendant et son gouverneur dans un arbre de dépendance. Nous prenons ici l'option d'une mesure qui cerne l'auto-enchâssement dans les phrases, dans le cadre d'une syntaxe de dépendance (Yan, 2021).

Les calculs sont effectués à partir de l'ensemble T des textes numérisés mais 'corrigés' en les segmentant pour que le parser puisse travailler correctement : on introduit par exemple de la ponctuation là où elle est manquante. Un parser ne prend pas non plus en charge des parties de textes, qui comportent par exemple des graphèmes inconnus comme des dessins en substitution de mots. C'est ce qui se passe dans le texte de la figure 5 ci dessous du scripteur 15B 3GB2 où l'on voit une casserole dessinée à la place du mot

pour signer dans l'espace, de plus le temps de brefs déplacements des mains et enfin l'expression du visage qui s'ajoutent au contexte en font dans les brefs instants multidimensionnels des échanges successifs d'un discours une langue bien différente de la réalité cognitive du message 'français écrit' dans son déroulement en écriture ou lecture. Il serait donc intéressant d'élargir ces calculs et l'étude à d'autres corpus de scripteurs sourds.

4. Conclusion et perspectives

Une conclusion émergeant de cette étude est la non homogénéité des scripteurs sourds, pourtant unis par leur caractéristique auditive, quant à la syntaxe de leurs productions écrites dans le corpus ALPHA. On peut en effet mettre en évidence quatre types caractéristiques que l'on peut décrire de façon statistiquement significative par les variables de l'étude.

Cependant cette constatation demande à être consolidée dans d'autre corpus à constituer.

Enfin une analyse contrastive avec des textes d'entendants en cursus standard, bilingue ou FLE est en cours par des méthodes de classification supervisée. On a ainsi déjà montré que les annotations en syntaxe ne sont pas que dans les syntagmes chez les entendants. Ce serait donc une première règle pour reconnaître un texte d'entendant : « si annotation en syntaxe hors syntagme alors texte d'entendant ».

References

- Aleksandrova T., Aykurt-Buchwalter S., Bernardon de Oliveira K., Felce C., Gómez L. et Naser Eddine A. (2024). « Présentation. — L'écrit en langue seconde en contexte universitaire : l'apport des recherches basées sur les corpus d'apprenants », *Lidil [En ligne]*, 70 | 2024, mis en ligne le 31 octobre 2024, consulté le 31 mars 2026. URL : <http://journals.openedition.org/lidil/12952> ; DOI : <https://doi.org/10.4000/12lm7>
- Balbi S. (2010) Beyond the Curse of Multidimensionality: High Dimensional Clustering in Text Mining. *Statistica Applicata - Italian Journal of Applied Statistics* Vol. 22 (1)
- Boivin, MC. and Pinsonneault, R. (2020). La catégorisation des erreurs linguistiques : une grille de codage fondée sur la grammaire moderne. *Le français aujourd'hui*. N°209. 89. 10.3917/lfa.209.0089.
- Durand, G. (2003). Écrire en L2 : bilan et perspectives des recherches. *Arob@Se volume 1-2*, pp. 6-21, 2003
- Gerdes K., Kahane S. and Chen X. (2021). Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics* (2021-...), 6(1):17.
- Lacheret-Dujour A., Bogliotti C., Jimenez A., Prenovec R. (2024). Guide d'annotations collaboratif Corpus Littératie chez les Sourds dit "Corpus ALPHA", *hal-04817848*
- Périni M. (2013). Que peuvent nous apprendre les productions écrites des sourds ? Analyse de lectures écrites de personnes sourdes pour une contribution à la didactique du français écrit en formation d'adultes. *Thèse de doctorat en Linguistique. Université Paris 8, Laboratoire SFL*
- Prenovec R. (2025). La littératie chez les sourds : états des lieux sociolinguistique et linguistique, perspectives didactiques. *Thèse de doctorat en Linguistique. Université Paris Nanterre, Laboratoire Modyco*
- Yan, C. (2021). Complexité syntaxique et flux de dépendance : études quantitatives dans les treebanks universal dependencies *Thèse de doctorat en linguistique, Université Paris Nanterre, laboratoire Modyco*

S07

Specialized Session
Finance and Textual Analysis

10-lug
09:30 - 11:00
Aula 11

Web Automatic Investigator (WAI): an Agentic Framework for Regulatory Multimodal Analysis of Web Contents

Andrea Coletta¹, Miriam Corrado², Gabriele De Capoa², Oliver Giudice¹,
Alessandro Maggi¹, Marco Mori¹, Anna Russo Russo¹, Valerio Signore²

¹Directorate General for Information Technology

²Directorate General for Consumer Protection and Financial Education
Banca d'Italia – [firstname].[lastname]@bancaditalia.it*

Abstract

The increasing adoption of Artificial Intelligence (AI) into Supervisory Technology (SupTech) has significantly transformed and accelerated many supervisory activities; yet, several tasks still require supervisors to manually inspect financial products and directly interact with financial institutions. Among these tasks, the assessment of institutional websites for regulatory compliance and usability remains largely manual, with supervisors navigating hundreds of websites using predefined regulatory checklists. Although data-driven solutions have been proposed for related SupTech applications, their applicability is limited by the dynamic and interactive nature of this inspection process, and by the heterogeneity of institutional websites. To address this gap, we introduce the Web Automatic Investigator (WAI), an agentic AI framework developed at the Central Bank of Italy that autonomously navigates, extracts information, and analyzes institutional websites in accordance with a supervisory checklist. Our approach consists of two main stages: first, the system navigates the target website to extract all available information and build a global representation; then, a set of agentic pipelines analyzes the extracted data and dynamically interacts with the website to address all items of the checklist. Empirical results, validated through end-user testing, show that our framework effectively supports supervisors by providing accurate quantitative compliance indicators and qualitative evidence (e.g., screenshots), thereby significantly speeding-up the compliance assessment of institutional websites.

Keywords: Information Retrieval, WebGraph, Website analysis, LLM, Multimodal LLM, Agentic AI, SupTech.

1. Introduction

During the past decade, the financial industry has undergone a major digital transformation, with banks and financial institutions now conducting most of their operations digitally (Tanda and Schena, 2019). In this context, the websites of financial entities have become increasingly central, not only for the daily operations of customers but also as repositories of product information, legal disclosures, transparency documentation, and complaint procedures.

As a consequence, institutional websites have become an increasingly important priority for banking supervision. In particular, Central Banks have focused on the assessment of institutional websites of financial entities to check for regulatory compliance and usability. The assessment of a website is typically carried out manually, using a structured checklist that includes all relevant requirements, covering both compliance and usability dimensions. For instance, usability criteria may include verifying whether contact information is provided, while

*Views and opinions are those of the authors and do not necessarily reflect the official policy or position of Banca d'Italia. Authors listed in alphabetical order.

compliance criteria may require checking adherence to specific European regulations governing disclosure and transparency obligations.

Although the increasing adoption of Artificial Intelligence (AI) in Supervisory Technology (SupTech) has significantly transformed and accelerated many supervisory functions (Barasa et al., 2024), the website inspection process, among several other tasks, still relies almost entirely on manual work by supervision experts and direct interaction with financial entities. In particular, website assessment requires supervisors to periodically analyze a large number of websites, ranging from a few dozens to several hundreds. Indeed, while data-driven SupTech solutions exist, the heterogeneity of institutional websites, together with the interactive and multimodal nature of the inspection process (e.g., assessing the usefulness of the website search engine), limits the applicability of existing approaches.

In this paper, we address the problem of automating the assessment of institutional websites by introducing the Web Automatic Investigator (WAI), an agentic framework developed at the Central Bank of Italy for the automated regulatory analysis of institutional websites. Our framework analyzes a given website, through the lens of the provided regulatory checklist, using two main core components: (i) an intelligent scraper that builds a unified representation of each investigated website; and (ii) a set of agentic pipelines designed to answer the compliance checklist while handling the heterogeneous nature of websites through interactive navigation and reasoning over the website content.

More specifically, the first component performs a browser-based crawling of the target domain, capturing all downloadable and dynamic content, which are then organized into a unified multimodal representation, along with a graph-based model of the website pages.

The agentic pipelines operate on top of this shared representation, decomposing each regulatory checklist item into a sequence of coordinated subtasks handled by specialized AI agents. For each checklist item, the pipeline produces a quantitative compliance score that reflects the degree to which the website satisfies the corresponding regulatory requirement. The score is accompanied by qualitative supporting evidence that justifies the assessment. For example, when evaluating transparency obligations for a specific product, the evidence may include a screenshot of the relevant transparency section along with its URL. Figure 1 provides an overview of the WAI architecture.

We empirically evaluate the WAI framework on a sample of real institutional websites, including 14 supervised entities (i.e., 7 banks, 2 payment institutions (PIs), 1 electronic money institution (EMI), and 4 non-bank credit institutions). We evaluate the framework by comparing the generated checklist answers against ground-truth annotations produced by domain experts and complement this evaluation with an end-user usability study. Our results confirm that WAI can effectively support the supervision of institutional websites, substantially reducing the time required for the assessment.

The main contributions of this work can be summarized as follows:

- we **propose WAI**, a novel agentic framework designed to automate the assessment of institutional websites against a supervisory checklist, producing both quantitative compliance scores and qualitative supporting evidence to ensure transparency and explainability of the resulting assessments;
- we **design an intelligent scraper** that constructs a unified multimodal representation of institutional websites along with a graph-based model of the website structure;

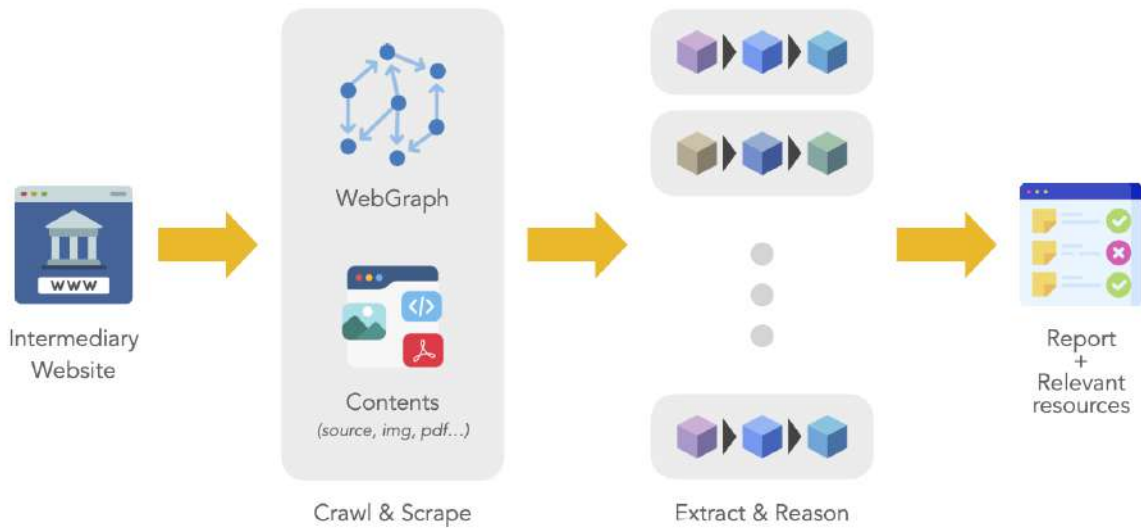


Figure 1: Sketch of the WAI framework architecture.

- we **develop a set of agentic pipelines** that operate over the multimodal representation and dynamically interact online with the “live” website to address each item in the supervisory checklist;
- we **validate WAI** empirically on a heterogeneous set of real institutional websites, complemented by an end-user study with domain experts, demonstrating the applicability and effectiveness of the framework in a real SupTech setting.

The remainder of the paper is organized into four sections. Section 2. introduces the WAI framework, with its two main components: the crawling module discussed in Section 2.1. and the agentic pipelines introduced in Section 2.2.. Section 3. reports the experimental evaluation, while Section 4. discusses the most relevant related work. Section 5. concludes the paper.

2. WAI Framework

In this section, we introduce the WAI framework — shown in Figure 1 — along with its two main building blocks. The first component, namely the **Crawl & Scrape** module, performs a browser-based crawling of the entire target website capturing all downloadable content, including HTML sources, full-page screenshots, and PDF documents. The extracted data is then organized into a unified multimodal representation, which comprises textual, visual, and meta-data layers, along with a graph-based model of the website. The second component, the **Extract & Reason** module, consists of a set of agentic pipelines designed to automatically compile a supervisory regulatory checklist. Each pipeline processes and reasons over the website representation produced by the first component, solving the checklist tasks through multiple heterogeneous agents. Agents are either *interactive agents*, *pattern-matching agents*, or *LLM-based agents*. Finally, for each checklist item, the pipelines produce a quantitative compliance score reflecting the degree to which the website satisfies the corresponding regulatory requirement, along with qualitative supporting evidence.

Finally, a web-based dashboard reports the compliance scores and their associated evidence (see Figure 3), enabling supervisors to efficiently inspect and audit the results.

2.1. Crawl & Scrape: Multimodal Representation of Institutional Websites

WAI represents each institutional website as a structured graph of webpages. Each webpage is represented as a multimodal entity that combines textual content, visual information, and meta-data. The website representation is constructed through an extensive browser-based crawling phase and enriched by processing static resources collected along the crawl. This results in a unified and structured representation that serves as input to the agentic pipelines.

Browser-Based Crawling and Content Acquisition. The construction of the website representation is initiated through an automated exploration of the target domain performed by a headless browser, orchestrated via Playwright (Microsoft, 2026). Unlike simple HTTP crawlers, this approach enables WAI to correctly handle dynamically generated content, client-side rendering, pop-ups, and interactive elements, thus supporting a more complete and faithful coverage of modern websites.

The crawling process follows a breadth-first traversal strategy over the target domain, systematically visiting all reachable webpages. On each webpage, WAI executes a standardized acquisition procedure. First, obstructive elements that prevent interaction with the underlying content, such as pop-ups, are automatically dismissed. Then, the system retrieves the HTML source code, captures a full-page screenshot and collects all outgoing hyperlinks, distinguishing between internal pages, external domains and static resources. Selected static resources are also acquired during the crawling process: in the current implementation, WAI focuses on PDF documents, as they often contain transparency documentation and contractual terms. PDF documents are downloaded either through the browser or via direct HTTP requests, depending on how the resources are exposed by the website. Out-of-domain pages are also visited with a similar acquisition procedure, but with the sole purpose of extracting their main textual and visual information, whereas hyperlinks and static resources are not collected further. To ensure termination and prevent infinite crawling caused by dynamically generated URLs and parameter variations, WAI adopts a deduplication mechanism based on content-level hashing.

The acquired HTML content is subsequently processed to separate the main body from navigational elements, such as the header and footer sections. The textual content is then extracted and converted into a Markdown format using *html2text* (Swartz, 2026). Textual content is extracted also from PDF documents, using *PyMuPDF* (Artifex Software, Inc., 2026). When direct text extraction fails, e.g. in image-based or scanned documents, or in files with non-standard encodings, WAI falls back on OCR using *pytesseract* (Smith, 2007). Simple heuristics on font size and position are applied to identify the document title, which helps in subsequent classification and retrieval steps.

Page-level representation. The multimodal representation of a webpage is organized into three complementary layers, each capturing a distinct aspect of the page content and context:

- the *textual layer* aggregates textual content extracted from HTML pages and linked PDF documents;
- the *visual layer* consists of full-page screenshots and selected visual crops, captured either after page loading or following interactions with interface elements;
- the *metadata layer* includes several attributes collected during the crawling process, such as the URL and its aliases, the distance from the homepage, indicators of the presence

of obstructive elements, the anchor text of incoming hyperlinks, and other contextual information.

Each layer is used across one or more agentic pipelines. The visual layer is inspected by visual agents to assess the layout properties and the presence or prominence of specific interface elements. The textual and metadata layers support a wide range of tasks, including page filtering and classification, information extraction, and multiple forms of content-based compliance assessment.

WebGraph and Site-Level Embedding. Beyond page-level content, WAI maintains an explicit structural representation of each website in the form of a directed graph, referred to as a *WebGraph*. In this representation, nodes correspond to webpages discovered during the crawling phase, while directed edges encode hyperlink relationships between pages. The WebGraph is constructed and managed using *NetworkX* (Hagberg et al., 2008).

A set of graph-level statistics is computed to obtain a compact characterization of the website topology; this set includes the number of nodes and edges, the maximum depth from the homepage, the average out-degree, and graph density. These descriptors are combined into a fixed-length vector that serves as a lightweight site-level embedding. Although intentionally simple, this embedding enables basic exploratory analyses such as projecting websites into low-dimensional space, identifying structural outliers, e.g., exceptionally shallow or overly complex websites, and grouping websites with similar navigational structures. In the current implementation, these features are primarily used for offline exploratory analysis and to monitor the effectiveness and coverage of the crawling process. However, the same representation could also be leveraged to adapt agents behavior, e.g., to tune navigation strategies based on identified website profiles.

2.2. Extract & Reason: Pipeline-Based Agentic Architecture

The **Extract & Reason** module implements an agentic architecture in which each item of the supervisory checklist is mapped to one pipeline. Each pipeline decomposes the corresponding supervisory question into a sequence of coordinated sub-tasks, each handled by a specialized agent. This design improves transparency in the assessment process, making explicit what evidence is collected and used, how each agent derives partial judgments, and how these are ultimately aggregated into a final outcome.

Agent Types. WAI employs more than thirty specialized agents, grouped into three main classes that reflect complementary capabilities and different levels of abstraction:

- *Interactive* agents operate on the live website through the browser interface, simulating typical user behavior across different tasks such as opening menus, following links, using the internal search engine, and traversing theme-specific sections. Their execution is driven by natural-language goals (e.g., “find the complaints section”), and their output is either a structured answer or plain text, along with a visual trace of the navigation session. These agents are implemented using *Browser Use* (Müller and Žunič, 2024).
- *Pattern-matching* agents carry out pattern-matching on textual artifacts extracted from HTML pages and PDF documents. They implement deterministic checks based on either regular expressions or simple structural heuristics. Among other objectives, they detect

contact information, mandatory phrases, expected section titles, and the presence of specific labels in transparency documents. These agents return as a result either a positive or a negative outcome along with text excerpts and a lists of URLs where the patterns were found.

- *LLM-based* agents process complex content, such as transparency documentation or full-page screenshots, to answer questions that require visual or semantic interpretation rather than simple pattern detection. Among others, typical tasks include classifying documents, assessing information disclosure with respect to regulatory requirements, and judging the clarity of rendered page layouts. These agents typically return a yes/no answer with structured judgments that include rationales and references to documents and webpages.

The agents' behavior is made tunable via multiple input parameters, in order to promote reuse across websites and pipelines.

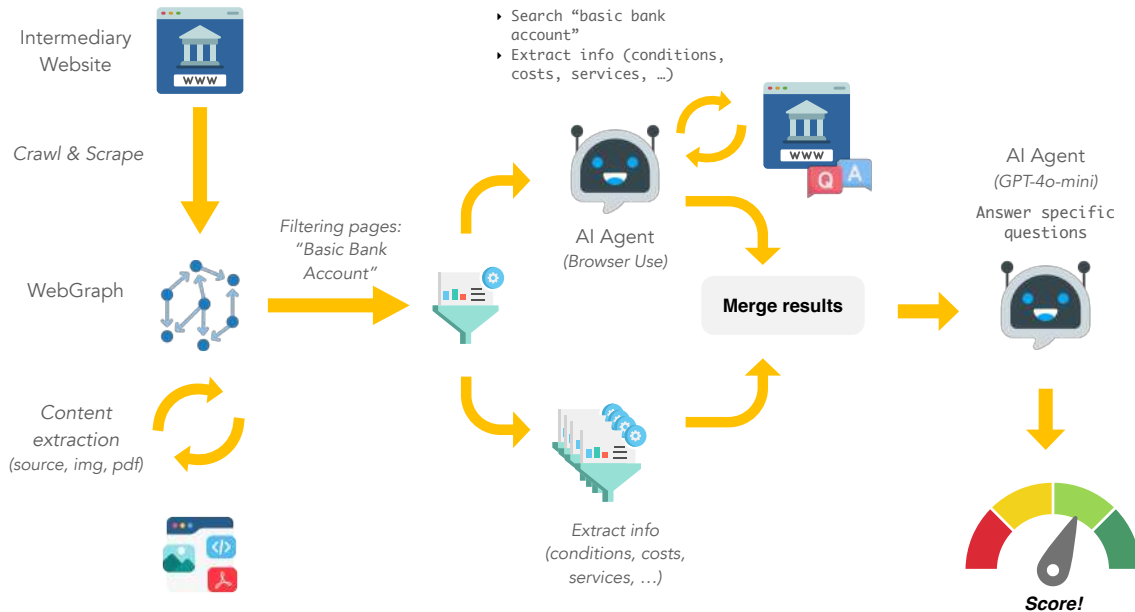


Figure 2: Workflow for the assessment of exposed information on the basic bank account. This task exemplifies how an agentic pipeline, operating on top of the multimodal representation produced by the Crawl & Scrape module, orchestrates multiple agents to produce a final compliance score for a checklist item.

Checklist-Guided Pipelines. Pipelines are structured workflows that orchestrate multiple agents to answer a supervisory question. They are implemented based on a common design pattern to support maintenance and extension: a *retrieval* phase exploits interactive and pattern-matching agents to identify webpages, PDF documents and text excerpts that are potentially relevant for the question at hand; a *reasoning* phase uses LLM-based agents to check the compliance of the retrieved content with the input requirements; then, a *scoring* phase consolidates intermediate results into a final judgment and stores supporting evidence.

Both the retrieval and reasoning phases of a pipeline may follow an ensemble-inspired strategy, employing different types of agents to process the same input or sub-task and jointly contribute

to the final outcome. This redundancy improves robustness and coverage by leveraging the complementary capabilities of the agents. AI agents, while flexible, are inherently non-deterministic and may occasionally produce unreliable outputs. Pattern-matching agents, by contrast, offer more stable and predictable behavior, but tend to be rigid when faced with diverse and complex inputs.

As an example, Figure 2 illustrates a pipeline designed to assess the compliance of the information provided for the offered “*basic*” bank account. During the retrieval phase, starting from the WebGraph representation generated by the Crawl & Scrape component, a pattern-matching agent identifies pages and documents mentioning the basic bank account. The process then follows two parallel paths: one uses an additional pattern-matching agent on the retrieved pages and documents, while the other relies on an interactive agent to collect information by interacting directly with the website. The output of this phase is a structured collection of heterogeneous data, such as contract terms and included services, which is then provided during the reasoning phase to an LLM-based agent to assess compliance with regulatory requirements.

Other pipeline examples might include a pipeline for *retrieving contact information*, using pattern-matching agents that search for relevant keywords, or a pipeline for *evaluating the possibility of subscribing on-line products/services*, using interactive LLM-based agents that actively attempt to subscribe for such products/services through the analyzed website.

Aggregation, Scoring, and Intermediate Evidence. Complex supervisory tasks often require verifying multiple distinct conditions, possibly across several objects, e.g., pages or documents; in addition, pipelines may include parallel paths to improve robustness. Consequently, at various stages of a single pipeline, multiple intermediate outputs may need to be aggregated into a single synthesized result. Intermediate results are combined using simple aggregation rules, depending on the supervisory question, such as the *any-match*, where the condition is considered satisfied if at least one result is positive, and *all-match*, where all results must satisfy the requirement.

The final output of a pipeline typically consists of: (i) a compliance score based on a codified categorical measuring system; (ii) a short natural-language justification; (iii) a curated set of evidence, including URLs, text snippets, screenshots, and navigation traces. This design choice is crucial to enable traceability of the performed evaluation as it allows supervisors to inspect how each assessment has been produced.

3. Experimental Evaluation

The WAI framework has been implemented as a Python prototype application. To simplify the evaluation process of the experimental results by domain experts, the prototype is equipped with a web-based interface that enables on-demand analysis of websites and inspection of results (see Figure 3).

Validation. The evaluation considered a heterogeneous sample of 14 institutional websites, including 7 banks, 2 payment institutions (PIs), 1 electronic money institution (EMI) and 4 non-bank credit institutions. This sample reflects a wide range of business models and website structures, representative of the target supervision task.

For each website, we considered two different checklists: (i) one for coherence and usability and (ii) one for regulatory compliance with transparency-related requirements. Each checklist consists of a set of questions to answer. An expert team manually answered the two checklists

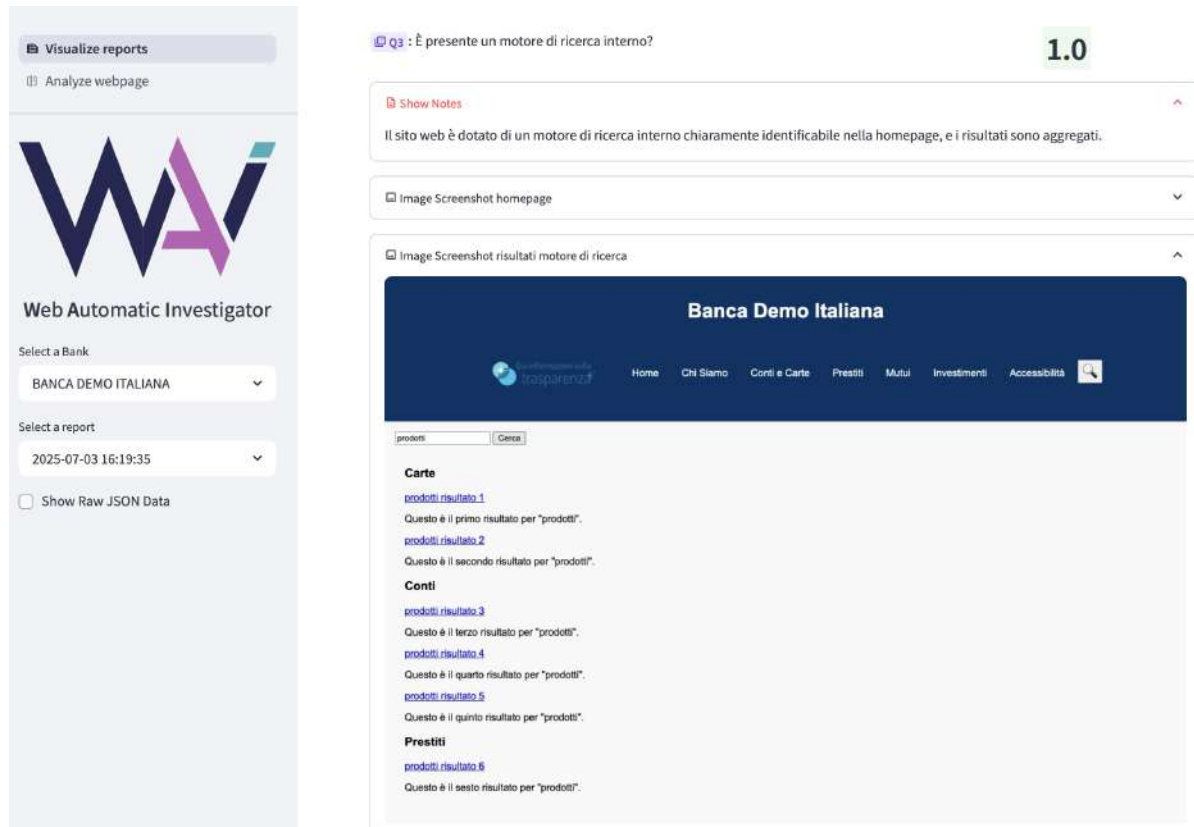


Figure 3: The WAI dashboard, showing the checklist-based evaluation of a fictitious bank’s website. The snapshot illustrates the assigned evaluation score together with its natural-language justification and one associated visual evidence.

for all the websites in the sample, thus building the *ground truth* reference against which WAI’s outputs were qualitatively compared.

Given the limited sample size and the exploratory nature of the prototype, the evaluation focuses on qualitative agreement against the *ground truth* and analysis of errors rather than on a single aggregate numerical metric. For each checklist item, the expert team examined: (i) whether the WAI score matched their own binary or ordinal judgment; (ii) whether the attached evidence (URLs, excerpts, screenshots, navigation traces) was sufficient to understand and correct the automatic decision, if needed.

The results indicate that WAI can approximate expert checklists with a generally good degree of precision, particularly on websites whose structure is relatively homogeneous and aligned with typical banking layouts. For such sites, experts reported that automatic answers for coherence/usability and compliance questions were often consistent with their own judgments and that the associated evidence was adequate to support review and possible adjustments.

On more heterogeneous or atypical websites — especially in the case of some payment institutions and the EMI — the behavior of the pipelines was more inconsistent. In these cases, specific structural features (e.g., unconventional navigation schemes, deep nesting of relevant sections and atypical document organization) occasionally led to incomplete coverage or to misinterpretation by agents, thus resulting in missing or unsatisfactory answers for certain checklist items. These failure cases were documented as candidates for future refinement of crawling strategies, along with content selection heuristics and prompt design.

Cost. In addition to accuracy-related observations, the evaluation also tracked execution time and approximate LLM costs for a subset of runs (OpenAI’s gpt-4o and gpt-4o-mini models were adopted as the main reasoning backends). This information was used to contextualize potential efficiency gains relative to fully manual inspections.

Our experiments show that, for the WAI prototype, the processing time of a single website ranges from fractions of an hour to several hours, depending on the complexity of the website structure and the content volume. Execution times are generally related to the variety and types of services offered, resulting in shorter execution times for small banks and specialized financial entities, while requiring more time and computation to complete the analysis of larger banks with richer offerings. The average execution time and LLM cost, for different types of supervised entities in the validation sample, are reported in Table 1. Overall, the results suggest that WAI can potentially speed up compliance assessment for a wide range of institutional websites at a reasonable cost.

Supervised Entity	Average Execution Time (Hours)	Average LLM Cost (\$)
Payment Service Provider	0.26	4.56
Small-Medium Bank	3.29	28.03
Large Bank	16.82	193.21
Financial Company	1.03	9.78

Table 1: Execution time and average LLM cost by type of supervised financial entity

4. Related Work

Early work on website evaluation has mainly addressed general quality dimensions such as usability, content, appearance, functionality and interactivity, typically in non-regulated domains. Allison et al. (2019) provide a comprehensive survey and the GoodWeb framework, showing that evaluation criteria and methodologies must be tailored to the specific purpose and context of the use of a website. However, these approaches do not directly address sector-specific regulatory requirements or the need to generate auditable evidence for supervisory decisions.

Supervisory authorities increasingly rely on technological tools to enhance data collection, monitoring, and analytical capabilities, recognizing the potential of automated analysis to increase the coverage and frequency of inspections (Financial Stability Board, 2020; World Bank Group, 2018). More recently, studies by the BIS (Bank for International Settlements) (Gambacorta et al., 2025; Prenio, 2024) have stressed that successful SupTech initiatives must move beyond experimentation and integrate into supervisory processes, with attention to explainability, robustness, and institutional constraints. Our work aligns with this view by targeting a concrete use case — website analysis of supervised financial entities for the Central Bank of Italy — and by designing outputs that can be reviewed and corrected by human experts.

Recent advances in multimodal foundation models and web agents are also relevant to our setting. Research on multimodal web navigation shows that realistic interaction with websites benefits from combining textual representations with visual signals, since many interface elements and layout cues are only reliably accessible in rendered form. Current research in the field of web accessibility has focused on the development of scalable audit frameworks that integrate automated checks, multimodal analysis, and human feedback. These contributions indicate that complex website assessments are better approached through structured workflows and tool orchestration than through single-step prompting (Gu et al., 2026; Wang et al., 2026).

5. Conclusion

This paper introduces the Web Automatic Investigator (WAI), an agentic framework developed at the Central Bank of Italy to support supervisors in analyzing institutional websites through the lens of a regulatory checklist, which covers both compliance and usability requirements. The framework combines a crawling module, which builds a multimodal representation of the entire domain, with agentic pipelines, which tackle the regulatory checklist while handling the dynamic and heterogeneous nature of institutional websites. We validated WAI on a heterogeneous sample of 14 institutional websites involving domain experts, who confirmed that WAI effectively supports supervisory activities, substantially reducing the time required for manual inspection. The experiments also highlight current limitations of the framework when handling atypical website structures and underscore the importance of keeping human experts in the loop to validate results and guide the refinement of the pipelines. Future work will focus on extending the robustness of the framework to a wider range of website designs, and on exploring the use of site-level structural embeddings, derived from the website graph, for dynamic tuning of pipelines behavior and proactive supervision alerts.

References

- Allison R., Hayes C., McNulty C. A., and Young V. (2019). A comprehensive framework to evaluate websites: literature review and development of goodweb. *JMIR formative research*, 3(4):e14372.
- Artifex Software, Inc. (2026). Pymupdf. <https://pymupdf.readthedocs.io/>. Version 1.25.5.
- Barasa M., di Castri S., Grasser M., Kiuhan S., Letsiou K., and Sousa Faria L. (2024). State of suptech report 2024. Available at SSRN 5518142.
- Financial Stability Board (2020). The use of supervisory and regulatory technology by authorities. *FSB Report*, Nov.
- Gambacorta L., Lauridsen N., Prenio J., and Kiuhan S. (2025). Making suptech work: evidence on the key drivers of adoption. Available at SSRN 6098842.
- Gu M., Wang Z., Lai S., Gao Z., Zhou S., and Bu J. (2026). Towards scalable web accessibility audit with MLLMs as copilots. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38515–38523.
- Hagberg A. A., Schult D. A., and Swart P. J. (2008). Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pp. 11–15.
- Microsoft (2026). Playwright. <https://playwright.dev>. Version 1.49.1.
- Müller M. and Žunič G. (2024). Browser use: Enable ai to control your browser. *browser-use.com*.
- Prenio J. (2024). *Peering Through the Hype-Assessing Suptech Tools’ Transition from Experimentation to Supervision*. Bank for International Settlements, Financial Stability Institute.
- Smith R. (2007). An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pp. 629–633. IEEE.
- Swartz A. (2026). html2text. <https://github.com/aaronsw/html2text>. Version 2024.2.26.
- Tanda A. and Schena C.-M. (2019). *FinTech, BigTech and banks: Digitalisation and its impact on banking business models*. Springer.
- Wang X., Wang S., Zhou A., and Hao H. (2026). See and remember: A multimodal agent for web traversal. *arXiv preprint arXiv:2603.02626*.
- World Bank Group (2018). *From spreadsheets to suptech: Technology solutions for market conduct supervision*. World Bank.

C06

Digital tools, annotation and corpus infrastructures

10-lug
09:30 - 11:00
Aula 12

A Text analysis application: a library course in Introduction to Text Analysis

Deborah Grbac¹, Valentina Schiariti²

¹Library, Università Cattolica del Sacro Cuore di Milano – deborah.grbac@unicatt.it - ORCID number: 0009-0005-4410-5503

²Library collaborator, Università Cattolica del Sacro Cuore di Milano – valentina.schiariti-collaboratore@unicatt.it - ORCID number: 0009-0004-1098-8769

Abstract

Following the conclusion of a partnership with JSTOR Labs (ITHAKA), which lasted from April 2020 to the 1st July 2025, that supported the implementation of the Constallete project, a service for teaching and learning how to perform text analysis, the Library at the Università Cattolica del Sacro Cuore di Milano launched, in November 2025, a course titled “*Introduction to Text Analysis*”, addressed to students, researchers, and professors who could be interested in learning about computational text analysis, and how to apply it to their research interests. The course follows to a data-driven approach to research and aims to answer to a research question both quantitatively and qualitatively. The course proposes to use textual data and text analysis to perform information retrieval on sources such as specific literature available through Library’s subscribed databases, and information extraction on retrieved or custom-collected *corpora* (pdf documents, digitized documents that are machine readable by OCR-Optical Character Recognition). The course is ongoing but it is already open to users. After the completion of the first module, two more modules are foreseen.

Keywords: text analysis pedagogy, information retrieval, information extraction, TDM literacy

1. Introduction

From 2020 to 2025, participating in the Constellate project offered librarians, faculty, and students a unique opportunity to learn, share, and apply Natural Language Processing (NLP) techniques to support research projects and enhance library services. This experience transformed how the library at Università Cattolica del Sacro Cuore di Milano approached information services - moving beyond traditional information retrieval (mainly done on metadata and on OPAC - Online Public Access Catalogue), to more advanced information extraction from textual data (extraction of textual statistics and lexicometry measures from collections of text).

Many are the applications of NLP tasks to the librarian work, such as for instance: helping students and researchers to find out suitable sources (querying databases, or surfing the web for web scraping), offering hints for research project planning (brainstorming, literature reviews), looking for information by using meta-data and textual data (information retrieval and information extraction), trying to assess research hypotheses by means of textual data as evidence (Data science, or a data-driven approach to research).

From a librarian perspective, this means working not only with texts and databases containing texts, but with textual data. The shift is from a classical library task, information retrieval — finding relevant literature and searching catalogues such as the OPAC — toward information extraction, namely the analysis of textual data in order to identify contents, meanings, lexical patterns and contextual relations within collections of documents. These collections may include born-digital texts, digitized documents made machine-readable through OCR, or larger database collections. The aim is also to develop a more autonomous approach to text analysis, one that is not entirely dependent on the search engines, AI tools or pre-packaged applications provided by commercial databases.

Copyright issues must always be kept in mind when dealing with scholarly texts written by authors. Text analysis should be carried out only when legal access to texts is granted, for example through a library subscription or because the texts are available in open access. Where applicable, research institutions such as universities may also rely on the Text and Data Mining (TDM) exception to copyright law for research purposes, with no commercial intent, and provided that no opt-out has been declared by the rights holders.

2. The Constellate project

The Università Cattolica del Sacro Cuore di Milano Library experience started in 2020 with a first collaboration as a beta tester for a new service, called TDM pilot, at the time under construction by JSTOR Labs,¹ part of the ITHAKA, a non-profit organization,² aimed at offering support for learning, teaching and spreading text analysis techniques. The first activity consisted in participating as an attendee to some experimental courses in Python. These classes were progressively integrated with practical applications and lessons on different text analysis tasks, such as topic modelling and regular expressions. The courses were offered in a learning environment organized with Jupyter notebooks. The partnership involved multiple activities: trialing the prototype, providing feedback on new features during regular check-ins, attending classes, promoting the tool on campus, and integrating it into institutional teaching through lessons, workshops, and educational materials.

At the end of 2021, the service was launched and renamed Constellate. The pedagogy package encompassed two semesters' worth of courses (fall and spring) and a summer school called TAPI (Text Analysis Pedagogy Institute) first, then TAP Institute. The TAP Institute regularly occurred each summer from 2021 to 2024. Classes during the year were dedicated to the subscribing institutions' users, but from 2022 on the TAP institute was open to all interested in, subscribers' users or not, thanks to public financings.

In 2022, Constellate made a cloud-based Jupyter environment available to users - called the Constellate Lab - providing 8 gigabytes of memory, 1 gigabyte for sharing, and approximately 30 gigabytes of disk space. The Constellate pedagogy package began to introduce new topics related to NLP tasks, like Named Entity Recognition extracted through Python's library and modules such as SpaCy (a Python library for NLP processing),³ lessons explaining how generative pretrained transformers and neural networks work, where to find and how to use free

¹ JSTOR Labs' website: <https://labs.jstor.org/>, retrieved on 23rd January 2026.

² ITHAKA is a learning organization whose aim is to make access to knowledge more affordable and spread worldwide, website: <https://www.ithaka.org/>, retrieved on 23rd January 2026.

³ SpaCy website: <https://spacy.io/>, retrieved on 23rd January 2026.

Large Language Models (in Hugging Face⁴), and how to do automatic text classification using Chat-GPT (Generative Pre-trained Transformer) via API (Application Programming Interface), how to build vector databases with embeddings and organize RAG (Retrieval Augmented) pipelines to work on customized contents, leveraging LLMs (Large Language Models) and limiting hallucinations.

In 2023 the Library at Università Cattolica del Sacro Cuore di Milano, after some trialing of the partnership and internal discussion, subscribed to the service. The subscription allowed all university users access to the Constellate Lab and to the pedagogy package.

Participating in the partnership became increasingly intriguing and demanding. This growing engagement was an occasion to gradually add building blocks to previous knowledge, through learning and sharing of best practices within the program's community. This brought about a shift in the education offered by the library. In 2023, after a full subscription to the service, part of the classes taught at the library focused on information literacy were switched into text analysis. Finally, at the end of 2024, the Constellate "Skill Build Program" was an opportunity to verify what was learned by attending classes and performing a practical research pilot project aiming at applying text analysis to a real project (see paragraph 5).

3. An empowering and unique experience

This experience was unique in the way it empowered librarians in multiple dimensions: becoming autonomous through learning, applying new skills in practical contexts, staying updated with technological developments, developing a critical perspective on commercial tools, and ultimately becoming more effective in supporting research. The Skill Build Program was an effective exercise in research support, producing some useful insights from JSTOR Archive contents on selected authors' publications, such as, for example, the pilot project on David Abosch.

The Constellate Lab was a practical space in which users could do analysis, build datasets (the Dataset Builder function), share code (the Snapshot function) and store and retrieve files and data (the storage space). Moreover, because it worked inside a Jupyter environment, it also provided users with the opportunity to gain experience using a cloud-based file system. This allowed saving memory in local machines in terms of CPU (Computer Processing Unit) and memory disk space.

Throughout this process, the library paid careful attention to copyright and intellectual property right in text analysis works. The Dataset Builder provided 'non-consumptive' datasets that allowed users to analyze texts while respecting authors' rights, operating within the legal framework of the Text and Data Mining (TDM) exception for research at universities and research centers. This approach not only ensured legal compliance but also taught users how to conduct text analysis while respecting important concepts like legal and authorized access and copyright protection, considerations that are extremely valuable when working on customized collections of texts, both locally or in a cloud-based system). Moreover, the Constellate platform provided the opportunity to have dedicated educators and a community within practice and

⁴ Hugging Face website: <https://huggingface.co/>, retrieved on 23rd January 2026.

sharing knowledge and methodology. This made the educational experience more complete and more effective, and it eventually led to user independence when performing text analysis tasks.

Constellate provided the opportunity to learn and to then immediately apply the lessons learned to customized content and not only proprietary ones. The service was connected to the JSTOR Archive and to other content repositories, making ready-made datasets available through the Constellate Python client in the Lab for teaching and learning purposes. At the same time, it also allowed users to import external and customized collections of texts. In this sense, Constellate functioned as a sandbox for beginners, while also supporting more advanced researchers, who could benefit from TAP Institute classes led by humanities researchers and instructors presenting up-to-date methods and techniques.

However, the most valuable lesson offered by Constellate was the opportunity to become more independent and to approach technological developments critically, especially those concerning Artificial Intelligence. As AI becomes increasingly pervasive, understanding how computational text analysis works enables users to evaluate outputs, identify limits, and decide how such tools can be responsibly integrated into different research and teaching activities. Finally, the announcement of Constellate's closure in November 2024 created the urgent need to make the text analysis work sustainable over time and to develop a local alternative solution.

4. The “Introduction to text analysis” course

Since July 2025 a course called “Introduction to text analysis” was conceived and implemented by Università Cattolica's Library, reusing Constellate open education materials and adapting them to the local environment. In November 2025, one year after the Constellate's closure announce, the first module was officially launched on the local Blackboard platform, immediately proposed in class and completed by January 2026.⁵ The course is aimed at transferring knowledge in text analysis and suggesting how to apply text analysis to scholarship communications publications by extracting information from *corpora* through a data-driven approach to research, contributing to create and support the institutional research enterprise.

The knowledge transfer will be done gradually, offering different modules, starting with basics concepts and definitions and a preliminary course in Python. A particular attention will be given to copyright issues as we propose to work with scholarly communication products, or Library subscribed resources. However, it is foreseen that users would be interested in also building their own *corpora*, using for example retrieved OCRed texts or other digital texts, making this point particularly relevant. A research methodology protocol is proposed, starting from the Constellate “Skill Build” experience, and outlining a possible research design based on available textual data. This protocol will cover a ray of topics such as: Sentiment Analysis, Named Entity Recognition, Automatic Classification using LLMs, RAG and vector databases.

The main difficulty encountered in creating the course was that, while the ITHAKA/Constellate repository provided a strong pedagogical foundation, the transition from a fully supported platform to a locally managed environment required a significant effort in both learning and teaching simultaneously. Moreover, working locally required acquiring new technical

⁵ Cfr. UC-2024-003 “Introduction to Text Analysis”, accessible only to Università Cattolica del Sacro Cuore users at: https://blackboard.unicatt.it/ultra/courses/_155998_1/outline, retrieved on 23rd January 2026. So far 20 students enrolled to the online course ongoing in its contents and already at disposal.

competencies related to setting up, configuring, and maintaining an appropriate working environment for text analysis. One of the first challenges was deciding where the computational work should take place. This issue was not purely technical but deeply connected to copyright and intellectual property considerations. Because the course explicitly addresses copyright issues and works with scholarly communication products, often protected by intellectual property rights, the use of cloud-based environments raised concerns about compliance and data governance. In order to avoid potential legal and ethical issues related to working with protected materials in external cloud services, the decision was made to rely on the local environment. While this choice ensured greater control over data and compliance with copyright regulations, it also made the setup process more complex and time-consuming, both for instructors and learners.

Another significant difficulty was related to the heterogeneity of the target audience. Potential users of the course notebooks include students, early-career researchers, and more advanced scholars, each with very different levels of technical knowledge, familiarity with programming, and research needs. Designing materials that could be accessible without being overly simplified, while still remaining useful for more advanced users, required careful planning. Understanding the knowledge base, expectations, and learning goals of such a diverse audience proved to be an ongoing challenge rather than a one-time task. To address this issue, multiple learning paths were designed and continue to be refined. The course materials were organized into what we defined as “workflows” and ready-to-use notebooks adaptable to different textual materials with minimal modifications to the code. This approach lowers the entry barrier for beginners, while still allowing more experienced users to adapt and extend the workflows according to their research questions.

Rather than simply replicating an existing platform, the library capitalized on the experience gained through Constellate to enrich the course content with additional details, practical considerations, and locally relevant examples. This process allowed the preservation of methodological continuity, while at the same time fostering independence from a single commercial service. The library’s knowledge also played a crucial role in guiding users toward appropriate sources and datasets that comply with copyright and intellectual property regulations. Librarians’ expertise in licensed resources, open access materials, and public domain texts made it possible to clearly indicate where text analysis activities can be carried out and with which data, without breaching copyright law. This guidance proved essential, especially when working with scholarly literature and digitized texts, and helped translate abstract legal frameworks into practical research workflows.

These interactions with users provided continuous feedback on research practices, informing them about the evolution of the course and allowing it to move beyond the original Constellate tracks. As a result, the course was progressively shaped around real research questions and concrete use cases emerging from the academic community, rather than following a predefined or platform-dependent learning path (see paragraph 5 for an example).

Finally, to support sharing and reuse of the materials, the code developed for the course was made available through GitHub repositories. This choice allowed notebooks, workflows, and examples to be accessed, reused, and adapted both within and beyond the institution. Using

GitHub also enabled inexperienced users to approach coding gradually through the browser-based preview of notebooks, which, although not interactive, allows their contents to be viewed without downloading files or installing software locally. From there, participants could download the materials locally, or share them through their own GitHub accounts, increasing opportunities for collaboration and joint projects. Moreover, hosting materials on an open platform promotes transparency and reproducibility, enabling others to understand analytical steps and build upon existing work. This approach aligned with the course's broader objective of fostering independent, reproducible, and collaborative research practices. For the library, this represented a sustainable solution for maintaining and updating teaching resources over time, while reducing dependence on proprietary platforms and encouraging an open approach to text analysis learning.

5. From course foundations to research workflows: the pilot project on David Abosch

In parallel with the development of the course, a separate research pilot project was carried out on a corpus connected to the scholarly production of David Abosch (1922-[2020?]), an American author working on Japanese history and publishing mainly in the 1960s. The project investigates whether American historical scholarship written in the same period of time incorporated the author's research interests, and in particular if the "political" theme was present in such a literature. In the course materials the pilot project is therefore not presented as a classroom exercise, but as an external research application used to assess how the foundational methods introduced in the course can be transferred, extended, and later generalized into reusable workflows.

The pilot used a dataset created during the Constellate Skill Build Program, before the discontinuation of the service. The dataset consisted of 4,979 documents of which we analyzed a sample of 1,500 JSTOR records and included metadata and textual information in the form of unigram, bigram, and trigram counts. Although the Constellate Dataset Builder is no longer available, the same dataset can be asked to the new Jstor Archive Service Text Analysis Support.⁶ The dataset interest lies in its analysis and in the proposal, via the course, of the reproducibility of the downstream analytical workflow used: data inspection, preprocessing, lexical analysis, TF-IDF, topic modelling, classification, and interpretation.

The preprocessing stage in the pilot project followed the same methodological principles introduced in the course. The initial cleaning relied on the NLTK English stopword list. As subsequent frequency outputs still contained residual noise, related to the scholarly and bibliographic nature of the records, terms such as *university*, *journal*, *volume*, *vol*, *issue*, *number*, *page*, *pages*, *pp*, *doi*, *stable*, *org*, *www*, *http*, and *https* were therefore added manually to a custom stopword list. This list was compiled through an iterative procedure: first running the analysis with the standard stopword list, then inspecting the most frequent terms, identifying words that reflected publication metadata or scholarly formatting rather than thematic content, and finally updating the list.

⁶ Requests can be presented at: <https://www.jstor.org/ta-support/?ref=labs.jstor.org>. Details about how the corpus has been extracted are available in the research protocol published in the course materials.

In the following tables, we present some selected outputs from the pilot project notebook here presented as examples of the type of evidence generated by the analytical workflow. They also illustrate how the foundational steps emphasized in the course, such as corpus inspection, preprocessing, stopwords customization, and frequency analysis, can be extended toward more advanced exploratory techniques, such as TF-IDF and topic modelling.

Table 1 reports one example of the analytical outputs produced by the notebook after preprocessing.

Table 1 – Result of Most Frequent Words analysis and Japan-related term targeted search

	term	frequency		term	frequency_original
0	war	24817	1	japanese	6953
1	states	22282	0	japan	5792
2	united	20551	7	asia	2785
3	state	18198	8	asian	1532
4	american	18101			
5	international	17842			
6	see	16207			
7	law	15647			
8	military	15093			
9	world	14609			
10	history	14231			
11	first	13682			
12	general	13637			
13	economic	13198			
14	government	12510			

The Most Frequent Words analysis shows that the original corpus already contains a visible political, institutional, military, and geopolitical vocabulary, with terms such as *war*, *states*, *united*, *state*, *american*, *international*, *law*, *military*, *government*, and *foreign* among the most frequent lexical items. At the same time, a targeted search for Japan-related terms shows that words such as *Japanese*, *Japan*, *Asia*, and *Asian* are also present in the corpus, although they do not dominate its overall frequency profile. This suggests that the corpus is connected to Japanese and Asian studies, but that its lexical structure is strongly shaped by broader political and international-relations vocabulary.

While preprocessing and frequency-based exploration represent the core of the current course, the pilot project also allowed us to test more advanced techniques on a real research corpus. The aim is not to complete research on the specific Abosch's case itself, but to extract from it a reusable methodological sequence: comparing raw and weighted frequencies, identifying distinctive terms, exploring latent thematic structures, and assessing how preprocessing and filtering decisions influence interpretation.

TF-IDF was then used as an exploratory extension of the frequency analysis. While raw frequency highlights the most recurrent terms in the corpus, TF-IDF helps identify terms that are more distinctive within specific documents or subsets. This step is useful pedagogically because it shows learners that “importance” in a corpus cannot be reduced to frequency alone. The combination of different techniques proved useful for examining the same research question from complementary perspectives. The Most Frequent Words analysis provided a first overview of the dominant lexical profile of the corpus. It showed that Japan-related terms did not appear among the highest-frequency words. This result suggested that Japan-related vocabulary was relevant, but not dominant at the aggregate level. TF-IDF was then used to move beyond raw frequency and verify whether some terms, even when not globally dominant, were distinctive of specific documents or subsets of the corpus.

Table 2 – Result of TF-IDF analysis

	term	tfidf_score_sum	document_count	tfidf_score_avg		term	tfidf_score_sum	document_count	tfidf_score_avg
0	law	19.730977	801	0.024633	15	urban	12.817802	423	0.030362
1	soviet	19.281422	688	0.028025	16	population	12.343080	672	0.018368
2	naval	18.775947	566	0.033173	17	migration	12.340648	203	0.060791
3	military	18.064719	989	0.018266	18	labor	12.318456	503	0.024490
4	nuclear	17.965171	438	0.041016	19	aircraft	11.482704	413	0.027803
5	president	16.897374	710	0.023799	20	research	10.954147	951	0.011519
6	air	16.715305	723	0.023119	21	treaty	10.882095	466	0.023352
7	army	15.820998	726	0.021792	22	arms	10.734600	472	0.022743
8	war	15.356312	1217	0.012618	23	british	10.770599	984	0.010895
9	navy	15.299859	524	0.029198	24	committee	10.642392	723	0.014720
10	chinese	14.380140	537	0.026779	25	ships	10.591485	505	0.020973
11	international	13.967610	1100	0.012698	26	defense	10.551516	626	0.016855
12	economic	13.783816	982	0.014036	27	policy	10.527251	1007	0.010454
13	social	13.422653	946	0.014189	28	japanese	10.499460	881	0.011918
14	weapons	13.052563	438	0.029800	29	india	10.452772	637	0.016409

The TF-IDF results confirmed the relevance of legal, military, geopolitical and international-relations vocabulary. Terms such as *law*, *soviet*, *military*, *nuclear*, *president*, *war*, *treaty*, *arms*, *defense*, and *policy* point to thematic areas that are not only frequent, but also distinctive within the corpus. These results support the interpretation that the corpus contains a visible political and geopolitical dimension, while also offering a bridge toward more focused subset analysis.

Topic modelling was then used as a further exploratory step to examine whether the lexical patterns identified through frequency analysis and TF-IDF could be organized into broader thematic structures. While Most Frequent Words analysis shows which terms are most recurrent, and TF-IDF highlights terms that are distinctive within documents or subsets, LDA topic modelling allows the researcher to explore whether groups of words tend to appear together in ways that may suggest latent themes. In the pilot project, this step was particularly useful because the research question did not concern the presence of isolated terms only, but the possible emergence of a political, institutional, military, or geopolitical dimension within the corpus.

The exploratory topic modelling output confirmed some of the patterns already suggested by the previous analyses. Across different trials, topics related to war and military history, government and foreign affairs, and law and political institutions emerged with relative stability. For example, terms such as *war*, *military*, *world*, *british*, *states*, and *united* pointed to a war/military history dimension, while terms such as *minister*, *state*, *president*, *foreign*, and *government* suggested a government and foreign affairs topic. Similarly, terms such as *law*, *international*, *states*, *united*, *nations*, and *foreign* pointed to a legal and political-institutional dimension. These results were treated as exploratory rather than definitive, since topic modelling requires interpretive decisions concerning preprocessing, the number of topics, and the assignment of provisional labels to word clusters.

Table 3 – Result of LDA Topic Modelling (5 topics)

Topic 0		Topic 2		Topic 4	
war	0.0078	minister	0.0041	law	0.0089
military	0.0048	reported	0.0038	international	0.0076
american	0.003	state	0.0036	states	0.0076
first	0.0028	president	0.0036	united	0.0068
world	0.0028	foreign	0.0035	see	0.0046
british	0.0027	history	0.0035	state	0.0039
states	0.0027	government	0.0034	nations	0.0034
history	0.0026	since	0.0034	foreign	0.0032
time	0.0026	east	0.0034	general	0.0031
united	0.0026	pakistan	0.0033	committee	0.0029
Topic 1		Topic 3			
social	0.0044	john	0.0042		
economic	0.0031	state	0.0042		
research	0.003	president	0.0035		
population	0.0029	asst	0.0032		
see	0.0027	washington	0.0031		
study	0.0024	robert	0.003		
political	0.0023	box	0.0028		
development	0.0022	library	0.0028		
american	0.0022	william	0.0027		
group	0.002	american	0.0027		

Even if the purpose of this pilot projet was not to provide definitive findings, but to support the methodological development of the course. It shows how a foundational workflow based on corpus inspection, preprocessing, stop word customization, and frequency analysis can be extended toward more complex exploratory techniques, while preserving transparency, reproducibility, and critical interpretation.

6. Conclusion

At our university, librarians were the first to use the TDM pilot prototype and not only did they improve their knowledge, but they helped local users to better understand text analysis. As the service Constellate closed, education in text analysis has been possible thanks to librarians' role

as educators. This required librarians to expand their knowledge and activity beyond their traditional roles and expected responsibilities.⁷

However, they were not alone but supported by the movement that took place in the United States around the Hathi Trust Digital Library first and the Jstor Labs then, and other universities. It was a practical experience whose results and conclusions were in part collected in the book “*Text and Data Mining Literacy for Librarians*”, a book published in September 2025, by ACRL (Association of College and Research Librarians), a division of ALA (American Library Association). ACRL was the author in 2015 of the “Framework for Information Literacy for Higher Education”⁸, and it is understandable that ten years after it, it was time to move from Information literacy to Text and Data Mining Literacy for librarians. A similar evolution is valid for researchers, going beyond their traditional roles and experiences and integrating interdisciplinarity and trans disciplinarity techniques (or new and complementary research methodologies) in their research methodology, also possible by cooperating with text analysis-aware librarians. This will lead to a mutually beneficial research enterprise, composed not only of researchers but also librarians. It is only through a dialogue with researchers that librarians will be able to participate effectively in the building of the common research enterprise.⁹

References

Grimmer, J., Robert, M. E., Stewart B. M. (2022). *Text as Data: A new Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press

Hovy, D. (2020). *Text Analysis in Python for Social Scientists. Discovery and Exploration*. Cambridge: Cambridge University Press

Hovy, D. (2022). *Text Analysis in Python for Social Scientists. Prediction and Classification*. Cambridge: Cambridge University Press

Tuzzi, A. (2024). *Fondamenti di analisi dei dati testuali*. Roma: Carocci

Acknowledgements: Our warm thanks go to former Constellate project manager Amy Kirchhoff, former Constellate educators: Nathan Kelber and Zhou Chen, for all that we learnt.

⁷ Cfr. “Bibliotecari oltre. Attitudine alla transdisciplinarietà e competenze collettive”, 13 marzo 2025, Convegno delle Stelline, “Biblioteche oltre. I nuovi scenari dell’interdisciplinarietà”, Milano, 12-13 marzo 2025. The dedicated page in the Conference program is at disposal at: <https://www.convegnostelline.it/evento/bibliotecari-oltre-attitudine-alla-transdisciplinaria-e-competenze-collettive/>, retrieved on 23rd January 2026.

⁸ACRL “Framework for Information Literacy for Higher Education”, available at: <https://www.ala.org/acrl/standards/ilframework>, , retrieved on 23rd January 2026.

⁹ Cfr. Lavoie, Brian “Moving the library beyond the library: reflections on an RLP leadership roundtable”, 22nd August 2024, in “*Hanging Together*”, the OCLC Research Blog, at disposal at: <https://hangingtogether.org/moving-the-library-beyond-the-library-reflections-on-an-rlp-leadership-roundtable/>, retrieved on 23rd January 2026. See also Lavoie, Brian “The library beyond the library”, in “*Hanging Together*”, the OCLC Research Blog, at disposal at: “<https://hangingtogether.org/the-library-beyond-the-library/>”, retrieved on 23rd January 2026.

JADT Anthology

Decades of textual data analysis in one corpus

Vanni Laurent¹

¹UMR 7320: BCL - CNRS - UnivCA — Laurent.Vanni@univ-cotedazur.fr

Abstract

Since 1990, the JADT have assembled a collection of nearly 1,500 articles in the field of text data analysis, forming a rich and evolving corpus. The archiving and analysis of these scientific publications present a critical challenge for the research community, not only to map the trajectory of a rapidly evolving discipline but also to anticipate the emerging questions it raises Beaudouin (2016); Iezzi (2010); Vanni and Mittmann (2016). This article introduces the JADT-Anthology, a new archival website built on top of the Hyperbase Web platform’s search engine. The platform offers both a publicly accessible database and integrated statistical tools, enabling users to explore the JADT corpus in depth. To introduce the corpus and database, we present a study that traces the evolution of themes, methods, and tools that have defined the identity of the JADT over time. By analyzing temporal trends, we employ textual statistics to identify key terms, co-occurrences, and linguistic patterns. The results illuminate how researchers adapt to emerging technologies and how these adaptations shape the evolution of related research themes. Specifically, Correspondence Analysis reveal the formation of distinct temporal clusters. These clusters underscore a marked technological acceleration, evident in both the evolution of available resources — from the corpora of the 2000s to the big data era of the 2020s — and the analytical tools employed, ranging from Benzécri’s statistical methods to AI-driven platforms like ChatGPT. Our critical analysis seeks to interrogate — and redefine — research priorities in an era where AI is fundamentally reshaping the interpretability of computational outputs.

Keywords: JADT-Anthology, Corpus, Hyperbase, Statistics, Correspondences Analysis

1. Introduction

For several decades, analyzing scientific publications has been an effective way to observe the trends and tensions that shape academic research. These publications reflect changing methods and resources, as well as research that adapts to address emerging scientific challenges and questions by employing new approaches. Texts published following conferences or in scientific journals and books offer a constantly evolving field of exploration that illustrates the shifting of scientific paradigms. Thus, corpora of scientific publications provide researchers with insights from many different angles. Measurable linguistic variations in the texts provide clues about authors styles and the registers of publications characterized by different themes or disciplines. These variations are the subject of various quantitative and qualitative analyses that address fundamental linguistic questions, such as those concerning argumentation Binder et al. (2022); Lauscher et al. (2018); Poudat, C. (2006) or topic analysis Basta and Costa-jussa (2021); Mariani et al. (2018); Mesmoudi et al. (2024). Recently, research on Large Language Models (LLMs) has shown particular interest in scientific publications, especially for NLP tasks that aim to automatically synthesize or summarize scientific contributions Takeshita et al. (2024). Although many corpora already exist, publications produced by JADT and digitized over the past 20 years have not been widely used by humanities researchers. However, the texts published by this international conference exemplify exceptional traditional methods and expertise

in multilingual textual data analysis Beaudouin (2016). Nevertheless, some studies have demonstrated the value of JADT analyses using topological measures (graph analysis) Iezzi (2010) or co-occurrence approaches Vanni and Mittmann (2016). Updating and disseminating the JADT corpus is therefore crucial for exploring the history and evolution of a discipline with a strong international presence. In this study, we aim to provide the scientific community with a multilingual corpus containing all contributions published and digitized by the JADT since the 2000s, totaling over a thousand scientific articles. We will illustrate the corpus with brief diachronic and synchronic analyses of the three languages with the strongest representation: French, English, and Italian. Then, we will present the *JADT-Anthology* website¹, which is dedicated to analyzing, archiving, and disseminating the corpus to the community. Finally, we will discuss the prospects and resources mobilized to continue this work and offer the community powerful analysis tools.

2. Corpus

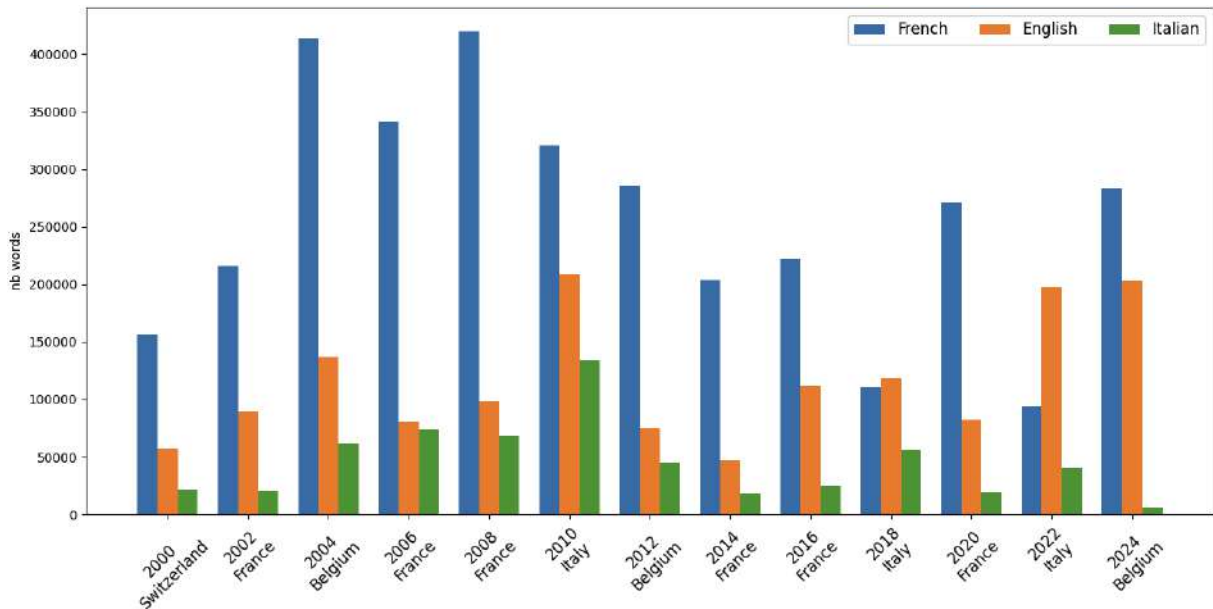


Figure 1 – Number of words per edition (year/location) and per language

Since its inception in the 1990s, the JADT has helped disseminate scientific research on text data analysis across multiple languages. Our corpus comprises over five million words spread across thirteen editions of the JADT from 2000 to 2024. The *.Doc* and *.PDF* formats, which have become widespread since the early 2000s, have allowed us to retrieve raw texts that we have standardized and tagged (encoding, cleaning, adding metadata, performing morphosyntactic analysis, etc.) in order to compare them. Editions published before 2000 are still available in print and must be scanned and OCR-processed before they can be incorporated into the corpus. To ensure consistency across the texts, this article organizes the data into three subcorpora divided into two periods. Each subcorpus represents one of the three languages we have chosen to analyze, totaling 5.4 million words: 3.3 million in French, 1.5 million in English, and 0.6 million in Italian. The distribution of word counts across the languages indicates that the French-speaking community behind the JADT has a strong presence in terms of publications

¹ Website accessible in 2026 via the URL : <https://jadtanthology.cnrs.fr>

in the various editions of this international conference. However, the reasons for this imbalance warrant further analysis, considering how word counts per language have evolved over the years and across host countries (Figure 1). First, we note that the distribution of word counts for certain languages is affected by the choice of host country. For example, English is more represented in the Belgian and Italian editions, while French editions favors French-language publications. Regardless of the host country, the use of French has been on a downward trend since 2012, offset by a marked increase in the use of English. Although English has not yet become as firmly established as it has in other international scientific communities, its use has steadily increased since the 2010s, if not the 2000s. This shift in communication strategy reflects an evolution in open research themes toward more Anglo-Saxon practices. This is the case with NLP, which has steadily increased in mention in the JADT since recent technological revolutions and the development of artificial intelligence (Figure 2). The next two sections use text statistics to accurately analyze the changes affecting scientific articles in the JADT. They examine the themes that unite the JADT community synchronically and track their evolution diachronically.

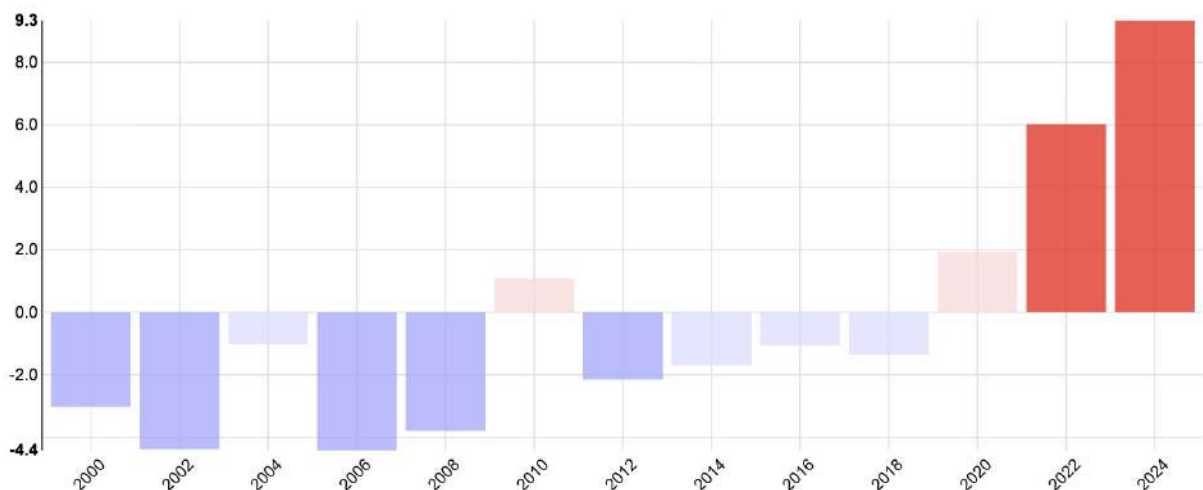
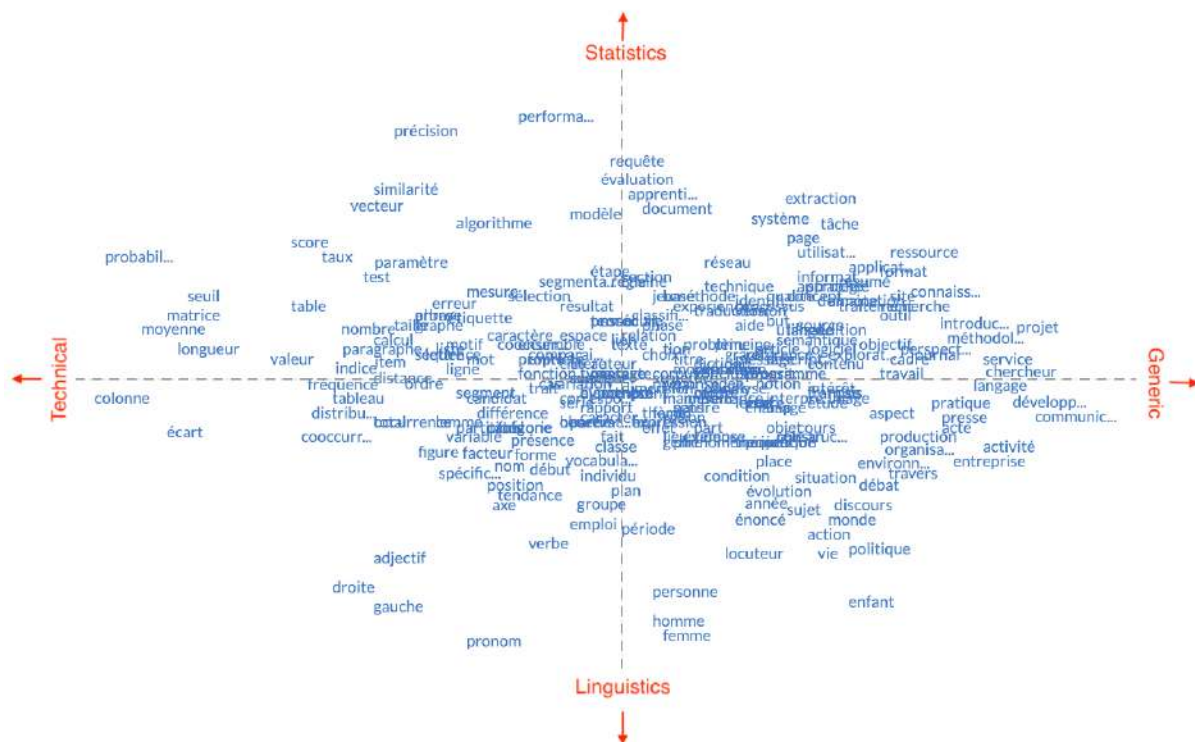


Figure 2 – Calculation of the Specificity, based on Fisher’s exact test (see Pincemin (2024)), of the term “NLP” (and “TAL” for French) in the entire JADT multilingual corpus.

3. Synchrony: A look at the intersection of linguistics and statistics

The analysis of themes within a corpus is the primary field of application for clustering methods such as hierarchical classification (ascending or descending), as illustrated by Reinert’s method Reinert, M. (1983), or Correspondences Analysis (CA) Benzécri (1973). We chose CA in particular because of its proven performance in generalised co-occurrence, which provides effective representations of isotopies Viprey (1997). Starting from a symmetric matrix representing the co-occurrence of the 300 most frequent nouns in each subcorpus (French, English and Italian) within a fixed 10-word window, we applied CA and projected the results onto the first two axes. Figures 3, 4 and 5 present the results for the three corpora.

Axis 1 and 2 are very similar across the three corpora analysed. Axis 1 (horizontal) contrasts technical vocabulary with general (or generic) vocabulary. On the other hand, axis 2 (vertical) contrasts vocabulary related to the natural sciences (textual statistics and computer science)



with that of the humanities (linguistics, sociolinguistics or social sciences). Specifically, the French corpus (Figure 3) shows that statistics are isotopic with words such as "probabilité" (probability), "colonne" (column), "écart" (deviation), "seuil" (threshold), "matrice" (matrix), "moyenne" (mean), "longueur" (length) extend towards the top of the graph, alongside the words "vecteur" (vector), "similarité" (similarity), "précision" (accuracy) and "performance" (performance). Further down, still in the technical section (on the left), are linguistic terms such as "adjectif" (adjective), "pronom" (pronoun), "verbe" (verb) and "nom" (noun). Note also the use of the terms "gauche" (left) and "droite" (right) which here refer to positions on graphs (represented by neighbour words such as "axe" (axis), "figure" etc.). The right side of the CA points to a broader vocabulary leaning toward statistics at the top - with the words "ressource", "extraction", "tâche" (task), "system" (system) and "apprentissage" (learning) — and toward linguistics and sociolinguistics at the bottom of the graph on the right - with "enfant" (child), "politique" (politics), "monde" (world), "discours" (discourse), "locuteur" (speaker), "énoncé" (utterance), "sujet" (subject), "personne" (person), "homme" (man) and "femme" (woman).

The English corpus (Figure 4) shows nearly identical isotopies to the French corpus. To the right of the CA, we see the technical vocabulary terms “row”, “column”, “weight”, “ratio”, “frequency”, “length”, “token”, “sequence”, and “matrix”. These terms extend downward in the graph toward statistical concepts, such as “precision”, “accuracy”, “classifier”, “training”, “query”, and “test”, and topward in the graph toward linguistic concepts, such as “adjective”, “noun”, “verb”, “lemma”, “form”, and “person”. To the left of the graph are broader terms that refer primarily to themes and generic terms related to scientific articles, such as “opinion”, “policy”, “medium”, “science”, “communication”, “life”, “perspective”, and “research” or “survey”. These terms lead to text mining at the bottom of the graph, which includes words

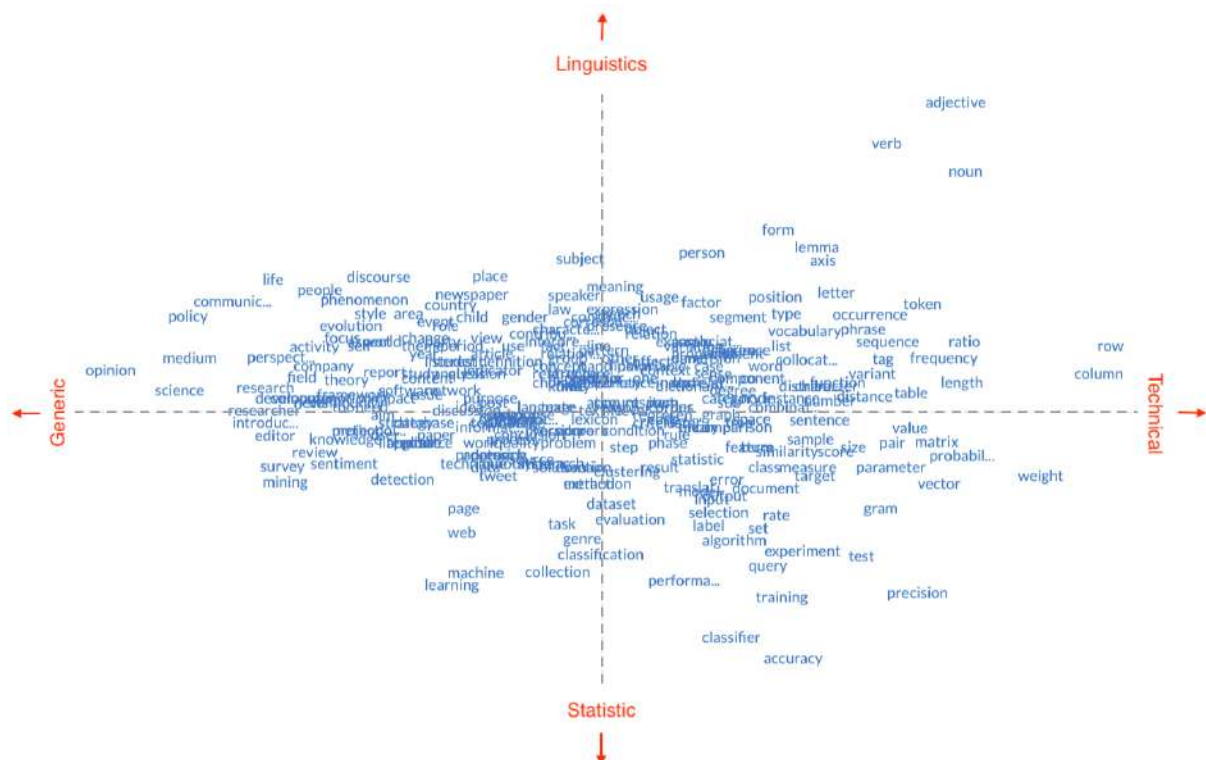


Figure 4 – CA of the co-occurrence matrix of the 300 most frequent nouns in the English corpus.

such as “mining”, “detection”, “machine”, and “learning”, or to linguistics at the top of the graph, which includes words such as “discourse”, “phenomenon”, “subject”, “speaker”, and “meaning”.

Finally, the CA of the latest Italian corpus (Figure 5) confirms the horizontal opposition (Axis 1) between technical vocabulary, represented by words such as “soglia” (threshold), “distribuzione” (distribution), “lemmo” (lemma), “lista” (list), “frequenza” (frequency) and “occorrenza” (occurrence), on the right, and generic vocabulary, represented by words such as “salute” (health), “disagio” (discomfort), “rischio” (risk) or “efficacia” (effectiveness), on the left. Axis 2, on the other hand, shows slight variation compared to the other corpora. The top of the graph leans mainly toward social science, with terms such as “destra” (right) and “sinistra” (left), which represent politics here, along with related words such as “politica” (politics) and “governo” (government), as well as other words such as “figlio” (son), “donna” (woman), and “giovane” (young). The bottom opens up to computer linguistic approaches with words such as “query”, “criterio” (criterion), “indice”, and “calcolo” (calculation) on the right, and “configurazione” (configuration) - which co-occurs statistically with “discorsiva” (discursive) (+14.18) - “repertorio” which co-occurs statistically with “discorsivo” (discursive) (+14.77) - and “efficacia” (effectiveness) - which co-occurs statistically with “valutazione” (evaluation) (+17.39) and “misurazione” (measurement) (+10.37) - on the left.

Although there are notable variations among the three CAs, the analysis of the three languages clearly demonstrates the homogeneity of the isotopies that define the JADT’s unique character. The highly technical statistical analysis fits perfectly with the more theoretical linguistics

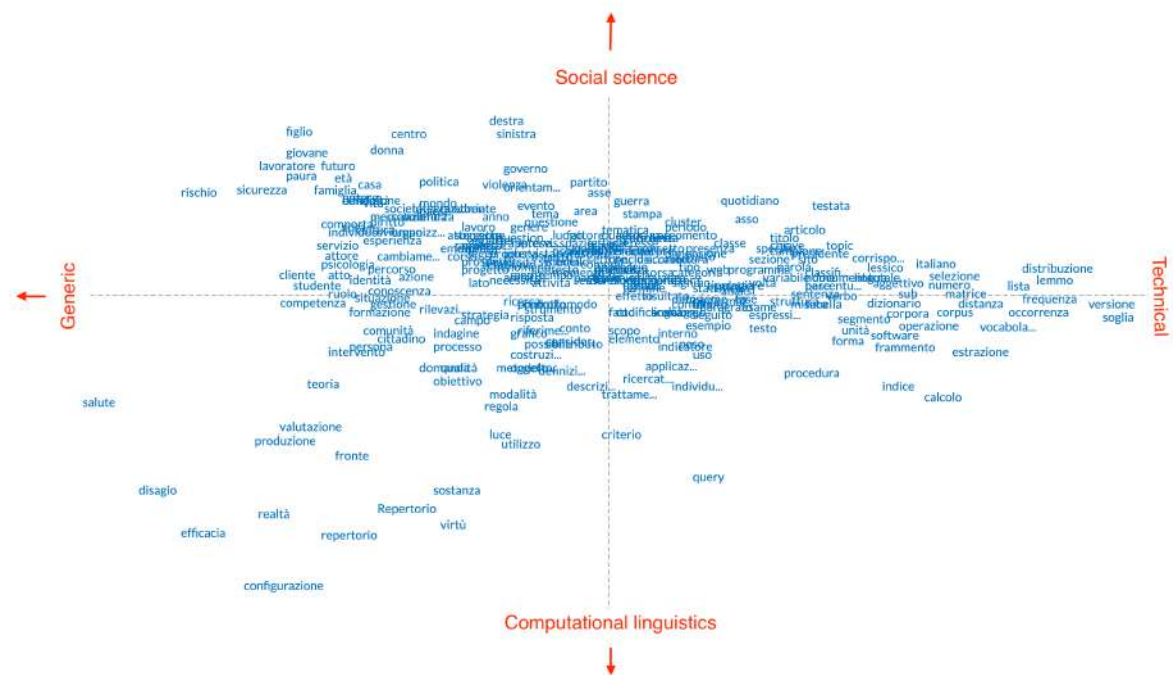


Figure 5 – CA of the co-occurrence matrix of the 300 most frequent nouns in the Italian corpus.

and sociolinguistics themes. For several decades, the JADT has been a privileged place where computational linguistics experts come together to discuss their work. This interdisciplinary approach offers unique perspectives and significant methodological proposals that contribute to the success of this international conference.

4. Diachrony: Toward a more generic discourse

The second analysis presented in this paper is diachronic. We divided each corpus (French, English, and Italian) into two comparable time periods. The first period covers six editions of the JADT spanning 12 years (2000–2012), and the second period covers five editions spanning 10 years (2014–2024). This analysis aims to explore the thematic variations present in the articles across these periods. To achieve this goal, we employed the historical calculation of specificities derived from Fisher’s exact test. The methodology and exploratory value of this test were recently reviewed in Pincemin (2024). Traditionally, this calculation yields a list of overused words in a subset of texts relative to the entire reference corpus. To summarize this list and align it with our topic analysis, we limited the calculation to nouns and selected the ten most distinctive terms that exceeded the minimum threshold of a +5 deviation from the corpus’s endogenous norm. The results are presented in Table 1.

Our proposed diachronic analysis reveals fairly consistent variations across the different corpora. The technical vocabulary associated with statistical methods and fundamental linguistics in the initial period is replaced by more general terms linked to the humanities and sociolinguistics, in particular. Of the ten most distinctive terms from the 2000–2012 period in the French corpus, seven are related to statistics and computational tools: “document”, “statistics”, “requête” (query), “automate” (controller), “information”, “vecteur” (vector) and “disponibil-

French		English		Italian	
2000-2012	2014-2024	2000-2012	2014-2024	2000-2012	2014-2024
document	classe	system	tweet	repertorio	topic
statistique	pétition	morpheme	topic	realità	cluster
requête	pourcent	tree	sentiment	configurazione	violenza
phrase	adolescent	noun	medium	modalità	web
automate	chanson	component	emotion	gesto	sentenza
information	immigration	register	embedding	psicologia	tweet
vecteur	figure	affix	network	intervento	e-mail
amorce	plongement	boundary	platform	cliente	pandemia
alignement	désaccord	code	review	psicoterapia	rilevatore
disponibilité	période	title	news	sostanza	sentiment

Table 1 – The ten most specific nouns from each language and period.

ité” (availability). Meanwhile, the other three words refer to linguistics: “phrase” (sentence), “amorce” (priming) and “alignment” (of multilingual corpora). In contrast, the second period (2014–2024) features significantly more thematic words, such as “petitions”, “adolescent” (teenager), “chanson” (song), “immigration” and “désaccord” (disagreement).

The picture is the same in English in the first period, with a slightly higher proportion of linguistic terms: “morpheme”, “noun”, “register” - which co-occurs statistically with “variation” (+14.34) and “linguistic” (+10.3) - and “affix”. The other words primarily point to the method: “system”, “tree”, “component”, “boundary” and “code”. As with French, the second period is clearly marked by research themes: “tweet”, “sentiment”, “emotion”, “review” and “news”.

Finally, the Italian corpus follows the same pattern with slightly more nuance. It features terms related to computational linguistics, such as “repertorio”, “configuration” and “modalità” (method) as well as thematic terms like “gesto” (gesture), “psicologia” (psychology), “cliente” (customer) or “psicoterapia” (psychotherapy) and a few more general terms “realità” (reality) and “sostanza” (substance). The second period features more thematic words, such as “violenza” (violence), “web”, “sentenza” (judgment), “tweet”, “e-mail”, “pandemia” (pandemic) and “sentiment” though technical and methodological terms like “topic”, “cluster” and “rilevatore” (detector) still play an important role.

The variations observed across the three corpora demonstrate a general shift toward more generic discourse that focuses on new themes and new corpora (“tweet”, “e-mail”). Twenty years ago, textual statistics and computational methods were prominent in demonstrations and discussions. Today, these methods and tools have diversified to incorporate machine learning and more NLP, as discussed in Section 2.. These new analytical tools are being applied for example to “embedding” methods (“plogement” in French list), as well as to the automatic detection of “topic”s or “sentiment” analysis in English and Italian. These new analytical tools are gradually shifting the focus of researchers towards applying increasingly complex methods in fields that extend beyond traditional linguistics and statistics, sometimes as a “black box”. However, as we saw in the previous section, the JADT’s identity continues to be defined by its unique methodological and interdisciplinary convergence, which brings together linguists, sociologists, mathematicians, and computer scientists around the same subject: text.

5. JADT-Anthology web site

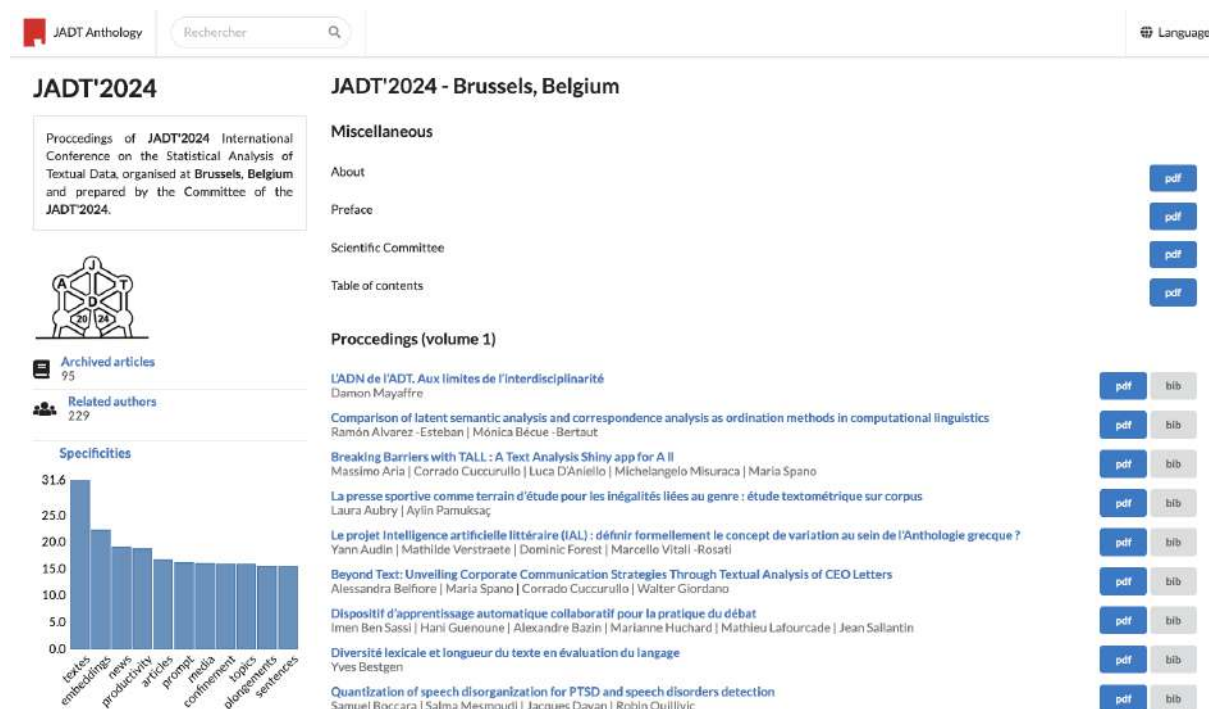


Figure 6 – JADT Anthology web site. Focus on 18th conference JADT'2024 - Brussels, Belgium : <https://jadtanthology.cnrs.fr/archives?edition=2024>

The French, English and Italian JADT corpora that we have produced are organised into three databases, plus a fourth multilingual one, all of which are accessible via the Hyperbase Web platform²³⁴⁵. Our work is accompanied by a website dedicated to archiving and analyzing publications in the JADT. We have named this website the “JADT Anthology”. It is hosted by the CNRS in France and can be accessed at : <https://jadtanthology.cnrs.fr>. This website aims to ensure the long-term success of the initiative launched by the founding members of the JADT⁶. It does so by making all of the papers presented at the various editions available to the community. The available format is *PDF*, as published in the proceedings. A set of metadata has been added to enable document search and retrieval. In addition to searching by year or title, you can search by author to access bibliographic references. The search engine uses a standard keyword-based system to organize results by edition and author. Each results page allows users to return directly to the text or view the abstract, which is automatically extracted from the *PDF* files. The search engine adds a statistical dimension by calculating lexical specificities, which enables users to explore how each metadata item (article, year, or author) relates to the reference corpus, or the entire collection of articles. Figure 6 shows an example focused on the 2024 edition of the JADT, which was held in Brussels, Belgium. The full contents of the proceedings are displayed, along with a bar chart showing the specific terms featured in this edition, such as “textes”, “news”, “embedding”, “productivity”, “articles”, “prompt”, “media”

² French Corpus : <https://hyperbase.unice.fr/jadtFR>

³ English Corpus : <https://hyperbase.unice.fr/jadtEN>

⁴ Italian Corpus : <https://hyperbase.unice.fr/jadtIT>

⁵ Multilingual Corpus : <https://hyperbase.unice.fr/jadt>

⁶ The proceedings of the JADT are available online at <http://lexicometrica.univ-paris3.fr>.

and “confinement” (lockdown). The article referencing system includes unique URLs for each article that can be shared, as well as suggested bibliographic citations in text and LaTeX formats. All of the results presented in this paper can be found on the JADT-Anthology website.

6. Conclusion

We have presented a corpus comprising all publications at the JADT since 2000, covering a period of over twenty years and 12 editions. This corpus, divided into sub-corpora by language, has revealed rich research opportunities from both a multilingual and a diachronic perspective. Lexical variations reflect thematic choices and registers (technical or generic) specific to each edition and each language. We have also proposed sharing this anthology of the JADT through raw textual data (available on the Hypebase Web platform), as well as via a website dedicated to archiving and preserving the proceedings of this unique international conference. However, the work reveals certain limitations in the quality of data automatically converted from *PDF* to *TXT* format. Several formatting effects, such as italics or bold, are not recognized by plain text, which ignores the corresponding words. Additionally, although headers and footers have been removed, they persist in some articles and may distort the measurements taken. To expand the analyses, it would be also advisable to digitize and incorporate the first four editions of the JADT (1990–2000) into the corpus. These editions mark an era when computing and internet access were still in their infancy. Analyzing this additional period would provide context for the past, present, and future work of the JADT.

References

- Basta, C. and Costa-jussa, M. R. (2021). Impact of COVID-19 in Natural Language Processing Publications: a Disaggregated Study in Gender, Contribution and Experience. In Chakravarthi, B. R., McCrae, J. P., Zarrouk, M., Bali, K., and Buitelaar, P., editors, *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 1–6, Kyiv. Association for Computational Linguistics.
- Beaudouin, V. (2016). Retour aux origines de la statistique textuelle : Benzécri et l’école française d’analyse des données. In *Statistical Analysis of Textual data - JADT 2016*, Statistical Analysis of Textual data - JADT 2016, pages 17–27, Nice, France. Mayaffre, D. Poudat, C., Vanni, L. et al.
- Benzécri, J.-P. (1973). *L’analyse des données: Benzécri, J.-P. et al. L’analyse des correspondances*. Dunod. Google-Books-ID: 0xTyzQEACAAJ.
- Binder, A., Hennig, L., and Verma, B. (2022). Full-Text Argumentation Mining on Scientific Publications. In Ghosal, T., Blanco-Cuaresma, S., Accomazzi, A., Patton, R. M., Grezes, F., and Allen, T., editors, *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 54–66, Online. Association for Computational Linguistics.
- Huguin, M. (2025). Comment créer, explorer et analyser un corpus de publications scientifiques avec Istex ? In *12èmes Journées de Linguistique de Corpus*, Lyon, France. Interactions, Corpus, Apprentissages, Représentations (ICAR) and Laboratoire de recherche en Linguistique et Didactique des Langues Étrangères et Maternelles (LIDILEM) and Dynamique Du Langage (DDL) and Praxiling, Université de Montpellier Paul-Valéry and Cognition, Langues, Langage, Ergonomie (CLLE).
- Iezzi, D. (2010). Topic connections and clustering in text mining: an analysis of the JADT network. In *10th Journées d’Analyse statistique de Données Textuelles (JADT’2010)*, pages 719–729, Rome, Italy.

- Lauscher, A., Glavaš, G., and Ponzetto, S. P. (2018). An Argument-Annotated Corpus of Scientific Publications. In Slonim, N. and Aharonov, R., editors, *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.
- Lebart, L., Pincemin, B., and Poudat, C. (2019). *Analyse des données textuelles*. Presses de l’Université du Québec.
- Lebart, L., Salem, A., and Berry, L. (1998). *Exploring Textual Data*, volume 4 of *Text, Speech and Language Technology*. Springer Netherlands, Dordrecht.
- Mariani, J., Francopoulo, G., and Paroubek, P. (2018). Measuring Innovation in Speech and Language Processing Publications. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Mayaffre, D. (2024). L’ADN de l’ADT. Aux limites de l’interdisciplinarité. In *JADT 2024*, number 10, pages 17–28, Bruxelles, Belgium. Presses universitaires de Louvain.
- Mesmoudi, S., Dégeilh, F., Cancino, W., Rodic, M., Peschanski, D., and Eustache, F. (2024). Transdisciplinary method for exploration and visualization of neuroimaging papers and databases: Application to PTSD. *Annales Médico-psychologiques, revue psychiatrique*, 182(10):908–917.
- Nozza, D. and Hovy, D. (2023). The State of Profanity Obfuscation in Natural Language Processing Scientific Publications. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3897–3909, Toronto, Canada. Association for Computational Linguistics.
- Pincemin, B. (2024). Specificities and other applications of the Fisher’s exact test to textual data: What’s the matter with lexical frequencies? In Dister, A. and Longrée, D., editors, *JADT 2024 : Mots comptés, textes déchiffrés*, volume 2 of *Cahiers du Cental*, pages 703–712, Bruxelles, Belgium. Presses universitaires de Louvain.
- Poudat, C. (2006). *Etude contrastive de l’article scientifique de revue linguistique dans une perspective d’analyse des genres*. ED Humanités et Langues (H&L), Université d’Orléans.
- Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l’analyse lexicale par contexte. *Cahiers de l’analyse des données*, 8(2):187–198.
- Takeshita, S., Green, T., Reinig, I., Eckert, K., and Ponzetto, S. (2024). ACLSum: A New Dataset for Aspect-based Summarization of Scientific Publications. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.
- Teich, E. (2010). Exploring a corpus of scientific texts using data mining. In *Corpus-linguistic applications*, pages 233–247. Brill. Section: Corpus-linguistic applications.
- Vanni, L. (2024). Hyperbase Web. (Hyper)Bases, Corpus, Langage. *Corpus*, 25.
- Vanni, L. and Mittmann, A. (2016). Cooccurrences spécifiques et représentations graphiques, le nouveau “Thème” d’Hyperbase. In *13es Journées internationales d’Analyse statistique des Données Textuelles (JADT 2016)*, volume 1, pages 295–305, Nice, France.
- Viprey, J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*,. Honoré Champion, Paris.

ActiveTigger: An open source collaborative text annotation software for computational social sciences

Émilien Schultz^{1,2}, Julien Boelaert³, Axel Morin², Annina Claesson^{2,4}, Emma Bonutti D’Agostini^{2,4}, Arnault Chatelain², Étienne Ollion^{5,2}

¹Groupe ENSAE-ENSAI – emilien.schultz@ensae.fr

²CREST

³Université de Lille

⁴Sciences Po Paris

⁵CNRS

Abstract

ActiveTigger is an open source software tool designed to support collaborative text annotation for computational social scientists. It implements in a user-friendly interface machine learning features such as training classifiers, fine-tuning language models, evaluating performance, and running predictions on large datasets. In particular, it implements active learning as an intermediate step between human annotation and machine learning. This article reviews its origins and developments, introduces its main features, and discusses its current limitations and future directions.

Keywords: document-level annotation, active learning, natural language processing, computational social sciences, open source software

Résumé

ActiveTigger est un logiciel open source conçu pour faciliter l’annotation collaborative de textes des chercheurs en sciences sociales computationnelles. Il implémente dans une interface conviviale des fonctionnalités de machine learning telles que l’entraînement de classificateurs, le fine-tuning de modèles, l’évaluation de leurs performances et la prédiction sur de grands corpus. En particulier, il propose l’apprentissage actif comme étape intermédiaire entre l’annotation humaine et l’automatisation. Cet article revient sur son histoire, présente ses principales fonctionnalités et discute ses limites actuelles et l’avenir du projet.

Mots clés : annotation à l’échelle du document, apprentissage actif, traitement automatique du langage, sciences sociales computationnelles, logiciel open source

1. Introduction

ActiveTigger¹ is an open source software tool designed to support collaborative text annotation for computational social scientists. It aims to streamline and accelerate the annotation of text corpora by offering a user-friendly interface for collaboration on a shared server. The interface is designed to promote best practices for supervised machine learning, as documented in both social science and computer science literature (Do et al., 2024).

The tool grew out of the day-to-day practices of social science researchers, responding to a growing need to bring structure and consistency to their workflows. In research settings, such

¹ An initial version was developed by J. Boelaert. The name is a pun that between the words "Tagger" and "Tigger".
We acknowledge funding from DRARI, ANR, and Progedo.

tools must meet standards of reproducibility and auditability, handle sensitive data, and run on the limited hardware typical of academic infrastructure. In the course of its development, ActiveTigger also proved valuable as a teaching tool, allowing students and researchers without coding experience to delve into the logic of machine learning.

Since 2023, ActiveTigger has been used in dozens of research projects. This community of practitioners was instrumental in the development of the current tool, which culminated in the release of a first public version (v1.0.0) in May 2026. This article reviews the developments of the software, presents its main features, and discusses its future.

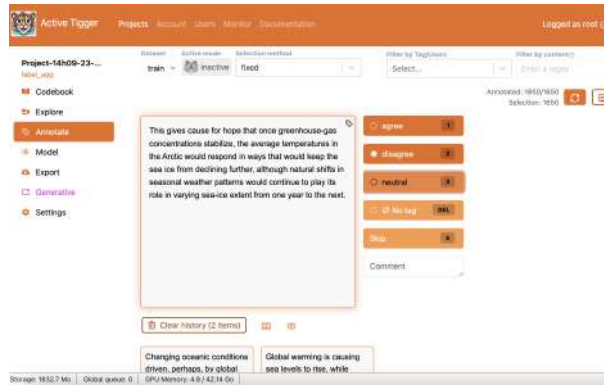


Figure 1: ActiveTigger main annotation panel

2. Motivations

Textual data sources are increasingly abundant, and so are the methods available to analyze them. Natural Language Processing (NLP) techniques for studying large-scale textual corpora have multiplied across a wide range of disciplines.

Released in 2017, the Transformer architecture (Vaswani et al., 2017) formed the backbone of contemporary language models² that changed the approach to text analysis. Together with the Python ecosystem and with expanding computational resources, these changes brought NLP within reach of researchers without formal training in computational linguistics. Dedicated methodological frameworks for the social sciences followed (Do et al., 2024).

Non-expert users often struggle with fundamental concepts in machine learning, such as the distinction between training, validation, and test phases. A dedicated software can lower these learning barriers by mediating between models and their users: “models and software are entangled in science, and software does critical work that models cannot perform on their own” (Hocquet et al., 2024).

Due to this growing interest, many social scientists have turned to generalist software to carry out their annotation tasks, and increasingly to easy-to-use generative models. This creates a heavy reliance on private companies, and raises concerns about the protection of sensitive data and GDPR compliance—concerns that call for on-premises solutions. Another reason for developing purpose-built tools is that, despite improvements in generative AI services, fine-tuned models trained on human annotations remain preferable for many use cases (Obeidat et al., 2025). By developing ActiveTigger, we hope to demonstrate that purpose-built tools — and the methodological choices encoded in their interfaces — are essential for making NLP-based research replicable (Barrie et al., 2024).

² Initially non-generative encoders.

3. Similar software for annotation

Annotation tasks in the humanities and social sciences encompass a wide range of research processes. Researchers have developed a variety of tools to support these activities (Colucci Cante et al., 2024). These tasks lie along a continuum, from the qualitative analysis of a limited number of long documents that require in-depth coding (such as interviews or reports), to the analysis of large datasets of short texts (such as newspaper articles or social media posts). This diversity also involves different levels of annotation (document-level, sentence-level, or token-level), all of which directly shape tool design.

Beyond differences in document type and granularity, annotation tasks themselves vary widely, including classification, span detection, relation extraction, and more. To achieve those annotations, even if spreadsheets are still useful (Moncada, 2025), a broad range of tools can be used, from Computer-assisted qualitative data analysis software (CAQDAS) such as QualCoder to computer-science-oriented annotation tools such as Label Studio, Prodigy, Doccano, or Argilla. More recently, even dedicated interface for zero-shot classification has been proposed by (Sprengholz, 2026).

Existing software tends to be general-purpose, offering a wide range of customizable options. However, highly configurable tools can overwhelm non-expert users in the social sciences, leading to suboptimal parameter choices or even causing users to abandon the tool altogether. An alternative approach narrows the option space by offering sensible defaults from the outset, lowering the barrier to entry before gradually exposing customization options.

More recently, this ecosystem of tools has begun to respond to advances in language models, covering both encoders and generative AI. Tools such as INCEpTION have developed dedicated “recommenders” to incorporate model predictions into the annotation process (Eckart De Castilho et al., 2024). However, integrating generative AI into annotation workflows remains challenging (Schroeder et al., 2025), particularly with respect to evaluating model performance.

ActiveTigger sits at the scaffolded, integrated, encoder-oriented corner of this space of existing tools. It aims primarily to provide the integration of all steps of the annotation process within a single platform: data import, exploration, annotation, model training, evaluation, and export. This end-to-end integration avoids the friction of switching between multiple tools and ensures methodological consistency throughout the workflow.

4. Main use cases

AT was initially built as a no-code interface to annotate textual data that includes ML-based strategies to accelerate annotations (active learning³ chief among them). It was meant to support the full pipeline of training, evaluating, predicting and exporting. Currently ActiveTigger allows to:

- **Optimally annotate a large corpus:** Users equipped with a few annotated documents can fine-tune models from within the annotation interface to optimally choose subsequent documents to annotate with active learning.
- **Supervise and coordinate multi-annotator campaigns:** Project owners can manage multiple annotators accounts at the same time. These annotators account can be given

³ Active learning selects the most informative instances for annotation during training (Ein-Dor et al., 2020; Miller et al., 2020).

restricted access.

- **Collaboratively stabilize a codebook:** Teams can work out a common codebook on the platform by organizing rounds of annotations and using the reported inter-annotator agreement scores.
- **Quickly identify and retrieve documents of interest:** ActiveTigger can output topic models (BERTopic) on the corpora and allows navigating them with regex filters.
- **Teach NLP:** ActiveTigger's design allows student to gain a quick hands-on experience of the different steps involved in an NLP research pipeline, without prior coding or machine-learning knowledge.

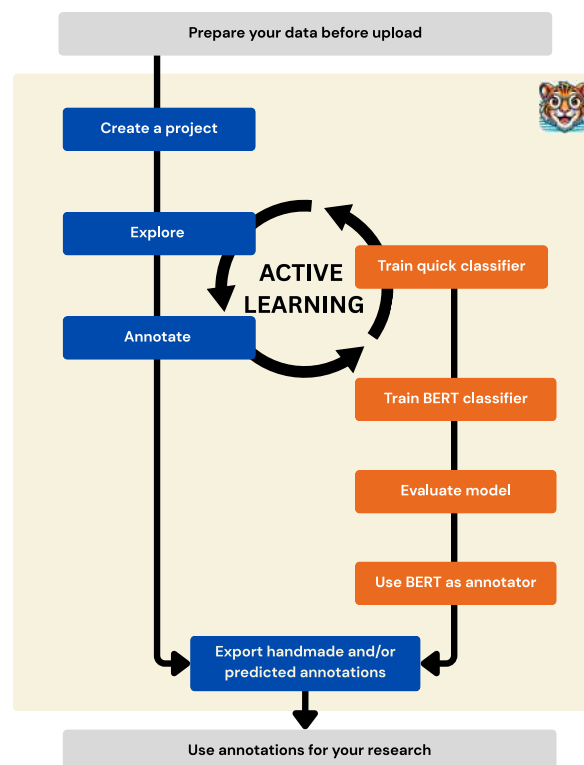


Figure 2: Main workflow of ActiveTigger

THE ANNOTATION PROCESS OF BOELAERT ET AL. (2025) TO DETECT GENDER-RELATED RESEARCH IN SOCIAL SCIENCE PAPERS' ABSTRACT.

The researchers started with uploading the data on ActiveTigger and splitting it into a train, a validation and a test set. Several co-authors then began annotating the documents following a first codebook on gender-related research. Several dozens of abstracts chosen at random were first annotated. Then a quick model was trained on them and the subsequent annotations were made on the abstracts the model found more difficult to classify (active learning). Over the course of several weeks the team iteratively stabilized the codebook, redefining the labels and documenting how to treat hard-to-classify examples. In a last step AT was used to FT a BERT model on the latest labeled examples. 600 annotations proved enough to build a classifier with an F1 score on the test set comparable to that of an expert annotator (above 90

5. Main features

The following subsections describe the main features, organised by functional area.

5.1. *Project setup and data management*

ActiveTigger is organized around projects. Each project is created from a dataset (Parquet, CSV or XLSX) and can include a configurable split into train, validation, and test sets—in keeping with best practices in supervised learning, whereby the split is performed *before* annotation begins. The platform supports dataset stratification to ensure representativeness across subsets. Projects can be shared across users, and existing annotations can be imported at project creation or added later.

5.2. *Annotation interface*

The annotation interface supports two types of schemes: multiclass (single label per text) and multilabel (several labels per text). Each scheme is associated to a codebook page that can be used to document the annotation guidelines. The codebook page can be viewed at any time while annotating documents. Contextual information from the original dataset can be displayed alongside the text to assist annotators. The interface maintains an annotation history and supports filtering by label, user, regex pattern, or model prediction. Comments can be attached to individual annotations.

5.3. *Text representation (features)*

Several methods are available to compute numerical representations of texts, which are used for visualization, topic modeling, and training lightweight models (scikit-learn classifiers, see Section 5.5): Sentence embeddings using transformer models via the sentence-transformers library, Word embeddings using fastText, averaged over all words in each document, Document-feature matrices (DFM), Boolean features based on regular expressions and Numerical or categorical variables from the original dataset.

5.4. *Selection strategies and active learning*

Multiple strategies are available to select the next element to annotate. Beyond random and sequential selection, the platform implements active learning methods based on model predictions: selection by maximum entropy in general or for specific labels (to focus on ambiguous cases), by maximum predicted probability for a given label (to find specific categories), or by minimum probability among predicted positives. Filters by regex, label, user, or prediction can

be combined with these strategies.

5.5. Model training

Two families of classifiers can be trained within the platform:

- **Quick models:** standard machine learning classifiers from scikit-learn (logistic regression with L1/L2 penalty, k-nearest neighbors, random forest, multinomial naive Bayes), trained on pre-computed features. They are lightweight and support 10-fold cross-validation.
- **BERT models:** transformer-based classifiers fine-tuned through the Hugging Face framework. The interface exposes key hyperparameters (learning rate, epochs, batch size, weight decay, gradient accumulation) and displays the training loss curve to help diagnose overfitting or underfitting. The list of available base models is configurable by the administrator.

5.6. Evaluation

Model performance is evaluated with precision, recall, and F1 score per class, as well as macro F1 across all classes. These metrics can be computed separately on the training, validation, and test sets. The interface displays confusion matrices and allows users to inspect misclassified examples. For BERT models, the loss curve is displayed for both the training and evaluation subsets. Predictions on the validation and test sets are only computed on demand to avoid influencing the annotation process.

5.7. Data exploration

The Explore page provides three complementary views of the corpus:

- A **tabular view** with filtering by label, user, and regex, allowing direct annotation from the table.
- A **2D projection** of the embedding space using UMAP or t-SNE, where users can select regions and lock them to constrain subsequent annotation to specific areas of the corpus.
- A **topic model** based on BERTopic (Grootendorst, 2022), with configurable parameters (UMAP neighbors, minimum topic size, outlier reduction). Topics can be converted into annotation schemes, and detailed reports can be exported.

5.8. Collaborative features and user management

The platform implements role-based access control at two levels: account type (root, manager, annotator) and project-level rights (manager, contributor, annotator). Multiple users can annotate the same dataset simultaneously, enabling inter-annotator comparison. The codebook serves as a shared reference for maintaining annotation consistency across team members.

5.9. Inference and export

Once a model is trained, ActiveTigger can apply it to the complete dataset or to an external file imported by the user. All project artifacts can be exported: annotated datasets (per set or globally, all schemes), computed features, 2D projections, trained BERT models, and predictions on the project or external datasets.

5.10. Generative models (experimental)

An experimental module allows users to query external large language models through their API (OpenAI, OpenRouter, Ollama, Hugging Face or Ilaas⁴) to annotate or extract information using custom prompts (similar to Sprengholz (2026)). Prompts can reference the text and contextual variables from the dataset. Results are stored and can be downloaded for further analysis. This feature is designed for exploration and comparison with human annotations rather than as a replacement for the supervised workflow.

6. Architecture description

The codebase is available in the GitHub repository: <https://github.com/activetigger/activetigger/>

The software follows a client/API architecture. The API is developed in Python, with FastAPI for the routes and a PostgreSQL database. The main client is a React/TypeScript frontend using the Vite framework. A Python client is also available to automate specific tasks such as account creation and project management.

The application is currently designed to be deployed either at laboratory scale or on premises (~30 simultaneous users / ~10 projects). ActiveTigger can be installed on premises or on a server using Docker and Docker Compose.

7. Validation

Since the workflow relies heavily on user interaction, the full pipeline cannot easily be evaluated end-to-end, nor directly compared to tools that address only a single step such as annotation or fine-tuning. We therefore restrict the reproducibility evaluation to the main computational task, namely fine-tuning a BERT model. Beyond this formal evaluation, ActiveTigger is already in use in several research projects, including published studies such as Boelaert et al. (2025) and Claesson (2026), which draw on different parts of the platform under real research conditions.

7.1. Reproducibility

The core workflow consists of fine-tuning a BERT model on annotated examples, then applying it to label the remaining dataset.

To validate this step, we use a published dataset for detecting stance on global warming in media (Luo et al., 2021): 2,050 documents annotated by multiple humans and reconciled by the original authors, with three labels (agree, neutral, disagree). The original study splits the data into 1,850 training and 200 test examples and reports a BERT-base⁵ baseline (learning rate $2e-5$, 7 epochs) with an average accuracy of 71% [0.64–0.77] and per-class F1 of 0.68 (agree), 0.76 (neutral), and 0.75 (disagree).

With the same parameters on our development instance (a slice of an A100 GPU, 40 GB), we trained two models⁶ using similar settings. The results are comparable: with accuracy scores ranging between 0.69 and 0.72 (macro F1 of 0.68 and 0.72); both runs fall within the confidence interval reported by the original authors. We also added this dataset to the project creation panel so that users can reproduce these results.

⁴ <https://www.ilaas.fr/>

⁵ <https://huggingface.co/google-bert/bert-base-uncased>

⁶ We did not use weighted data, since weighted cross-entropy loss is not currently implemented in ActiveTigger

7.2. Robustness check

To test whether the software scales, we deployed it on a new OVH virtual server (Cloud GPU with an RTX5000, 16 GB VRAM, 28 GB RAM, 4 vCores, with Ubuntu installed). The complete installation (Docker, drivers, and the app) took about 13 minutes.

We used the Python client to conduct the following stress tests⁷:

Test scenario	Result
Response time (no active users)	Typical response time: 30–50 ms
Create 10 projects	All projects were created and deleted without errors
Create 10 projects, each triggering BERT training	The job queue handled the sequence correctly Maximum delay observed < 0.8 s
Create 1 project with 10 users annotating simultaneously	All operations completed successfully Maximum delay observed < 0.4 s

Table 1: Stress test results

8. Discussion and future of the project

8.1. ActiveTigger in the research community

Over the past two years, a growing community of early adopters has formed around ActiveTigger, spanning multiple research teams and disciplines. Several projects conducted within the Computational Social Science team have already led to publications: Boelaert et al. (2025) analyzed the evolving prominence of the concepts of gender, class, and race in the French social sciences literature, while Claesson (2026) examined online interactions with members of parliament on Twitter. These studies share a common methodological approach: grounding a research question in the relevant literature and then using supervised classifier training to scale researcher annotations to large corpora.

Beyond the core team, researchers and master's students from various institutions are currently using the software to explore new approaches to their data. A frequent use case involves training classifiers to reduce large corpora to more manageable subsets — for example, extracting articles on a specific topic from a comprehensive press database. Thanks to the development instance hosted at CREST, users can access a GPU remotely to train their models, and then continue their analysis locally on a personal computer. This lightweight infrastructure lowers the barrier to entry and encourages experimentation across a wide range of projects.

ActiveTigger has also served as a focal point for broader methodological discussions in the social sciences. ActiveTigger's workflow proved consistent with the social science tradition of engaging closely with one's data, while harnessing the power of open-weight, pre-trained encoder models. It has sparked discussions on several fronts: how to design research-oriented software tools, how the boundary between qualitative and quantitative methods is shifting, and what kinds of AI tools are best suited to social sciences. New features are tested along research projects to expand the tool's capabilities.

⁷ <https://github.com/activetigger/at-cli-client>, with different test playbooks in /test

8.2. Current limitations and future roadmap

ActiveTigger is first and foremost a research tool operating in a rapidly evolving field. This position has two implications: the tool must continually adapt to the changing practices of the research community, and its development must keep pace with evolving software engineering best practices. To support long-term sustainability, the project is developed openly on GitHub, with a public issue tracker, contribution guidelines, and a roadmap that invites participation from both researchers and developers.

The evolving nature of the software poses a challenge for systematic evaluation, as the tool itself has changed between different projects, limiting direct comparisons across versions. Establishing a stable evaluation framework—ideally including reference datasets and versioned benchmarks—remains a key priority going forward.

At the same time, recent advances open up promising avenues for the project, including zero-shot and few-shot text classification, stronger text embeddings, and cross-modal text-image annotation. Because ActiveTigger builds on generic Python libraries such as transformers and sentence-transformers, integrating newer embeddings and models is straightforward. This nonetheless raises the question of how to balance model size against the computational cost of fine-tuning, particularly in the resource-constrained settings typical of academic research.

A related question is which new tasks to prioritize, such as Named Entity Recognition, classical lexicometric features, and multimodal workflows. The Generative Panel, still under development, will need to be developed in close dialogue with the social science community to meet the specific needs of social science research. A further step would be to experiment with fine-tuning generative models within the platform.

On the software engineering side, ActiveTigger currently supports single-GPU deployments serving up to a few simultaneous projects and a few dozen users. We are refactoring the task management system to improve robustness and better allocate computational resources—a prerequisite for scaling to larger user communities and supporting multi-GPU configurations. This effort will be accompanied by expanded unit and end-to-end testing to ensure software quality and prevent regressions. Reproducibility is another priority: project exports are currently limited to specific contents (models and annotations), and we aim to extend them to capture full project state, including annotation history, trained model checkpoints, and configuration parameters, so that entire annotation campaigns can be archived and re-executed.

Taken together, these directions reflect the broader ambition of the project: to provide the social science community with NLP tooling that is both methodologically rigorous and operationally reproducible, while remaining accessible to researchers without deep machine learning expertise.

9. Funding & Acknowledgements

ActiveTigger was supported by a grant from DRARI Île-de-France. It also benefited from the support of Progedo and from the ANR (Pantagruel) for its dissemination.

Ouestware (Paul Girard) contributed to software development and key technical decisions.

The *Groupe ENSAE-ENSAI* hosted a development instance that was made available to early users.

References

- Barrie, C., Palmer, A., and Spirling, A. (2024). Replication for Language Models. Preprint.
- Boelaert, J., Coavoux, S., Delaine, E., Despres, A., Gollac, S., Keyhani, N., Momméja, A., and Ollion, É. (2025). La part du genre: Genre et approche intersectionnelle dans les revues de sciences sociales françaises au XXI^e siècle. *Actes de la recherche en sciences sociales*, N° 258-259(3-4):126–145.
- Claesson, A. (2026). Le prix de la visibilité: Une analyse computationnelle des interactions en ligne avec des député-es français-es. *Revue française de science politique*, Vol. 75(3):IX–XVII.
- Colucci Cante, L., D'Angelo, S., Di Martino, B., and Graziano, M. (2024). Text Annotation Tools: A Comprehensive Review and Comparative Analysis. In Barolli, L., editor, *Complex, Intelligent and Software Intensive Systems*, volume 87, pages 353–362. Springer Nature Switzerland, Cham.
- Do, S., Ollion, É., and Shen, R. (2024). The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. *Sociological Methods & Research*, 53(3):1167–1200.
- Eckart De Castilho, R., Klie, J.-C., and Gurevych, I. (2024). Integrating INCEpTION into larger annotation processes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 110–121, Miami, Florida, USA. Association for Computational Linguistics.
- Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., and Slonim, N. (2020). Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure.
- Hocquet, A., Wieber, F., Gramelsberger, G., Hinsén, K., Diesmann, M., Pasquini Santos, F., Landström, C., Peters, B., Kaspróicz, D., Borrelli, A., Roth, P., Lee, C. A. L., Olteanu, A., and Böschén, S. (2024). Software in science is ubiquitous yet overlooked. *Nature Computational Science*, 4(7):465–468.
- Luo, Y., Card, D., and Jurafsky, D. (2021). Detecting Stance in Media on Global Warming.
- Miller, B., Linder, F., and Mebane, W. R. (2020). Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches. *Political Analysis*, 28(4):532–551.
- Moncada, M. (2025). Should we use NVivo or Excel for qualitative data analysis? *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 165–166(1-2):186–211.
- Obeidat, M. S., Sultan Al Nahian, M., and Kavuluru, R. (2025). Do LLMs Surpass Encoders for Biomedical NER? In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, pages 352–358, Rende, Italy. IEEE.
- Schroeder, H., Roy, D., and Kabbara, J. (2025). Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks.
- Sprengholz, P. (2026). Annotaid: A browser-based tool for LLM-assisted qualitative coding of open text. *SoftwareX*, 34:102702.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need.

ParaTaxe: a Suite of Tools for Processing and Publishing Multi-parallel Corpora

Olivier Kraif¹, Cyrille François², Elnaz Jalilian¹, Natacha Rimasson-Fertin³,
Pascale Roux⁴

¹ LIDILEM, Univ. Grenoble Alpes – olivier.kraif@univ-grenoble-alpes.fr,
elnazjalilian@gmail.com

² CIEL, University of Lausanne – cyrille.francois@unil.ch

³ ILCEA4, Univ. Grenoble Alpes –natacha.rimasson-fertin@univ-grenoble-alpes.fr

⁴ Passages XX-XXI, Univ. Lyon 2 – pascale.roux@univ-lyon2.fr

Abstract

Over the past fifteen years, digital tools for parallelizing and aligning textual corpora have developed significantly within the field of digital humanities. Several projects and software tools have been created to compare different translations or versions of texts. Despite these technical advances, many parallel corpora produced in academic contexts quickly become inaccessible online due to the lack of maintenance or funding. To address this gap, we present a suite of open-source tools developed within the ACR Grimm Tradalign and ParaTaxe projects. This suite includes the AAlign aligner, based on multilingual embeddings, a Python processing pipeline for converting and aligning corpora, a web interface called ParaTaxe for managing collections of parallel corpora, and a web interface for exploring aligned corpora built with TEI-Publisher.

Keywords: multi-parallel corpora, parallel texts alignment, XML-TEI publication platform.

Abstract

Au cours des quinze dernières années, les outils numériques permettant de paralléliser et d'aligner des corpus textuels se sont largement développés dans les humanités numériques. Plusieurs projets et logiciels ont été créés pour comparer différentes traductions ou versions de textes. Malgré ces avancées techniques, de nombreux corpus parallèles produits dans un cadre académique deviennent rapidement inaccessibles en ligne, faute de maintenance ou de financements. Pour répondre à cette lacune, nous présentons une suite d'outils open source développés dans les projets ACR Grimm Tradalign et ParaTaxe. Cet ensemble comprend l'aligneur AAlign, basé sur des embeddings multilingues, une chaîne de traitement Python pour convertir et aligner des corpus, une interface web de gestion de corpus nommée ParaTaxe, et une interface d'exploration de corpus alignés basée sur TEI-Publisher. Une première application a permis de publier un corpus parallèle des contes des frères Grimm.

Mots clés : alignement de textes parallèles, corpus multiparallèles, plate-forme de publication XML-TEI.

1. Introduction

Over the past fifteen years, digital tools for corpus parallelization have developed significantly in the field of Digital Humanities: examples include the HyperMachiavel software (Gedzelman & Zancarini, 2011) for a corpus of translations of Machiavelli's *The Prince*, the alignment tool from Marianne Reboul's *Odysseus* project, which studies translations of *The Odyssey* into French from the 16th to the 20th century (Reboul, 2022), the CollateX software by Dekker & Middel (2011), which allows multiple alignments to be observed at the lexical level, and the Variance platform (see <https://variance.unil.ch>), dedicated to comparing different versions of the same edition.

Despite the maturity of alignment techniques and the large number of projects carried out in this field, it is clear that many parallel corpora developed in an academic context are not (or no longer) available online due to a lack of maintenance or funding to develop *ad hoc* sites (such as Hyperprince). For example, the CARMEL parallel corpus (El Bèze et al., 2006), a vast corpus of travel stories in four languages, was only accessible for two years, following the withdrawal of the industrial partner involved at the time. On the Ortolang Platform (<https://www.ortolang.fr/>), there are several parallel corpora available for download (GIEC corpus, UN Security Council Resolutions 1946-2015 corpus), but none of them can be consulted directly in a simple manner.

To fill this gap, we present a suite of open source tools developed as part of the *Grimm Tradalign* (funded by the ACR group) and *ParaTaxe* (funded by IDEX UGA) projects. There are four tools: the *Allign* aligner (Kraif, 2025), a robust, state-of-the-art aligner based on multilingual embeddings; a Python processing chain that converts parallel corpora in TXT or DOCX format into XML-TEI, then aligns them (in TSV, TXT, and XML); a web interface called *ParaTaxe Dashboard* for managing collections of parallel corpora and launching this pipeline; and finally, a web interface for exploring aligned multilingual corpora, developed using TEI-Publisher, an instance of which has already been used to deploy a parallel corpus of Grimm's fairy tales (<https://grimm.unil.ch/>).

2. Main concepts and existing tools

By multi-parallel texts, we mean a generalization of the concept of parallel texts, which originally applies to pairs combining source texts and their translation into a target language. The alignment of such a parallel corpus is generally performed by an alignment software or an aligner, which, as formalized by Isabelle (1992), relies on two steps : *segmentation* (at varying degrees of granularity, paragraphs, sentences, phrases, or words) and *matching* equivalent segments. We also use the definition of parallelism given by Langé & Gaussier (1995), who based this concept on two criteria related to the structure of equivalence relations: 1/ quasi-bijectivity: segments of the source text generally have a corresponding segment in the target text; 2/ quasi-monotonicity: the order of the target segments generally respects that of the source sentences.

Based on these formal properties, we can generalize the concept of a multi-parallel corpus to a triplet $\langle T, f, A \rangle$, where T is a collection of texts (which may be in one, two, or several

languages), f is a segmentation function (which may operate at the level of paragraphs, sentences, verses, etc.), and A is a set of pairings between equivalent segments, with the concept of equivalence extended beyond that of translational equivalence: it can concern successive rewritings of the same text in the case of textual genetics or variations between several editions, different translations from the same source text, or even a mixture of these phenomena (as is the case in the Grimm corpus described below). The notion of a multi-parallel corpus *ultimately* becomes a formal property based on the constraints of quasi-bijection and quasi-monotonicity linked to alignment A .

For the publication of multi-parallel corpora defined in this way, there are two types of tools:

- Automatic aligners, which aim to produce alignments generally at the sentence level. Following the generation of surface aligners such as Gale & Church (1991), BMA (Moore, 2002) and Hunalign (Varga et al. 2007), semantic aligners based on sentence embeddings (vector representations derived from neural models) have shown greater accuracy and robustness with regard to phenomena such as parallelism breaks or significant distance between languages with different alphabets. Among these, Thomson & Koehn (2020), with the Vecalign system, and then Liu and Zhu (2022), with BertAlign, have produced a new state of the art. At the lexical level, we can mention statistical aligners such as FastAlign (Dyer et al. 2013), and more recently, based on transformer architectures, SimAlign (Jalili Sabet et al. 2020) or awesome-align (Dou & Neubig 2021). It should be noted that, with a few exceptions, such as the multi-aligner JAM (Kraif, 2015), these tools most often operate on pairs of texts. The WebAlignToolkit¹ interface provides access to various online aligners and allows users to obtain a sentence-level multi-alignment for two or more texts. In the field of aligning versions of the same text from a genetic perspective, the CollateX software (Dekker & Middel 2011) allows users to view correspondence graphs at the character string level. Finally, with Mkalign, Zimina & Fleury (2015) present a comprehensive environment for the alignment, visualization, and statistical exploration of aligned texts – the tool can be used locally on a desktop computer, but is not designed for online use.

- The second family of tools concerns interfaces for publishing parallel corpora. It is clear that a large number of parallel editions are based on *ad hoc* sites rather than generic tools: we can cite the examples of *HyperPrince*², *Translatoscope*³, and *Odysseus Polymetis*⁴. There are platforms dedicated to the publication of different corpora, such as the very rich Opus (Tiedmann 2012), the Variance platform (Mahrer & Zufferey, 2019), and the collaborative Ugarit (Yousef et al. 2022), which allows users to edit word-by-word alignments and share them. Also noteworthy is the ParCoLab project (Stosic et al. 2024), which is based on the open source NoSketch Engine, but aims to publish a specific parallel corpus.

¹ Available at : <http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit/>

² Translations of Machiavelli's *The Prince*: <http://hyperprince.ens-lyon.fr>

³ Translations of Aristophanes' *Ploutos*: <https://elan.gricad-pages.univ-grenoble-alpes.fr/translatoscope>

⁴ Translations of Homer's *Odyssey*: <https://odysseuspolymetis.github.io/paralogos/>

In this article, we propose ParaTAXE, a suite of generic, open-source tools integrated into a single platform that allows users to align and publish their own editions of a multi-parallel corpus without investing in new developments.

3. The ParaTAXE suite

The ParaTAXE (Parallel Text Alignment eXploration and Edition) suite aims to integrate the entire workflow, from importing source files to viewing texts in parallel, with an unlimited number of columns (except for reading comfort, depending on screen size).

3.1 The Dashboard

The dashboard is the corpus management interface⁵. Corpora are organized into *collections* (e.g., *Homère*), each of which contains *documents* (e.g., *Iliade1*), each corresponding to a list of *files* (e.g., *ILIADE1_1530_Samxon_fr.xml*). Files can be in different formats: DOCX, TXT, XML-TEI.

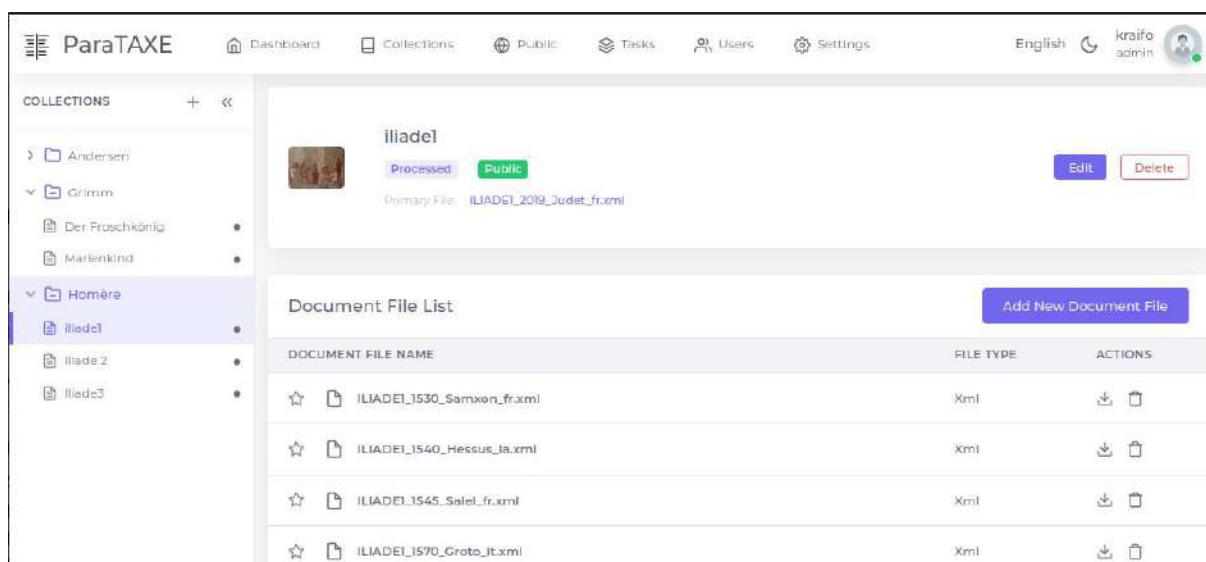


Figure 1: Management of collections / documents / document files

3.2 Launching the pipeline

For each collection, a metadata file is defined to specify information such as ‘title’, ‘author’, ‘year’, ‘translator’, etc. for each file.

- File conversion

The pipeline involves conversion to XML-TEI if necessary, with metadata injection in the header and segmentation at the paragraph and sentence level. If the imported files are already in XML-TEI, it is necessary to indicate which segmentation tag should be taken into account for alignment (e.g., <l> for the verses of the Iliad).

⁵ An instance of the tool is currently hosted here: <http://phraseotext.univ-grenoble-alpes.fr:5000>

- Alignment

Alignment is performed using the Align aligner (Kraif, 2025), a state-of-the-art aligner that performs a preliminary step of detecting alignable areas (e.g., by eliminating untranslated prefaces or chapters, or by identifying a different order in the publication of stories or short stories in a collection). The aligner itself is available on a git repository⁶ for standalone use (command line interface). The aligned files are saved in different formats (TSV, TXT, XML-TEI) and can be downloaded by the user.

To simplify alignment, one file must be designated as the pivot: all other files are then aligned with the pivot (which implies, for n files, $n-1$ aligned pairs).

3.3 Visualization

Once the processing is complete⁷, the documents can be viewed in a multi-parallel visualization. Clicking on a segment synchronizes the display with all other displayed versions (synchronization is achieved by alignment with the pivot, even if it does not appear in the columns).



Figure 2: Parallel display of aligned documents

3.4 Publication of corpora

Collections can be public or private: a public collection is accessible for consultation by other registered users of the platform. In addition, once the processing is complete, it is possible to make a document completely public, without authentication, by clicking on "Publish." This publication mode involves displaying an illustration icon, a presentation text, and multi-parallel columns. A search field allows the user to extract parallel concordances for a given expression.

⁶ At the address: <https://gricad-gitlab.univ-grenoble-alpes.fr/kraifo/align>

⁷ The pipeline jobs are executed using Redis, a queue management system for tasks in progress.

In addition, we have also developed TradAlign, a publishing interface based on TEI Publisher⁸. Independent of the ParaTAXE platform, it also allows multi-parallel publication and metadata search options. We used it to publish a translation corpus of the Grimm brothers' fairy tales⁹ (Rimasson-Fertin et al. 2024). This corpus gather various editions of a selection of 10 Grimm tales, from the first edition of 1812 to the last of 1857, which has been taken as a pivot. French translations from the 19th century have been aligned with this pivot, as well as the previous original editions.

Note that the files obtained by the aligning pipeline of ParaTAXE are compatible with the format required for TradAlign.



Figure 3: Grimm TradAlign interface, based on TEI Publisher

4. Conclusion and future works

The ParaTAXE suite is available on a Git repository under open source licence¹⁰. There are two branches: in the main branch, publication is automatically directed toward the TradAlign tool, which is integrated; in another lighter branch, without TEI-Publisher and Exist-db module, the aligned view is integrated into the dashboard.

The platform can be installed on a development or production server using a single command that creates docker containers¹¹.

In future versions, we plan to add a module (which is almost finished) for manual editing of automatic alignments, via a drag-and-drop system from a two-column view.

We also plan to add a lexical level alignment step (which already exists in AAlign) and a module to extract translation and textometric statistics (such as sentence lengths, vocabulary

⁸ The implementation has been conducted by the Jinntec company under our supervision (<https://jinntec.de/>)

⁹ The site is hosted by the University of Lausanne and can be viewed here: <https://grimm.unil.ch>

¹⁰ Available here: <https://gricad-gitlab.univ-grenoble-alpes.fr/jaliliel/parataxe-admin>

¹¹ Docker is a technology that creates a kind of virtual environment to ensure the application's portability across different systems. The install procedure just consists in one command line (*sudo make dev* or *sudo make file*), with the *makefile* installing the dependencies and launching the various docker containers hosting the necessary services.

sizes, hierarchical indexes, occurrence and cooccurrence frequencies, correspondence frequencies, etc.)

Acknowledgments

The Grimm corpus and the Grimm Tradalign interface were funded by ACR (Project ANR-15-IDEX-02) and UNIL. The ParaTAXE project received IDEX funding from UGA (IRGA2023 –ParaTAXE – G7H-IRG23D42).

References

- Dekker R. H., Middell G. (2011). Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. *Supporting Digital Humanities 2011*. University of Copenhagen, Denmark.
- Dou Z.-Y., & Neubig G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2112–2128.
- Dyer C., Chahuneau V., Smith N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648.
- El-Bèze M., Richard C., Meyer R. (2006). CARMEL Project: Travel Stories. In *Proceedings of TALN 2006*, Louvain-la-Neuve.
- Gedzelman S., Zancarini J.-C. (2011). HyperMachiavel: a translation comparison tool. *Lingua e stile*, 2011, vol. XLVI, no. 2, pp. 247-266.
- Isabelle P. (1992). Bi-Textual Aids for Translators. In *Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research*, University of Waterloo, Waterloo, pp. 76-89,
- Langé J.-M., Gaussier É. (1995). Alignement de corpus multilingues au niveau des phrases. *T.A.L.*, vol. 36, n. 1-2, pp. 67-79.
- Jalili Sabet, M., Dufter, P., Yvon, F., & Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1627–1643).
- Kraif, O. (2015). Multialignement vs bialignement : à plusieurs, c’est mieux ! In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*, pp. 345–356.
- Kraif O. (2025). Adaptive bi-textual alignment based on multilingual embeddings. *Proceedings of CORIA-TALN-RJCRI-RECITAL 2025*, Marseille, July 1-5, 2025. <https://aclanthology.org/2025.jeptalnrecital-taln.4.pdf>
- Kraif O., Roux P. (2022). Comparison of an original text and its back-translations: what do textometric measures tell us?, in Meng Ji Christine and Michael Oakes (Eds.), *New methodologies in corpus-based translation studies: The empirical revolution in translation studies, Meta*, v. 67, no. 1, Les Presses de l’Université de Montréal: Montreal.
- Liu L. & Zhu M. (2022). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, Volume 38, Issue 2, June 2023, 621–634.

- Mahrer R. & Zufferey J. (2019). *Variance: A digital platform for the study of textual rewriting across editions*. In W. Scholger & M. Büchler (Eds.), *Digital Scholarly Editions as Interfaces*. Norderstedt: Books on Demand.
- Moore R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144, Tiburon, USA, October. Springer.
- Stosic, D., Marjanović, S., Bernhard, D., Bach, X., Bras, M., Kevers, L., Retali-Medori, S., Vergez-Couret, M., & Werner, C. (2024). The ParCoLab parallel corpus and its extension to four regional languages of France. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16014–16023).
- Reboul M. (2022). Semi-automatic comparison of French translations of Homer's Odyssey (1547-1955). In Didier Alexandre and Glenn Roe (eds.), *Cultures et pratiques savantes du numérique*, Classiques Garnier (10.48611/isbn.978-2-406-12961-5)
- Rimasson-Fertin N., François C., Kraif O., Roux P., Jalilian E. (2024). *Grimm Tradalign: plateforme de comparaison de traductions de contes des frères Grimm*. UGA-UNIL.
- Thomson B. & Koehn P. (2020). Vecalign: Improved Sentence Alignment in Linear Time and Space. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1342–1348, Hong Kong, China, November 3–7, 2019.
- Tiedemann J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.
- Varga D., Halácsy P., Kornai A., Nagy V., Németh L. & Trón V. (2007). Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.
- Yousef T., Palladino C., Shamsian F., & Foradi M. (2022). Translation Alignment with Ugarit. *Information* (2022) 13(2): 65.
- Zimina, M. & Fleury, S. (2015). Perspectives de l'architecture Trame/Cadre pour les alignements multilingues. *Nouvelles perspectives en sciences sociales*, 11(1), 325–353.

Analyse de textualités multiples avec TXM : du texte brut aux éditions TEI, transcriptions synchronisées, traductions alignées et micro-textes

Serge Heiden

ENS de Lyon – slh@ens-lyon.fr

Abstract

The fundamental choice of using a standard XML-TEI pivot representation of texts in the TXM platform has enabled the processing of a wide range of textual modalities. After a brief overview of the textual formats used in textometry, we characterize our use of the XML-TEI standard and describe how it is employed to encode the various textualities handled by TXM. We then present the main configuration options available for importing a corpus into TXM in order to obtain an analyzable corpus. We conclude by highlighting the FAIR benefits of the interoperability of the resulting corpora.

Keywords : digital text, multimedia text, synchronized text, aligned texts, annotation, textometry, XML, TEI, TXM

Résumé

Le choix fondamental d'utiliser une représentation pivot XML-TEI standard des textes dans la plateforme TXM a permis de développer une capacité de traitement d'une grande diversité de modalités textuelles. Après un petit historique des formats textuels utilisés en textométrie, nous caractérisons notre utilisation du format standard XML-TEI et décrivons comment nous l'utilisons pour encoder les différentes textualités gérées par TXM. Nous présentons ensuite les principaux paramétrages disponibles pour importer un corpus dans TXM afin d'obtenir un corpus analysable. Nous concluons sur l'intérêt FAIR de l'interopérabilité des corpus obtenus.

Mots clés : texte numérique, texte multimédia, texte synchronisé, textes alignés, annotation, textométrie, XML, TEI, TXM

1. Introduction

La textométrie s'applique aujourd'hui à une grande variété de textualités : textes numériques d'origines diverses (OCR, HTR, épigraphie, traitement de texte...), messages asynchrones (emails), réseaux interactifs (chat, sms, twitter), sous-titres ou transcriptions synchronisés d'enregistrements audio ou vidéo, réponses à des questions ouvertes d'enquêtes, traductions alignées (e.g. au niveau du paragraphe).

Historiquement, le format texte brut (une séquence de caractères) a toujours servi à représenter l'ensemble de ces textualités pour les exploitations informatiques.

Dans les années 80, des systèmes de balisage simples (non hiérarchiques et avec très peu de balises) ont été développés pour coder des délimitations textuelles (texte, prise de parole...) et quelques métadonnées (émetteur, date...) au sein du texte brut. Par exemple l'encodage MA-CHINAL du laboratoire de Saint-Cloud (Lafon et al., 1985) (analogue à l'encodage COCOA anglo-saxon (Wikipedia contributors, 2026)) utilisé par les logiciels Lexico et Saint-Chef, ainsi

que les balises Hyperbase ou encore Alceste, encore utilisées aujourd’hui par le logiciel Iramuteq (<https://pratinaud.gitpages.huma-num.fr/iramuteq-website/>).

Dans les années 90, le système d’encodage SGML a servi de premier support aux recommandations de balisage de la Text Encoding Initiative (TEI) (TEI Consortium, 2023). Ces recommandations sont alors devenues le standard international reconnu pour l’encodage des textes numériques.

Enfin dans les années 2000, l’adoption par la TEI du système d’encodage simplifié XML (Bray et al., 2008) - accompagné du standard UNICODE d’encodage universel des caractères (Unicode Consortium, 1996) - a permis une large diffusion de ce standard dans les projets de recherche en humanités numériques.

C’est pour cela que, dans les années 2010, la plateforme TXM (Heiden, 2010) s’est donné pour objectif d’être capable d’exploiter l’encodage XML-TEI des textes sources pour construire les corpus à analyser.

C’est un choix assez exigeant car dans l’encodage XML-TEI, le seul format reste le XML. La partie TEI (qui dénomme et structure les balises XML) reste très ouverte ontologiquement pour faciliter son adoption par les chercheurs. Mais cette liberté d’encodage empêche l’interprétation automatique de n’importe quel fichier XML-TEI par les logiciels. Parmi les difficultés rencontrées on trouve la possibilité d’encoder un même phénomène textuel de différentes façons ou encore la possibilité d’encoder une alternative entre deux encodages différents.

La réponse apportée dans TXM à cette problématique générale consiste à fournir au sein de l’environnement d’importation des fichiers sources des outils de paramétrage et de migration vers une représentation XML-TEI unifiée propre à TXM qu’il peut interpréter entièrement automatiquement.

Ce faisant, différentes balises de la TEI (la TEI propose environ 600 balises relevant de nombreux domaines textuels) nous ont permis d’intégrer au sein d’une même architecture numérique les différentes modalités textuelles que nous allons maintenant présenter.

1.1. Quatres modalités textuelles à exploiter

À ce jour quatre grands types de données textuelles sont traités par TXM en étant standardisés en XML-TEI lors de l’import dans la plateforme :

1. des écrits allant du texte brut aux éditions critiques richement encodées en XML-TEI, pouvant intégrer des données multimédia comme des images et pouvant être affichées côte à côte avec les images de fac-similé des sources (ouvrages ou manuscrits numérisés en image) (Heiden and Lavrentiev, 2012)
2. des transcriptions d’enregistrements audio ou vidéo encodant les tours de parole de chaque locuteur et synchronisées au tour de parole ou plus finement encore pour pouvoir jouer dans une fenêtre au sein même de TXM le média source au moment où des mots sont prononcés (Pincemin et al., 2020)
3. des textes en relation de traduction ou de version alignés au niveau d’une structure textuelle comme la phrase ou le paragraphe. Ces corpus multilingues permettent des exploitations textométriques basées sur l’alignement comme l’analyse de résonances (Zimina, 2005)
4. des micro-textes de la taille d’une seule à plusieurs phrases ayant beaucoup de méta-

données comme les échanges dans les réseaux sociaux ou les réponses à des questions ouvertes d'enquêtes. La géométrie spécifique de ces corpus est en général transformée de tableaux vers des textes structurés

1.2. Textes écrits

Dans le domaine des sources écrites, l'encodage XML-TEI permet bien sûr de représenter du simple texte brut (appelé aussi "texte au kilomètre"), ce qui est la forme textuelle la plus traitée par les logiciels de TAL et d'analyse textuelle.

La TEI permet ensuite d'encoder divers niveaux structurels à la demande, c'est à dire sans qu'il y ait d'obligation à tout encoder, en fonction des besoins d'analyse et de la disponibilité des informations.

La TEI permet également d'encoder des notes critiques ainsi que des informations pouvant piloter le rendu des textes lors du rendu des textes dans un format de lecture donné. Par exemple les limites de paragraphes peuvent servir à formater les paragraphes ou les sauts de page à gérer une pagination à la lecture. Les sauts de page permettent également de lier chaque page de texte à l'image de la numérisation de la source d'origine comme un ouvrage ou un manuscrit, qu'il soit en ligne ou sur la machine où s'exécute TXM. Ceci permet à TXM de proposer des éditions appelées 'fac-similé' affichant côté à côté l'image source d'une page et la transcription de cette page. C'est un outil très utile pour vérifier la qualité des exploitations réalisées à partir des transcriptions, notamment en cas d'erreurs d'OCÉRisation ou d'HTÉRisation.

1.3. Transcriptions d'enregistrements

Dans le domaine des transcriptions d'enregistrement (audio ou vidéo), le format des textes source dépend fondamentalement des logiciels permettant de produire les transcriptions (soit par saisie assistée - avec les logiciels CLAN (<https://dali.talkbank.org/clan/>) ou Elan (<https://archive.mpi.nl/tla/elan>) par exemple - soit automatiquement - à l'aide des logiciels Whisper (<https://openai.com/fr-FR/index/whisper/>) ou VoxSigma (<https://www.vocapia.com/speech-to-text-technology2.html>) par exemple).

Le logiciel open-source Transcriber (<https://trans.sourceforge.net/en/presentation.php>, qui n'est plus développé activement) est à l'origine d'un format basé sur XML qui a eu beaucoup de succès pour les transcriptions simples (utilisées lors d'enquêtes en sociologie par exemple) délimitant et attribuant les prises de parole de chaque locuteur et les synchronisant avec les enregistrements source aux tours de parole près ou plus finement encore.

Pour la gestion de ce type de textes dans TXM nous avons créé une représentation XML TEI qui encode l'intégralité des informations du format Transcriber. Cela permet de produire un rendu des textes formatant les prises de parole en paragraphes et permettant de naviguer entre les transcriptions et les enregistrements pour les jouer au moment où certains mots sont prononcés dans une fenêtre au sein de TXM.

1.4. Textes alignés

Les corpus de textes composés de traductions en plusieurs langues ou de plusieurs versions d'un même texte peuvent aligner des textes entre plusieurs langues ou entre différentes versions au niveau de certaines structures textuelles, comme le niveau des paragraphes ou des phrases.

Ces corpus peuvent être exploités en recherchant simultanément des mots dans deux langues ou deux versions à la fois tout en tenant compte de l'alignement des structures dans lesquelles se

trouvent les mots.

Comme pour les corpus oraux, le format source de ces corpus dépend des outils qui ont permis de les produire, par calcul d'alignements automatiques et vérification assistée par une interface d'édition.

Le logiciel open-source OmegaT (<https://omegat.org/fr/>) est très populaire pour l'édition de mémoires de traduction. Les mémoires de traduction qu'il produit sont encodés dans un format XML avec des alignements simples appelé TMX.

Pour la gestion de ce type de corpus dans TXM nous avons créé une représentation XML TEI qui encode l'intégralité des informations d'alignement du format TMX.

1.5. Micro-textes

Les textes de réponses à des questions ouvertes d'enquêtes ou encore d'échanges sur les réseaux sociaux (Twitter, Instagram...) sont caractérisés par une petite taille et un grand nombre de métadonnées. Le format d'export privilégié par les plateformes qui les produisent est celui de tableaux (Calc, Excel ou CSV).

Pour représenter ces corpus de façon efficace dans TXM, les tableaux sont converti en fichiers XML où le contenu textuel provient de certaines colonnes et la structuration et les métadonnées proviennent d'autres colonnes. Ce XML est encodé en TEI pour bénéficier des services de formatage des éditions à l'import, comme la production d'une page par répondant contenant un tableau de métadonnées du répondant affiché en début de page suivi des différentes sections correspondant à chaque réponse.

2. Importation de fichiers source dans TXM

Les sources d'un corpus sont lues par un module d'import qui construit un corpus TXM. Un corpus TXM est composé de sa représentation unifiée XML-TEI et des différents objets utilisés par les outils d'analyse : les index de mots du moteur de recherche, les éditions (éventuellement paginées) et les fichiers multimédia.

À chaque modalité textuelle correspond un module d'importation dédié.

2.1. Import de textes écrits

Les modules d'import généralistes de textes écrits sont, du plus élémentaire au plus complet :

- le module TXT+CSV¹ pour les textes bruts
- le module XML/w+CSV pour les textes XML, pré-encodant optionnellement des mots à l'aide d'une balise <w>
- le module XML TEI Zero/w+CSV pour les textes XML encodés selon un jeu minimal d'une dizaine de balises TEI

Ils sont complétés par des modules d'import dédiés à des formats spécifiques : module ODT+DOCX+CSV, module Iramuteq pour le format Alceste, module Hyperbase.

¹ le '+CSV' indique que chaque texte peut recevoir des métadonnées optionnelles encodées dans un fichier tableur, nommé 'metadata'.

Tous ces modules passent par une représentation XML-TEI unifiée avant de produire les objets internes d'un corpus TXM.

Le module XML-TEI Zero/w+CSV offre des services d'adaptation du XML-TEI qui seront présentés dans une section dédiée.

2.2. Import de transcriptions d'enregistrements

Le module d'import XML Transcriber+CSV lit le format XML Transcriber pour produire une représentation XML-TEI encodant les éléments de la transcription, dont la synchronisation temporelle des tours de parole, des énoncés ou des mots.

Il organise également l'accès aux fichiers média (audio ou vidéo) pour le lecteur média des corpus, sur la machine en local ou bien à distance (par Internet, avec possibilité d'identification automatique pour les ressources sous droits d'accès).

2.3. Import de corpus parallèles

Le module d'import XML TMX+CSV lit le format XML TMX pour produire une représentation XML-TEI encodant les structures de base des textes et les alignements entre textes en autant de fichiers que de textes alignés pour qu'ils puissent disposer de leur propre édition.

2.4. Import de corpus de micro-textes

Le module d'import Tableau lit un format de tableur (Calc, Excel ou CSV) pour produire une représentation XML-TEI encodant les structures de base des textes et leurs métadonnées en fonction de paramètres désignant les colonnes de contenu textuel et les colonnes d'encodage de métadonnées.

3. Paramétrage du module d'import XML-TEI Zero/w+CSV

Nous limitons notre présentation au paramétrage de ce module car il en offre un nombre représentatif, sachant que de nombreux autres modules partagent les mêmes paramétrages.

3.1. Paramétrage des métadonnées externes

Les métadonnées externes correspondent aux propriétés des textes (titre, auteur, date, genre, etc.). Elles servent à construire les différentes configurations de corpus (sous-corpus et partitions), ainsi qu'à participer à la création des références de concordances qui permettront de situer les occurrences.

Dans la représentation XML-TEI unifiée, chaque fichier XML correspond à un texte. La stratégie d'encodage des métadonnées de textes dans TXM est de s'appuyer sur des tableaux en format tableur (Calc, Excel ou CSV). Avec une ligne par texte et autant de colonnes que de métadonnées à encoder. C'est un format particulièrement pratique pour maintenir les métadonnées d'un nombre important de textes.

Pour les sources XML-TEI, un utilitaire permet d'extraire les métadonnées des textes depuis leur <teiHeader> pour les stocker dans un tableau de métadonnées avant l'import dans TXM. Comme les métadonnées encodées dans les entêtes TEI <teiHeader> sont souvent d'ordre documentaire et moins d'ordre analytique, les données extraites sont souvent retravaillées pour obtenir des données analytiques, c'est à dire permettant de réaliser les configurations de corpus nécessaires aux analyses. Par exemple des dates sont converties en décennies ou en dates simplifiées, l'orthographe des noms d'auteurs est unifiée sur l'ensemble du corpus, etc. Si les

tableaux de métadonnées sont beaucoup travaillées, un utilitaire inverse permet de re-projeter au sein des entêtes TEI <teiHeader> les valeurs de métadonnées encodées dans les tableaux pour archivage.

3.2. Paramétrage des métadonnées internes

Les métadonnées internes correspondent aux propriétés de structures internes aux textes (identifiant de locuteur, sexe de locuteur, âge de locuteur, etc.). Elles servent à construire les différentes configurations de corpus (sous-corpus et partitions) basées sur des structures internes, ainsi qu'à participer à la création des références de concordances qui permettront de situer les occurrences. Elles servent par exemple à encoder les propriétés de locuteurs de corpus de transcriptions d'enregistrements. Dans ce cas, les métadonnées externes encodent les propriétés des enregistrements et les métadonnées internes encodent les propriétés des tours de paroles ou d'énoncés voire de mots.

3.3. Paramétrage des plans textuels

Les plans textuels permettent de désigner dans le XML-TEI unifié ce qui relève ou non de l'indexation des mots qui participeront aux calculs textométriques.

- le plan 'hors texte' désigne les structures qu'il ne faut ni indexer ni éditer. Il désigne par défaut la structure <teiHeader> qui contient les métadonnées de textes en TEI. Ce plan est très utile pour exploiter les textes encodés en XML-TEI car on y trouve souvent de nombreux encodages documentaires qui ne doivent pas participer aux analyses textométriques. Fondamentalement, il permet surtout de remplacer toute logique de 'nettoyage préalable des textes' avant exploitation (avec potentielle perte des données textuelles complètes) par une logique de sélection de la surface textuelle à exploiter au moment de la construction d'un corpus. Cela permet de capitaliser tout l'historique d'encodage des textes tout en étant capable de ne sélectionner que ce qui doit faire partie d'une étude textométrique en fonction des objectifs de recherche.
- le plan 'hors texte à éditer' désigne les structures dont les mots seront affichés dans les éditions mais pas indexés pour participer aux calculs textométriques. Il sert aux situations où l'on a besoin de lire certains mots du texte sans qu'on veuille les faire participer aux calculs textométriques. Par exemple, dans une transcription d'enregistrement, on peut indexer les mots des tours de parole des répondants pour que l'étude porte sur leurs mots et afficher les tours de paroles de enquêteurs dans l'édition, sans les indexer, pour que l'on puisse suivre les échanges à la lecture de l'édition.

3.4. Pré-encodages optionnels de mots et leurs annotations

Les textes sources peuvent pré-encoder les mots ou non selon trois modes :

- aucun mot n'est pré-encodé, comme pour du texte brut. Dans ce cas TXM procède à la segmentation lexicale en fonction des caractères se trouvant dans le texte.
- certains mots sont pré-codés. Dans ce cas la segmentation lexicale de TXM prendra les mots pré-encodés tels qu'ils sont définis, en plus des unités issues de la segmentation lexicale du texte brut. Si les mots encodent des attributs ils sont automatiquement convertis en propriétés de mots dans l'indexation.

- tous les mots sont pré-encodés. Dans ce cas TXM ne procède à aucune segmentation lexicale. Les mots sont indexés tels quels avec leurs propriétés éventuelles.

3.5. Paramètres de lemmatisation et autres outils de TAL

TXM permet d'appliquer des outils de lemmatisation automatique sur les textes importés. La lemmatisation s'applique aux mots déterminés par la segmentation automatique ou par une pré-encodage.

Son résultat est encodé dans les éléments <w> de la représentation XML-TEI unifiée.

Ce paramétrage permet notamment de choisir la langue qu'il faut utiliser pour lemmatiser.

3.6. Ressources multimédia pour afficher des images

Le module XML-TEI Zero interprète les images d'illustration encodées au fil d'un texte, qui sont intégrées aux éditions de textes.

Il interprète également les images de fac-similé de pages, ces images sont désignées par l'attribut @facs des éléments <pb/>.

3.7. Paramètres de construction des éditions de textes

La stratégie de construction des éditions de ce module est d'associer un grand nombre des balises HTML qui encodent l'édition à des éléments XML-TEI se trouvant dans la représentation unifiée source. Par exemple, par défaut une mise en évidence XML-TEI <hi> est transférée en balises HTML , une mise en évidence XML-TEI <emph> est transférée en balises HTML <i>, un saut de ligne forcé XML-TEI <lb/> est transféré en balise HTML
, etc.

Les paramètres permettent de préciser par exemple la stratégie de pagination : basée sur le repérage de balises <pb/>, en utilisant éventuellement une numérotation encodée dans un attribut @n, ou bien sur un nombre maximum de mots par page.

4. Étapes d'intégration d'un corpus dans TXM

Le module XML-TEI Zero+CSV donne accès aux différentes étapes de l'intégration d'un corpus dans TXM en offrant la possibilité de manipuler le XML-TEI produit à chaque étape.

Pour cela on crée un dossier nommé du nom de l'étape dans le dossier source et on y place des feuilles de transformation XSLT qui seront appliquées dans l'ordre lexicographique de leur nom.

4.1. Étape 1-split-merge

Cette étape est activée avant tout traitement par TXM.

Elle permet de fusionner ou d'écarter les fichiers source selon les besoins d'encodage du corpus, sachant que dans l'architecture de TXM, un fichier correspondra à un texte du corpus construit.

Cette étape est utile pour adapter des corpus XML-TEI où les fichiers ne correspondent pas forcément à un texte à traiter. Par exemple un fichier XML-TEI qui contient plusieurs textes ou encore un texte éclaté en plusieurs fichiers XML-TEI.

4.2. Étape 2-front

Cette étape permet d'adapter des encodages ou pour de migrer des encodages dans différents niveaux de structures, une fois l'étape précédente réalisée - c'est à dire où les fichiers de textes sont construits.

Cette étape est utile pour construire finement les références de concordances à partir de toutes les informations disponibles dans l'encodage XML-TEI. Par exemple les numéros de page encodés dans des attributs @n des balises <pb/>.

4.3. Étape 3-posttok

Cette étape applique les transformations sur les textes segmentés en unités lexicales, c'est à dire une fois que la segmentation lexicale a été réalisée.

Cette étape est utile pour redresser certaines segmentations automatiques incorrectes.

4.4. Étape 4-edition

Cette étape applique des transformations pour obtenir des fichiers HTML correspondant à des éditions des textes.

Elle fonctionne en deux passes :

- une passe pour produire autant de fichiers HTML que de pages indiquées par des balises <pb/>
- une passe pour produire le contenu des fichiers HTML

4.5. bibliothèque XSL de TXM

TXM est livré avec une bibliothèque d'une cinquantaine de feuilles de transformation XSLT pouvant servir dans les différentes étapes de traitements de l'import.

Par exemple les feuilles 1-default-html.xsl et 2-default-pager.xsl sont utilisables dans l'étape 4-edition pour produire une édition HTML générique, en plus de l'édition par défaut produite par TXM.

5. Conclusion : Interopérabilité et Archivage

Nous avons montré comment une représentation XML-TEI interne propre à TXM a permis de construire une architecture de corpus permettant de traiter un large spectre de textualités.

Pour conclure nous voudrions insister sur un dernier intérêt - externe à TXM - de l'utilisation d'une telle représentation XML-TEI unifiée.

En effet, comme TXM transforme dans une représentation XML-TEI interne toutes les sources de corpus à traiter, quel que soit leur format d'origine (texte brut .txt, MS Word .docx, MS Excel .xlsx, etc.), il se comporte comme un outil générique de standardisation de sources de corpus pour l'analyse de corpus textuels.

Dans une optique de pratiques FAIR (Findable, Accessible, Interoperable, Reusable), les représentations standard XML-TEI des corpus TXM sont donc également une garantie d'interopérabilité entre outils d'analyse et d'archivage pérenne des corpus.

Le logiciel Iramuteq, par exemple, est capable de lire la version XML-TEI des textes d'un

corpus TXM. Ceci permet aux utilisateurs d'Iramuteq d'appliquer ses outils, comme la classification Reinert, sur n'importe quel corpus TXM.

De façon plus générale, n'importe quel logiciel interprétant le standard XML-TEI peut donc exploiter les données textuelles exportées depuis TXM.

En termes d'archivage, le format XML est ouvert et lisible par l'homme et par la machine. La validation de la conformité du XML au schéma TEI est une garantie minimale de structuration cohérente. La sémantique TEI sépare la logique fondamentale de la structure des textes de leur représentation physique. La terminologie des balises TEI auto-documente cette structure et enfin, l'entête de texte <teiHeader> encode des métadonnées d'archivage fondamentales : auteur, date, source, conditions de droits, etc.

Références

- Bray T., Paoli J., Sperberg-McQueen C. M., Maler E., and Yergeau F. (2008). Extensible markup language (XML) 1.0 (Fifth Edition). Technical report, W3C.
- Heiden S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Ryo Otoguro K. I., editor, *24th Pacific Asia Conference on Language, Information and Computation*, pp. 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University.
- Heiden S. and Lavrentiev A. (2012). The TXM Portal Software giving access to Old French Manuscripts Online. In *Proceedings of the 1st Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects, Seventh International Conference on Language Resources and Evaluation*, pp. 29–35, Istanbul, Turquie.
- Lafon P., Lefevre J., Salem A., and Tournier M. (1985). *LE MACHINAL : Principes D'enregistrement Informatique Des Textes*. Collection "SAINT-CLOUD". Klincksieck, Paris.
- Pincemin B., Heiden S., and Decorde M. (2020). Textometry on Audiovisual Corpora. University of Toulouse.
- TEI Consortium (2023). *TEI P5 : Guidelines for Electronic Text Encoding and Interchange*. Version 5.1.0 edition.
- Unicode Consortium, editor (1996). *The Unicode standard*. Addison-Wesley Developers Press, Reading, Mass, version 2.0 edition.
- Wikipedia contributors (2026). COCOA (digital humanities). [https://en.wikipedia.org/w/index.php?title=COCOA_\(digital_humanities\)](https://en.wikipedia.org/w/index.php?title=COCOA_(digital_humanities)). Consulté le 18 mars 2026.
- Zimina M. (2005). Exploration textométrique de corpus de traduction. *Meta : journal des traducteurs / Meta : Translators' Journal*, 50(4).

[THIRD PAGE]

C07

Statistical methods and pattern extraction

10-lug
09:30 - 11:00
Aula Multimediale C

The UMCC: Usage and Measurement of Connectors in Context. A methodology and interface for measuring the use of connectors in different genres of corpora

Corinne Rossari¹, Claudia Ricci¹, Linda Sanvido¹, Jonathan Schaber², Gerold Schneider²

¹University of Neuchâtel – corinne.rossari@unine.ch / claudia.ricci@unine.ch / linda.sanvido@unine.ch

²University of Zurich – jonathan.schaber@uzh.ch / gerold.schneider@uzh.ch

Abstract

This article analyses linguistic forms in relation to their statistically significant context, aiming to identify similarities or differences according to their lexical distribution across corpora representing different discourse genres. To this end, we have developed a visualization tool “translating” the proximity/distance between connectors and other lexical items – based on statistical measures known from collocation research – into graphical representations consisting of rays whose length indicates their mutual degree of attraction. The object of our study is discursive connectors, for two reasons. First, the collocations of this category of forms have not been studied: previous research mainly focused on verb + noun or noun + adjective pairings. Second, the typologies of connectors and the description of their contexts of use in relation to their functions pose a problem of circularity: the function is deduced from the context of occurrence, and the context is what serves as the basis for assigning a function. Our methodology, combined with the visualization tool, will highlight the properties of connectors based on their pairings with other forms according to the corpus taken into account. The more these pairings recur in the different corpora, the more they indicate conventional properties of connectors.

Keywords: discourse connectives, corpus-driven approach, specificity index, textometry, variation

1. Introduction

The description of discourse connectors is based on an implicit methodological principle: a connector is defined by the types of coherence relations in which it is considered acceptable, and coherence relations are identified by the type of link that can unite the constituents surrounding the connector. Thus, in the approach of Pander Maat and Degand (2001), a connector belongs to the category of subjective causality if it is compatible with a relation interpreted as subjective, and it is excluded if it is deemed unacceptable in this type of configuration. However, a risk of circularity arises because the connector and the relation it conveys are not treated independently of their contexts of occurrence: relational categories are inferred from the contexts in which connectors appear, and the connectors are then classified according to those same categories.

This difficulty is part of a more general problem in the description of discourse markers. It is linked to the lack of a clear distinction between the content surrounding the marker, whose semantics are difficult to conceptualize, and the functions attributed to it. In other words, when a marker x appears in contexts a , b , c , or d , we tend to conclude that it has the functions a , b , c ,

or *d*. Contextual properties based on a qualitative study – even if accompanied by a large sample of uses – are then reinterpreted as functional properties of the marker itself. Abouda's (2022) analysis of the discursive marker *du coup* illustrates this descriptive problem particularly well. According to this author, the diversity of environments in which *du coup* appears (causal, consecutive, inferential, or organizational contexts, etc., illustrated by a sample of uses of *du coup* in a given corpus) leads him to attribute a plurality of functions to it. The issue, then, is to determine what is intrinsic to the functioning of the connector itself and what sets it apart from other discourse markers, which likewise occur across heterogeneous contexts.

These findings invite us to reconsider the methodological foundations of connector analysis. Rather than starting from the coherence relations used to classify them, we start from the statistically significant environments of their use in different corpora, covering different diamesic and diaphasic dimensions. The aim is to deduce their properties from the linguistic and discursive configurations in which they preferentially appear, independently of any projection onto the coherence relation they are supposed to convey. From this perspective, the proximity or distance between connectors is determined by the similarity or divergence of their contextual lexical profiles, regardless of the relation linking the segments they connect. In other words, it is distributional regularities that enable us to map their behaviour.

In order to make these regularities visible and systematically compare usage profiles, we use a visualization tool which we call Usage and Measurement of Connectors in Context (UMCC). This instrument allows us to visualize the behaviour of connectors in their attested environments, thus providing an empirical methodological basis for rethinking their functional description.

2. Methodology

In line with an approach that seeks to describe the functioning of forms through the specific linguistic environments in which they appear (see for instance Firth, 1957; Sinclair, 1991; Biber, 2015), we rely on the concept of collocation (cf. Evert, 2008) to identify these environments. To establish the collocates of connectors, we have used the TXM platform, a tool for the analysis of structured and annotated textual corpora (Heiden et al., 2010).

2.1. Identification of collocates

The identification of collocates of a given connector is based on the calculation of a specificity index (Lafon, 1980; 1981) implemented in the TXM tool. Lafon's index is based on a probability calculation and relies on four elements: the frequency of the co-occurring item in a given text span, the total frequency of the co-occurring item in the corpus, the size of the span, and the size of the whole corpus. It is calculated using a hypergeometric model that compares the observed co-frequency of the connector and the form with the expected co-frequency based on their overall distribution in the corpus – which is neither affected by the size of the corpus, nor by low-frequency bias (for a discussion concerning associations measures with regard to frequency, see Evert, 2008: 1226 ff.)¹. We only retain items whose specificity value exceeds a

¹ The least frequent connector (*par conséquent*) included in the analysis reaches 130 occurrences in one of the two corpora selected for this paper, the press corpus *Le Monde*, and was not taken into account in the analysis of the other corpus, *ESLO2*, as it occurs only once in it.

statistically significant threshold (index $\geq 2^2$). After the lists of collocates thus obtained have been manually checked (removing duplicates or irrelevant items such as proper nouns, acronyms, etc.), they are integrated into the UMCC tool.

2.2. Grouping into morphosemantic families

To assess not only the attraction of connectors to specific lexical items but also their tendency to co-occur with particular semantic or grammatical classes, we complement the analysis of collocates by grouping them into broader semantic or morphosemantic families. This grouping makes it possible to capture the regularity of collocational patterns at a more abstract level, thereby avoiding the masking effects of lexical dispersion. Accordingly, alongside the extraction of word lists, we carry out targeted annotations in TXM for selected categories, such as modal verbs (*pouvoir, devoir, falloir*) or deictic forms (*je, me, tu, te*).

2.3. Corpus selection

Our choice of corpora is based on a central principle: to increase the variety of uses of forms in order to capture connectors in the diversity of their actual uses. The aim is to select a range of discourse genres – such as press, encyclopaedias, computer-mediated communication, and spontaneous spoken interactions – that reflect diverse contexts of production across spoken, formal written and spontaneous written modes, to capture variation along both diamesic and diaphasic dimensions. We thus adopt a methodology that allows for the gradual expansion of both the range of corpora used and the set of languages considered.

In this study, we focus on French and select two corpora that differ markedly along the diamesic and diaphasic dimensions – written press and spontaneous spoken data – in order to distinguish properties inherent to the functioning of connectors from those dependent on corpus-specific, genre-related parameters. Table 1 presents the two corpora used.

Diamesic/diaphasic dimension	Genre	Name of the corpus	Language	Corpus information
Formal writing	National press	Le Monde	French	Tokens: 17'895'009 Year: 2010
Informal speech	Spontaneous conversation	ESLO2	French	Tokens: 1'976'506 Year: 2005-2014

Table 1: presentation of the corpora.

3. Visualization tool

Once the relevant corpora and connectors have been selected and the statistical metrics for the collocates, or groups of co-occurring items, have been computed, the resulting data can be examined to identify patterns in the contextual profiles of the connectors.

Conceptually, the data obtained by the method just described above can be thought of as a collection of lists, each list holding the computed metrics for all items *I* for a connector *c* in a corpus *d*, where *I* are the (group of) collocates, *c* is one of the pre-defined connectors (in this

² This threshold indicates that there is less than a one-percent likelihood that the collocation is due to chance.

study: *mais*, *pourtant*, *donc*, *du coup*, *par conséquent*³), and *d* is one of the selected corpora (*Le Monde*, *ESLO2*). Table 2 lists the first 10 collocates with the highest computed metric for the connectors *mais* and *donc* in the corpus *Le Monde* (the columns are *lemma*, *PoS*, *specificity index*):

<i>mais</i>			<i>donc</i>		
surtout	ADV	247.25	être	VER:pres	192.88
cela	PRO:DEM	238.69	il	PRO:PER	155.12
ce	PRO:DEM	232.38	.	SENT	142.07
seulement	ADV	217.36	falloir	VER:pres	106.74
je	PRO:PER	177.91	ne	ADV	103.43
elle	PRO:PER	175.64	pas	ADV	80.27
certainement	ADV	146.13	voilà	ADV	55.36
bien	ADV	144.01	ce	PRO:DEM	43.44
rester	VER:pres	127.89	être	VER:cond	41.70
être	VER:pres	127.00	devoir	VER:pres	35.87

Table 2: first ten most strongly attracted collocates of *mais* and *donc* in the corpus *Le Monde*.

Obviously, it is very tedious to manually inspect such lists side-by-side and try to identify possible patterns between them (items that occur in both lists are highlighted in the table above – still it is difficult to immediately set the numbers in relation to each other etc.). Therefore, we decided to visualize these tables in an interactive tool for aiding the researchers with formulating hypotheses regarding the context profiles of the connectors⁴.

Our guiding principle is to convert the lists into a network of nodes and edges, where the nodes are the connectors and items, and the edges connecting them represent the computed metric between them (i.e. between the connector and its collocates). To visualize the property of indicating strong attraction between connector and item, the metric is inverted, such that items that display high attraction are also visually close to the corresponding connector.

In the simplest case where one connector (in this case *mais*) in one corpus (*ESLO2*) is displayed, the connector would form the “root” or centre of the network and the collocations would be dispersed around the connector, connected to it by links that are lengthwise adjusted to faithfully represent the inverse of the metric. Figure 1 shows this case.

³ The selected connectors were chosen because they allow for the identification and representation of a diverse range of relations. However, the choice of connectors does not influence the methodology itself, as the same approach could be applied to other connectors.

⁴ Publicly available at: <https://connectives.linguistik.uzh.ch/>.

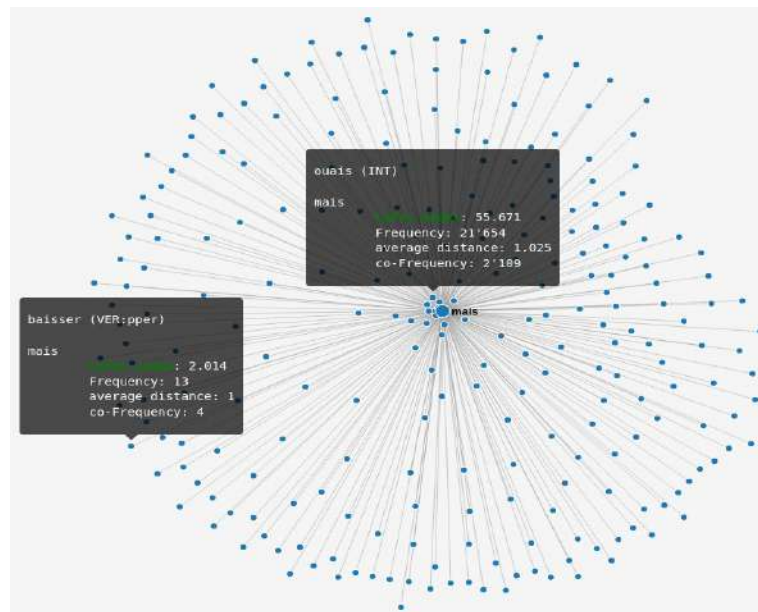


Figure 1: screenshot showing a single connector with its collocates clustered around it. The length of the links between the connector and its collocates corresponds to the inverted attraction metric. Two collocates have been clicked to expand their information pop-up.

The tool also allows for the representation of collocates shared by multiple connectors (Figure 2 displays the collocates shared by at least two connectors). It allows nodes to be clicked to open a pop-up containing more information about them (e.g. the precise metrics that were computed for it etc.), and they can be filtered to only display specific items such as PoS type *VER* etc.

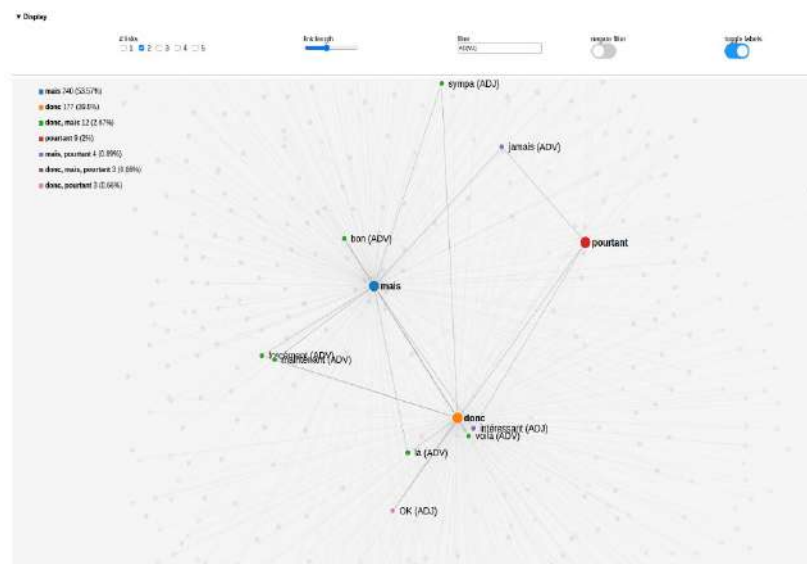


Figure 2: screenshot of a network of two connectors with their collocates clustered around them in ESLO2. The controls on top of the graph indicate that only collocates related to 2 connectors (cf. links) and matching the RegEx "AD[JV]" (adjectives and adverbs) (cf. filters) are shown with their labels.

The user can interact with the rendered data: nodes can be dragged to different places in the network, causing the simulation to re-run and settle in a new equilibrium that resembles this initial state. To find an optimal layout for the network where the nodes are spread out as uniformly as possible (to facilitate “reading” the network), as well as keep the link length as close as possible to the inverse to the pre-computed metric (i.e. a collocate node x with half the specificity strength to connector c compared to collocate node y should be twice as far away from c), the application runs a force simulation on the data. Initially, the nodes get randomly positioned around the centre of the network. Then, periodically, physical-like forces get applied to the network, i.e. nodes that are linked get attracted along the link with a force equivalent to the inverse of the metric and additionally, nodes repel each other to force the network to disperse uniformly in the two-dimensional space. This periodic force-application runs until the network reaches a state of equilibrium, i.e. a layout where the forces (more or less) balance each other out⁵.

The consequence of this approach is, however, that the resulting network layout may not always represent the original information (i.e. the computed attractions between items) perfectly. For example, if an item is very strongly attracted to two connector nodes, faithfully representing these strong attractions would require both connectors to be positioned extremely close to that item – and, consequently, close to each other. This, however, would force other nodes into unnaturally tight configurations, conflicting with the repulsive forces acting across the network. Instead, the layout settles into a configuration that balances these competing forces by minimizing the system’s overall energy. This means that while the global arrangement is optimized, individual relations (e.g., between specific pairs of nodes) may not be represented at their ideal distances. For this reason, namely the force simulation to render the network⁶, the UMCC application is rather a tool for getting intuitions about tendencies in a network than for retrieving precise numerical information from it.⁷

4. Results

Two levels of analysis are conducted: a macro level, which groups certain forms into functional families (cf. section 2.2), and a micro level, which examines individual collocates (cf. section 2.1). Bar graphs displayed in Figures 3 and 4 display the results for the associations between two families of forms (modal verbs and deictic pronouns) and the selected connectors across the two corpora.

⁵ We use the D3 JavaScript library for computing the force simulation and rendering the network (<https://d3js.org>).

⁶ Note that due to the random initialization of the nodes also the layout of the exact same network will always be slightly different from previous runs.

⁷ For example, the tool can reveal a proximity between two connectors and make it possible to formulate a hypothesis about a shared functioning, which can then be verified through specificity indices and a qualitative analysis of the lists of co-occurrences (see Table 2).

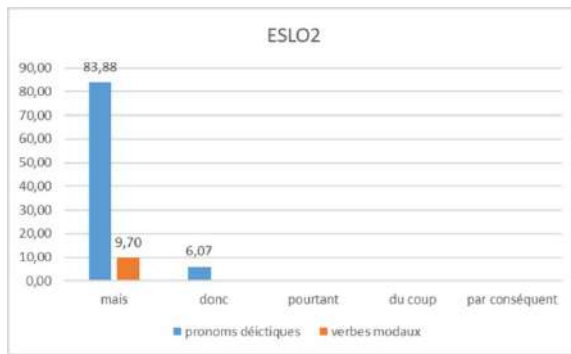


Figure 3: macro level in ESLO2.

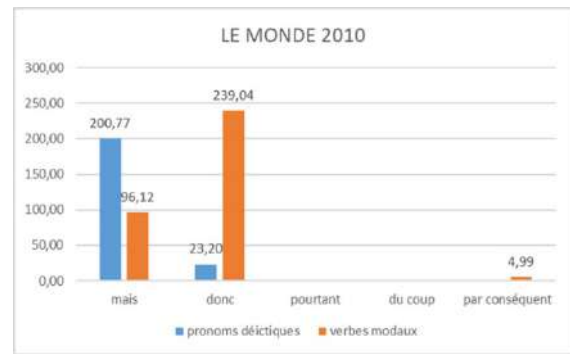


Figure 4: macro level in Le Monde.

A first striking result concerns *mais*, which shows a specific association with both families in both corpora. In other words, regardless of the corpus considered, *mais* appears in proximity to both deictic pronouns and modal verbs. This stability across corpora that are very different in terms of medium and register suggests a stable contextual profile for this connector. The behaviour of *donc* is more differentiated. In both corpora, it is specifically associated with deictic pronouns, indicating a stable affinity with environments involving enunciative anchoring. On the other hand, the association with modal verbs only appears in *Le Monde*. This result suggests that the proximity between *donc* and modality markers is sensitive to the discourse genre. Conversely, the distributional profiles of *pourtant* and *du coup* appear relatively neutral from the point of view of these two families. This does not mean that they are devoid of contextual constraints, as we will see in the following sections. As for *par conséquent*, it appears to be a connector of formal writing – there is only one occurrence in *ESLO2* which makes the examination of the collocates irrelevant. In *Le Monde* it shows only an attraction with the family of modal verbs, unlike *mais* and *donc* which are attracted to both families. This attraction of the three connectors to modal verbs in the written press may offer a way to see these connectors as privileged markers of written argumentation. Overall, these results illustrate the value of an approach based on co-occurring environments, highlighting that *mais* and *donc* share characteristics and that *mais* exhibits stable regularities across very contrasting corpora.

At the micro level, we can examine each co-occurring form, including those belonging to each of the two families introduced earlier. For each connector, we extracted all collocates within a six-item window (three items to the left and three to the right⁸), excluding proper nouns and acronyms. Figures 5 to 13 show a selection of the collocates for each connector. Our selection is partially driven by the corpus: we observed, within the lists of all collocates, recurring items that pertain to the same semantic field. In the figures, the collocates are colour-coded in relation to the type of semantic indication they convey. The blue and orange collocates pertain to the semantic field that has been computed as two families (cf. macro level). The deictic *hier* (light blue) represents a temporal deictic form and has therefore been considered separately; the pronoun *on* (dark blue), having a significant proportion of its uses that are not deictic, has also been considered apart. The yellow collocates pertain to the field of negation, the light green collocates to the field of modal adjectives, the dark green collocates to the field of modal adverbs, the red collocates to the field of interjections.

⁸ The relevant span depends on each connector and is defined following a series of tests (see Rossari et al. 2024). For this study, we chose the smallest possible span to ensure a semantic association between the connector and the form being examined, without reducing it to a one-item span, which would have caused us to miss too many relevant results.

Occ	Fréq	CoFréq	Indice	DistMoy
PRO:PER_je	51068	4586	73,4799	1,3378
PRO:PER_me	8720	757	10,0846	1,7543
PRO:PER_moi	7321	651	10,4375	0,977
PRO:PER_toi	1327	120	2,785	1,35
VER:cond_devoir	161	21	2,4377	1,5238
VER:pres_falloir	3178	337	14,1775	1,3412
VER:pres_pouvoir	4270	362	4,4363	1,7873
ADJ_possible	335	44	4,4822	1,0909
ADV_forcément	602	77	6,7984	1,9351
ADV_peut-être	1942	248	19,8042	1,5403
ADV_non	15883	3131	127	0,5755
ADV_pas	29655	3582	230,6713	1,8554
ADV_rien	1217	163	15,1531	1,7117
INT_non	241	34	4,2275	1,4118
ADJ_nul	92	15	2,8188	1,2
ADV_jamais	1203	121	4,5999	2,0248
INT_hé	527	64	5,046	0,8594
INT_hein	9149	816	13,1114	1,2365
INT_hm	40600	3012	4,9697	1,1531

Figure 5: collocates of mais in ESLO2.

Occ	Fréq	CoFréq	Indice	DistMoy
VER:impf_falloir	404	37	2,2187	1,3243
ADV_forcément	602	59	3,9103	1,322
PRO:PER_on	28355	1968	12,5534	1,1784
INT_euh	76431	8303	127	0,6574
INT_hm	40600	4244	278,2463	1,278

Figure 7: collocates of donc in ESLO2.

Occ	Fréq	CoFréq	Indice	DistMoy
NOM_Vous	40	3	3,45	1
PRO:PER_je	26613	195	20,978	1,2769
PRO:PER_me	8020	44	2,7101	1,3409
PRO:PER_nous	25520	182	18,4983	1,2253
PRO:PER_te	477	6	2,1961	1,3333
PRO:PER_vous	9937	61	4,7949	1,0328
VER:cond_devoir	6818	98	30,2967	0,0612
VER:cond_falloir	920	23	12,3993	0,087
VER:cond_pouvoir	7483	87	21,0292	0,1034
VER:futu_devoir	1928	41	18,9334	0,2195
VER:futu_falloir	1029	21	9,7842	0,3333
VER:impf_falloir	1034	18	7,4514	0
VER:infir_devoir	630	9	3,4211	0,1111
VER:infir_pouvoir	1388	14	3,404	0,4286
VER:pper_devoir	3026	24	3,7469	0,4167
VER:pper_falloir	669	15	7,7273	0
VER:pper_pouvoir	3168	24	3,464	0,5833
VER:pres_devoir	11048	138	35,8748	0,1377
VER:pres_falloir	8549	213	106,7378	0,108
VER:pres_pouvoir	15938	153	27,4699	0,1961
ADV_éventuellement	154	5	3,6925	1
ADV_forcément	739	8	2,3636	0,75
ADV_potentiellement	185	8	6,5106	0,875
ADJ_nécessaire	2250	25	6,2915	0,96
ADJ_possible	3155	31	6,4716	0,7742
ADJ_potentiel	548	9	3,8628	2,1111
ADJ_probable	497	9	4,1809	0,3333
INT_ah	121	3	2,0741	2
PRO:PER_on	37542	213	11,4888	1,4225
ADV_ne	117348	901	103,428	1,2442
ADV_pas	80157	641	80,2719	0,4275
NOM_rien	2567	25	5,2975	2
PRO:IND_aucun	6035	50	7,583	1,1

Figure 8: collocates of donc in Le Monde.

Occ	Fréq	CoFréq	Indice	DistMoy
VER:infir_devoir	630	2	3,1854	2
VER:pres_devoir	11048	4	2,3724	0,25
VER:cond_nécessiter	41	2	5,5587	1

Figure 13: collocates of par conséquent in Le Monde.

Occ	Fréq	CoFréq	Indice	DistMoy
PRO:PER_je	26613	1350	177,9097	0,5948
PRO:PER_me	8020	285	15,6709	1,8491
PRO:PER_moi	2299	106	12,3973	1,0943
PRO:PER_te	477	25	4,3613	1,48
PRO:PER_tu	977	47	6,4741	1,2766
VER:cond_falloir	920	53	9,8192	1,434
VER:cond_pouvoir	7483	196	2,7051	1,5
VER:futu_falloir	1029	82	23,0603	1,2683
VER:impf_falloir	1034	38	3,0066	1,4211
VER:pres_devoir	11048	358	13,8085	1,5559
VER:pres_falloir	8549	488	80,9085	1,4119
VER:pres_nécessiter	279	14	2,5472	1,0714
VER:pres_pouvoir	15938	680	61,3535	1,7397
ADJ_nécessaire	2250	92	8,2964	1,1196
ADJ_certain	1789	70	5,8436	1,9286
ADJ_probable	497	26	4,4948	2,0385
ADV_certainement	380	35	12,1252	1,2571
ADV_peut-être	2759	230	66,2405	1,5913
ADV_probablement	747	37	5,5426	1,7297
ADV_forcément	739	47	10,1922	1,617
ADJ_nul	929	45	6,3123	0,5778
ADV_guère	1289	43	2,5265	2,1395
ADV_non	9113	477	67,6522	1,6709
ADV_nullement	120	15	7,3474	1,6
ADV_pas	80157	4903	127	2,0577
KON_ni	5951	184	6,2383	2,0924
ADV_jamais	6412	336	48,1177	1,628
ADV_rien	4450	333	83,2988	1,3634
PRO:PER_on	37542	1487	108,6129	0,9213

Figure 6: collocates of mais in Le Monde.

Occ	Fréq	CoFréq	Indice	DistMoy
PRO:PER_tu	13042	94	9,089	0,9043
ADV_ben	2531	25	4,9977	0,56
INT_bah	12969	94	9,2075	1,3511
INT_euh	76431	339	4,082	1,1534
NOM_beh	60	3	2,8626	1

Figure 9: collocates of du coup in ESLO2.

Occ	Fréq	CoFréq	Indice	DistMoy
PRO:PER_on	37542	21	4,5614	0,9048

Figure 10: collocates of du coup in Le Monde.

Occ	Fréq	CoFréq	Indice	DistMoy
VER:infir_hier	49	2	3,1867	0
ADV_jamais	1203	10	7,3707	2,2
PRO:PER_on	28355	44	5,012	1,3636

Figure 11: collocates of pourtant in ESLO2.

Occ	Fréq	CoFréq	Indice	DistMoy
VER:infir_falloir	162	3	2,289	0
ADJ_nécessaire	2250	12	2,4286	1,4167
ADJ_probable	497	5	2,3494	1,6
ADJ_nul	929	7	2,3868	1,1429
ADV_guère	1289	13	5,2347	0,5385
ADV_jamais	6412	40	8,3494	0,95
ADV_ne	117348	610	83,4656	1,5197
ADV_pas	80157	328	26,6592	0,6829
ADV_rien	4450	35	9,9136	0,9429
KON_ni	5951	25	2,8829	1,52
NOM_rien	2567	17	4,2712	1,4118
PRO:PER_on	37542	122	5,2559	1,4754

Figure 12: collocates of pourtant in Le Monde.

In *ESLO2*, *mais* has a profile that highlights the speaker's voice. It is associated with deictic pronouns, modal verbs (particularly *falloir* and *pouvoir*), modal adverbs, forms of negation and interjections. These patterns confirm its role in spoken discourse characterized by subjectivity and interactive management of the conversation. In *Le Monde*, *mais* shows a comparable profile that also foregrounds the speaker's voice in a discourse marked by subjectivity and dialogue management. Except for not having interjections as collocates and for showing an attraction with the pronoun *on*, it is associated with the same categories of forms (deictic pronouns, modal verbs, modal adverbs, and forms of negation).

In *ESLO2*, *donc* is associated with the verb *falloir*, modal adverbs, the pronoun *on* and interjections. These associations show a very similar profile to that of *mais*, with a strong link to forms highlighting the speaker's subjectivity. In *Le Monde*, *donc* presents a very similar profile, associating with modal verbs, modal predicates, modal adverbs, negation and the pronoun *on*. Both *mais* and *donc* show a very stable profile in the two corpora, despite the different genres represented, which emphasises the convergence between the two connectors.

In *ESLO2*, *du coup* is mainly associated with the deictic pronoun *tu* and various interjections. This profile highlights its dialogical anchoring in spontaneous contexts. In *Le Monde*, *du coup* is only associated with the pronoun *on*, showing no traces of such a profile⁹. Interestingly, in this corpus, it is *donc* that is associated with an interjection and not *du coup* – contrary to the representation we have of the use of *du coup*, showing a posture of spontaneity in discourse.

Pourtant has a profile sharing similarities with both *mais* and *donc*. In *ESLO2*, it is associated with the pronoun *on*, a form of negation (*jamais*) and a temporal deictic (*hier*). It thus seems to show an anchoring in the situation of enunciation – even though this must be verified by sorting the occurrences of the deictic *on* only – and a potentially dialogical anchoring through the strong negation staged by *jamais*. In *Le Monde*, the profile of *pourtant* is much more strongly characterised by an association with negation, with a whole series of items including *jamais*. This makes it similar to the profile of *mais*. There is also an attraction to the pronoun *on* and to deontic and epistemic modality, similar to what we have noted for *donc* and *mais*. Its combinatorial profile makes it stand out as a strong dialogical connector, even in written language. In this respect, *pourtant* is comparable to *mais*, which is the only other connector to be attracted to negation in both corpora.

Par conséquent stands out as it is attested only once in *ESLO2*, making it impossible to extract collocates. In *Le Monde*, it is associated with deontic modality: *devoir* and *nécessiter*. This makes it the connector most closely associated with a construction characteristic of formal written discourse, setting it apart from the profiles of *du coup* and *donc*.

5. Conclusion

The study based on the statistically significant combinatorial properties of the connectors considered highlights:

- properties that are stable (attraction to deixis or negation in both corpora for certain connectors);

⁹ At first glance, one might think that this is due to the fact that *du coup* is a marker typically associated with spoken language. However, in *Le Monde*, the number of occurrences of *du coup* (437) is largely higher than the number of occurrences of *par conséquent* (130), traditionally associated to a written register.

- properties that bring together connectors traditionally associated with different relations (*donc* and *mais*);
- properties highlighting peculiarities of certain connectors (*par conséquent*);
- properties that bring together connectors sharing a relation based on a form of opposition (*mais* and *pourtant*).

These profiles make it possible to anchor the study of connectors to effective properties derived from real and representative uses, which would not emerge without the statistical calculations measuring the deviation between the expected use and the observed use. The combination of the collocate lists and the UMCC allows us to obtain the most complete view of how connectors function. On the one hand, the UMCC visualization tool shows the major trends in the functioning of connectors and facilitates comparisons between connectors. On the other hand, the collocate lists provide a more precise view of the functioning of a given connector, showing its collocations with items that can be grouped into families, which can in turn be integrated again into the visualization tool. The proposed methodology showed that it is possible to go beyond the circularity bias regarding the classification of connectors. Such a methodology can be extended to other forms and other languages, providing access to the functioning of a form, e.g. pointing out what is genre- or language-related and/or what is governed by cognitive mechanisms.

References

- Abouda, L. (2022). L'émergence du marqueur méta-discursif du coup: de la conséquence à l'actualisation énonciative. *Langages*, 226(2), 99-116.
- Biber, D. (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge Univ. Press.
- Evert, S. (2008). Corpora and collocation. In A. Lüdeling & M. Kytö (éds.), *Corpus Linguistics*. An International Handbook (pp. 1212-1248). Berlin: Mouton de Gruyter.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in linguistic analysis*, Oxford: The Philological Society, 1-32. Reprinted in Palmer (1968). *Selected Papers of J. R. Firth 1952-59*, London: Longmans, 168-205.
- Heiden Serge, Magué Jean-Philippe, & Pincemin Bénédicte. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, 1021–1032.
- Lafon, P. (1981). Analyse lexicométrique et recherche des cooccurrences. *MOTS*, 3, 95-148.
- Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, 1, 127-165.
- Pander Maat, H. & Degand, L. (2001). Scaling causal relations and connectives in terms of Speaker Involvement. *Cognitive Linguistics*, 12, pp. 211-245.
- Rossari, C., Montrichard, C., Ricci, C. et Sanvido, L. (2024). Vers un schéma cognitif de l'exemplification : étude textométrique et contrastive de formes de l'exemplification en français et en italien. *Langue française*, 222(2), 59-76.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Corpus

Laboratoire Ligérien de Linguistique - UMR 7270 (LLL) (2025). *ESLO-FLEU : Enquêtes Sociolinguistiques à Orléans - Fle et Linguistique pour l'Enseignement Universitaire* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v2, <https://doi.org/10.82270/eslo-fleu/v2>.

The Motif Work Bench Pipeline: From Sequential Pattern Mining to Multidimensional Statistics

Hugo Dumoulin¹, Timothée Premat²

¹MoDyCo, Université Paris Nanterre – hdumoulin@parisnanterre.fr

²Neuchâtel University/SNSF – timothee.premat@unine.ch

Abstract

This paper presents the Motif Work Bench (MWB) pipeline, a textual data analysis application designed for the analysis of discursive routines using textual corpora. Situated within the field of discourse analysis, the MWB approach relies on the notion of motif. Motifs are defined as statistically significant linguistic sequences capable of carrying discursive functions characteristic of a particular genre. MWB aims to fill a gap in the state of the art: the lack of a tool bridging advanced *corpus-driven* (Longrée & Vanni 2025) algorithms for the automatic extraction of sequential motifs like CloSPEC (Béchet et al. 2015) with a multidimensional statistical framework (Lebart & Salem 1988) enabling their discursive interpretation.

Keywords: pattern mining, textometry, multidimensional statistics, discourse genres, discursive routines

Résumé

Cette communication présente la chaîne Motif Work Bench (MWB), un dispositif de linguistique outillée destiné à l'analyse des routines discursives à partir de corpus textuels. Inscrite dans le champ de l'analyse du discours, cette démarche s'appuie sur la notion de motif, entendue comme une séquence linguistique statistiquement significative susceptible de porter une fonction discursive caractéristique d'un genre. MWB vise à combler un manque dans l'état de l'art : l'absence d'outil réalisant l'articulation entre des algorithmes avancés et *corpus-driven* (Longrée & Vanni 2025) d'extraction automatique de motifs séquentiels comme CloSPEC (Béchet et al. 2015) et un dispositif de statistiques multidimensionnelles (Lebart & Salem 1998) permettant leur interprétation discursive.

Mots clés : fouille de motifs, textométrie, statistiques multidimensionnelles, genres de discours, routines discursives

1. Introduction¹

Discourse analysis nowadays relies heavily on the observation of discursive *routines*. Understood as “the articulation between a more or less fixed lexicosyntactic pattern and textual or discursive functions” (Née, Sitri & Vénard 2016, our translation), routines thus materialize as a linguistic feature. This matches what Bakhtin identifies as the defining feature of discourse

¹ This work was supported by a postdoctoral fellowship funding from University Paris-Est Créteil (for the second author), as part of the ArchivU project.

genres, namely their reliance on “relatively stable types of utterance”, characterized by a “selection made within the resources of language—lexical, phraseological and grammatical means—but also, and above all, [...] compositional construction” (1984: 265, our translation).

Within this framework, a decisive issue emerges: identifying the routines that may be said to characterize a genre. Given the “two-sided” nature of the notion of discourse genre, between language and discourse (Branca-Rosoff 1999: 116), identifying the discursive routines specific to a genre involves both analyzing the repetition and fixation of certain linguistic forms and interpreting their discursive function in context. To this end, methods from computational linguistics are particularly valuable.

In this context, the notion of sequential “motif” can be used to define the formal component of what may be interpreted as a discursive routine. This notion comes from stylistic studies of literary texts (Mellet & Longrée 2012) and computer science (Agrawal & Srikant 1995). One key challenge lies in articulating the various automatic motif-extraction methods developed in natural language processing (NLP) with the visualization of their distribution across a corpus, which can support the interpretation of their function within the genre. One way to address this challenge is to use multidimensional statistics (Benzecri 1973; Lebart & Salem 1988), long-established methods in textometry recognized for their contribution to discourse analysis (Lethier 2009). As a result, the concept of motif now appears as a meeting point between NLP and textometry, brought together by the perspective of discourse analysis. This concern for interdisciplinary dialogue is particularly evident in research within a field that has become especially active in recent years (Legallois, Longrée, Novakova, Silvestre de Sacy 2025).

In this respect, a methodological issue proves decisive and motivates the present study (Longrée & Vanni 2025). Many sequential motif-extraction and visualization tools adopt a *corpus-based* perspective, in which the researcher identifies motifs following a query built around a pivot form that embodies a research hypothesis; this is for instance the logic behind the concordancer or index functions of state-of-the-art software such as TXM (Heiden, Magué & Pincemin 2010) or Hyperbase (Vanni 2024). Another approach is possible, described as *corpus-driven*. This alternative would not rely on a pivot form but would instead captures all motifs present in the corpus given formal conditions of frequency or length, thereby reducing the initial role of research hypotheses. This is the logic underlying software such as SDMC (Béchet et al. 2013), which, however, does not provide tools for multidimensional statistical visualization, unlike Hyperbase or TXM, and tends to generate redundant results, as already noted (Longrée & Vanni 2025). It is that context that motivates the current proposal: the Motif Work Bench (MWB) pipeline, designed to contribute to bridging the gap between *corpus-driven* sequential motif extractors and multidimensional statistical visualizations.

2. From set-motif to sequential multidimensional motifs

A motif is defined as a type of set. Before being adopted by linguists, the concept was used in marketing, for instance to define shopping baskets by gathering items frequently bought together. The representation of such motifs is not necessarily structured. Transposed into linguistic research, the motif *on the basis of*, can be represented such as (1), where no linear or hierarchical structure is applied.²

² These examples are from de Brown corpus (Francis & Kučera 1964).

- 1) Set-motif $\left\{ \begin{array}{c} \textit{basis} \\ \textit{of} \\ \textit{on} \\ \textit{the} \end{array} \right\}$

Given the limitations of a representation such a (1) for linguistic studies, structured-motifs are preferred to set-motifs. Motifs are usually structured based on sequentiality ('unit α precedes unit β ') and/or on hierarchical, morphosyntactic structure ('unit α is dominated by unit β '), such as in (2-3).

- 2) Sequential pattern: $\{\textit{on the basis of}\}$
 3) Hierarchical motif: $\{\textit{on} [\textit{the} [\textit{basis}] \textit{of}]]\}$

The motifs implemented in MWB are sequential motifs; they can also incorporate syntactic structure but not as their structural defining feature. In MWB, syntactic information is implemented as one level of representation among others, through dependency relations (Dep). Units can also be defined by other levels, such as their lemma or their part of speech (POS). If a motif is represented through different levels, the motifs become multidimensional sets, such as the one in (4).

- 4) Sequential, multidimensional motif:
 $\{ \{ \text{pos=PRE, lemma=on} \} \{ \text{pos=DET} \} \{ \text{pos=NOUN, lemma=basis} \} \{ \text{pos=PRE} \} \}$

With a representation such as (4), each property defining a component of the motif is called an *item*, and the collection of properties defining a component are called *itemsets*. A multidimensional sequential motif is thus defined as an ordered collection of itemsets, each one comprising items. Combining different abstract levels of representation allows the capture of motifs despite surface variation. (5), a version of (4) where the first itemset is only specified as a preposition (without lemma specification), would capture not only the previous pattern (5.a), but also the patterns (5.b-d).

- 5) The motif representation:
 $\{ \{ \text{pos=PRE} \} \{ \text{pos=DET} \} \{ \text{pos=NOUN, lemma=basis} \} \{ \text{pos=PRE} \} \}$
 captures:
 a. *on the basis of*
 b. *as a basis for*
 c. *as the basis for*
 d. *as a basis of*

This is one of the differences between MWB and the approach taken by Silvestre de Sacy & Legallois' *PyMotifs* and *MotiveR* scripts, which reduce *a priori* the multidimensionality of itemsets by representing each word through only one level (e.g., some function words are only represented by their lemma, some lexical words only by their POS; Silvestre de Sacy & Legallois 2025). In contrast, MWB remains agnostic toward which level of representation better captures motifs and leaves this decision to frequency-based criteria. Indeed, MWB uses the CloSPEC extraction algorithm (Béchet et al. 2015), which uses a frequency criterion (only collect patterns with a minimal support frequency) and a closure criterion (do not collect a smaller pattern contained in an equally frequent larger pattern).

However, MWB does not follow Mekki et al. (2020) in using growth rate (the ratio between motifs' support frequency in two corpora) as a method to qualify motifs characteristic of a corpus compared to another, in a one-versus-one logic. Instead, we propose to rely on multidimensional statistics to explore the relevance of motifs in several contrastive subcorpora,

a strategy that enables many-to-many comparisons and is widely used in the textometry community.

3. Description of the pipeline

To use multidimensional statistics onto the extracted motif, MWB uses a Corpus Work Bench (CWB)-encoded (Evert & Hardie 2011) version of the corpus in which it looks for the extracted motifs. MWB's main architecture is a Python script, but it uses Perl for the CWB-encoding and R for more accurate statistics.

The typical workflow of the MWB pipeline is represented in Figure 1, when the script is run on a new corpus (i.e., if MWB does not have access to previously generated results). It starts by tagging and parsing the corpus into CoNLL-U files using Stanza (Qi et al. 2020). Then, the pipeline splits in two paths: from the CoNLL-U files, MWB produces VRT files for CWB encoding and DMT4 files (Nanni & Rigotti 2007) for motif extraction. CloSPEC extracts motifs from the DMT4 given the settings of the user (see next section), and a clustering operation is applied based on motif similarities, in order to reduce the redundancy in extracted motifs (Mekki 2022); clusters are represented as their medoid using the S2MP method (Saneifar et al. 2008). Finally, the list of the medoids is queried through CQP (from CWB) to produce a contingency table of motifs against texts. This contingency table is sent to an R script which uses a Graphic User Interface (GUI) produced by the package Shiny to allow the user to explore different multidimensional statistical approaches, using mainly the *FactoMineR* (Lê, Josse & Husson 2008), *factoextra* (Kassambara & Mundt 2020) and *textometry* (Loiseau et al. 2026) packages. The columns of the contingency table can be modified based on the one of a metadata.csv file, for instance to observe the distribution of motifs against texts, genre, periods, etc.

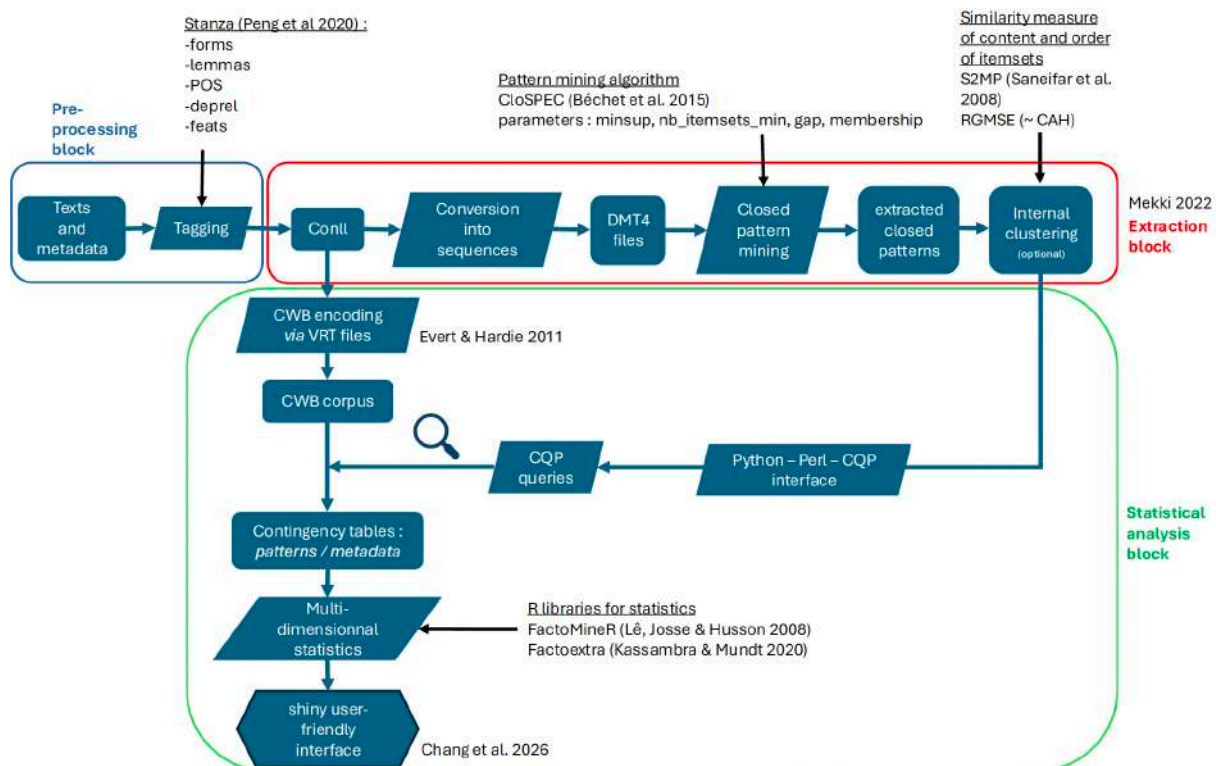


Figure 1. Technical diagram of MWB.

4. Settings and multidimensional statistics

MWB allows the user to pass some settings to the pipeline, including those of the CloSPEC extractor. These settings include the level(s) of representation for itemsets (typically lemma, POS and Dep), the minimal size requirement for motifs (*itemset_min*: number of itemsets needed for a motif), the minimal and maximal gap allowed in a motif (*gapmax* and *gapmin*), and the use of clustered motifs or raw ones. One key setting is the *minsup*, the minimal support frequency. It requires that a given motif be present in a percentage of all supports (i.e., what is tokenized as a sentence in the CoNLL-U file) in the corpus. If the *minsup* is set to 3, only motifs present in at least 3% of all sentences will be kept. It is possible to pass several values of *minsup* to MWB (as a comma-separated list), in which case MWB computes different datasets for each *minsup*. It is also possible to apply a membership constraint, forcing motifs to contain lemmas selected based on decreasing frequency, specificity (Lafon 1980; Loiseau et al. 2026) or a user input list. The settings for *itemset_min*, *minsup*, *gapmin* and *gapmax* allow multiple values. For each combination of values, a different dataset of motifs is produced, and the user can switch between these datasets in the exploratory statistics interface.

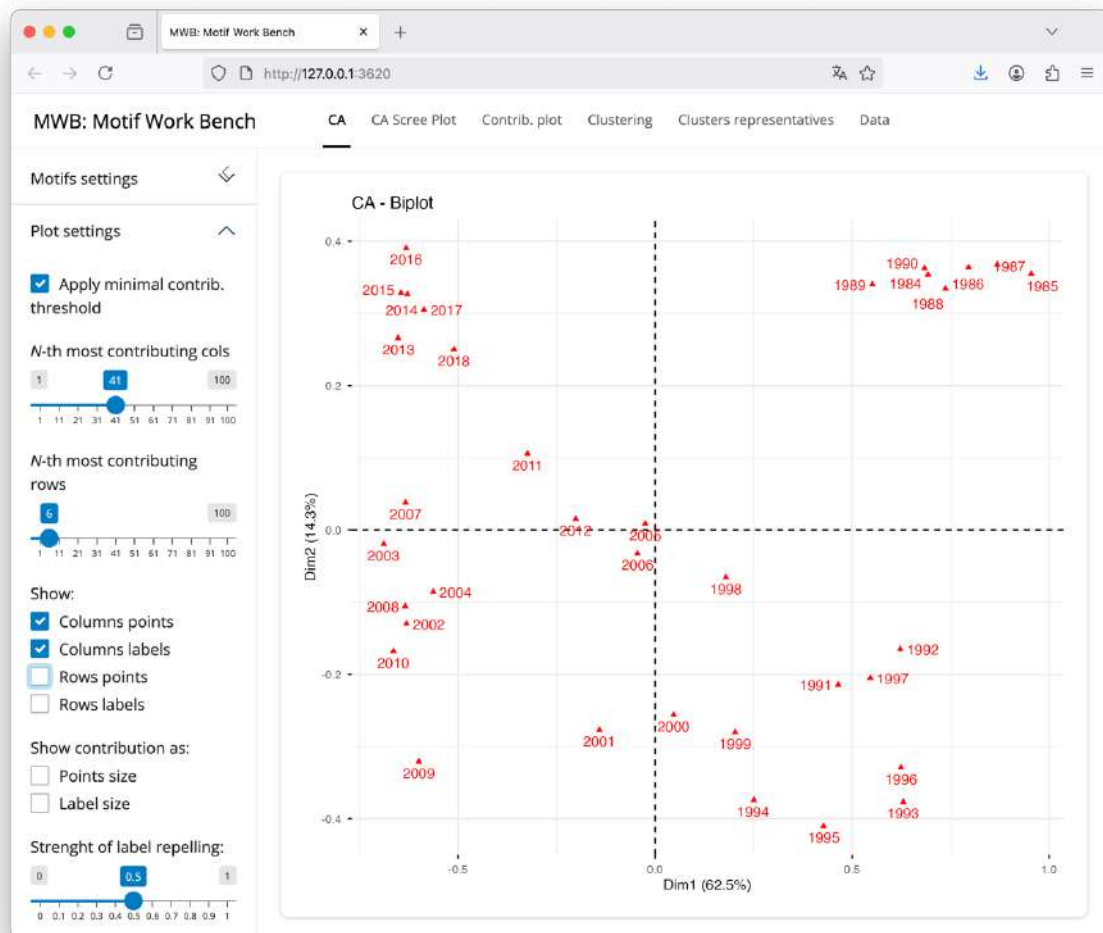


Figure 2. A glimpse at MWB's statistical exploratory interface

At the end of the pipeline, MWB offers a graphic user interface that allows the user to navigate between several statistical views: a correspondence analysis (Benzecri 1973) shows the distribution of motifs between subcorpora and their statistical association; an agglomerative clustering (Ward method) allows the user to identify similarities between groups of motifs or subcorpora; and the cluster representative tab enables the visualization of the representatives of each cluster. For all these views, the user can compare the effect of different metadata or motif extraction settings. For instance, in figure 2, the data is split according to the year of the texts, while in figure 3 below the splitting metadata is the decades.

5. Use case: exploring routinization

Here we examine an example of the range of uses enabled by Motif Work Bench in the exploration of the ArchivU corpus (Dumoulin, Sitri, Mellet 2022), a diachronic corpus of minutes from university council meetings at Paris Nanterre University covering the period 1984–2018. These minutes, a genre of grey literature within professional writing in the administrative sphere, prove to be particularly conducive to the production of discursive routines.

An instantiation of MWB is carried out by selecting the linguistic parameters of lemma and dependency relation (Dep) as the description of each itemset, and by choosing a minimum itemset size of 5 items, a minsup of 0.5 (meaning that 0.5% of the sentences in the corpus must contain a motif for this motif to be selected). A membership constraint is used to ensure that motifs include at least one of the 1,000 most frequent lexical words (adjectives, nouns, or verbs) in the corpus, which helps obtaining more directly interpretable motifs.

The extracted motifs are then represented according to their distribution across texts by decade using correspondence analysis (CA; Benzecri 1973), in which, through a set of parameters in the Shiny graphical interface, the 41 most contributive points are highlighted (figure 3).

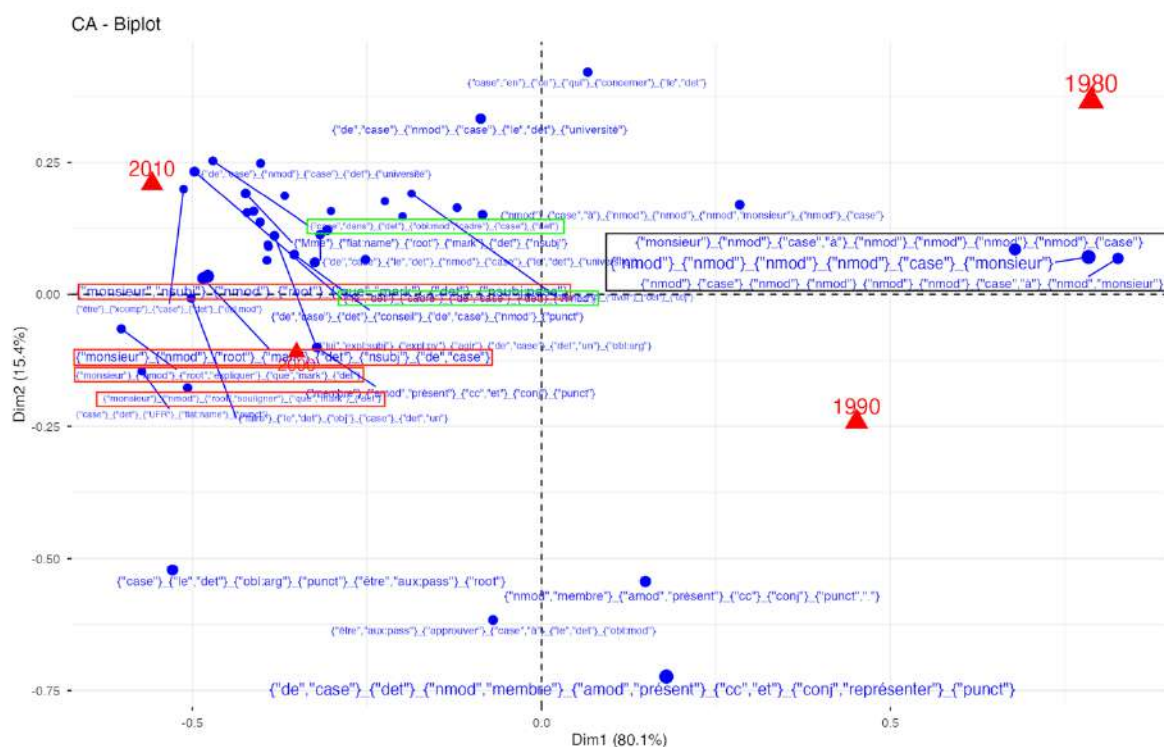


Figure 3. Correspondence analysis of motifs by decade

This CA makes it possible to represent routinization phenomena occurring at different levels. We can already note the overall structure of this CA, which clearly contrasts the 1980s–1990s with the 2000s–2010s along the principal (east–west) axis, accounting for 80.1% of the total inertia of the corpus.

5.1. Textual practices

First, if we focus on the motifs represented on the eastern side of the graph, we observe the following set of motifs (highlighted in black on the CA in figure 3):

- 6) {"monsieur"} {"nmod"} {"case", "à"} {"nmod"} {"nmod"} {"nmod"} {"nmod"} {"case"}
- 7) {"nmod"} {"nmod"} {"nmod"} {"nmod"} {"case"} {"monsieur"}
- 8) {"nmod"} {"case"} {"nmod"} {"nmod"} {"nmod"} {"nmod"} {"case", "à"} {"nmod", "monsieur"}

These motifs, which may at first appear cryptic, in fact correspond to different realizations of the same initial textual structure found in the minutes from the 1980s–1990s, a structure devoted to the list of proxies:

- 9) [*PROCURATION EN DÉBUT DE SÉANCE*] M. REY à M. PIN M. HARTWEG à M. HIMY (1988-09-26)
'[Proxies appointed at the start of the session] Mr. Rey to Mr. Pin Mr. Hartweg to Mr. Himy'

During these decades, the proceedings open on a page of vertical lists (sometimes with several columns), gathering the names and status of the members of the board, including the list of proxy appointments. These raw lists are not triggered by any full text phrase but are governed by low-level titles (*Procuration en début de séance* in 11). What motifs (6-8) capture is the singularity of the proxies' list in these proceedings, represented under the abstraction of a sequence of nominal modifying relations (nmod) with or without prepositions (case). This prototypical, formal sequence contrasts sharply with its counterpart in the 2000s–2010s, where the proxies appear more often integrated in the text, and in fully syntactic phrases, such as in (10):

- 10) *Mme Effosse donne procuration à M. Ragot ; M. Polit donne procuration à M. Fouque ; Mme Cheze donne procuration à M. Esteban* (2016-04-18)
'Ms. Effosse appoints Mr. Ragot as her proxy; Mr. Polit appoints Mr. Fouque as his proxy; Ms. Cheze appoints M. Esteban as her proxy'

Here, MWB has captured differences in writing practices that can be understood as routines at the level of textual organization. This is a clear case of corpus-driven serendipity, as we did not focused on the proxies' list during our close readings.

5.2. A standardization of reported speech forms

Next, if we focus more specifically on the motifs located on the western half of the graph—thus associated with the 2000s–2010s—we can observe the following series of motifs (11-14), highlighted in red on the CA:

- 11) {monsieur, "nsubj"} {"nmod"} {"root"} {"que", "mark"} {"det"} {"nsubj:pass"}
- 12) {monsieur, "nsubj"} {"nmod"} {"root"} {"que", "mark"} {"det"} {"nsubj"} {"de", "case"}
- 13) {monsieur, "nsubj"} {"nmod"} {"root", "expliquer"} {"que", "mark"} {"det"}
- 14) {monsieur, "nsubj"} {"nmod"} {"root", "souligner"} {"que", "mark"} {"det"}

These are introductory forms of indirect reported speech, of the type "*Monsieur Balaudé explique que le...*" ('Mister Balaudé explains that the...'). What makes this observation

noteworthy is that the entire corpus – including the 1980s–1990s – is written in indirect reported speech (IS). Thus, the striking phenomenon is the statistical association, with the 2000s–2010s, of this specific configuration of indirect reported speech: using a *that*-clause rather than an infinitival construction, and favouring the verbs *expliquer* and *souligner* (‘to underline’).

These motifs point to a discursive routine in minute writing that emerges around the turn of the 2000s, characterized by a rigidification of the forms of indirect speech employed as well as a standardization of the reporting verbs used (*expliquer*, *souligner*). This observation obtained with MWB aligns with and confirms similar findings derived from the same corpus using other statistical tools (Sitri 2022, Lethier & Wang 2022).

5.3. Indicators of grammaticalization

Finally, it is interesting to interpret the position of the motif (15-16), highlighted in green on the CA:

- 15) {“case”, “dans”} {“det”} {“obl:mod”, “cadre”} {“case”} {“det”}
 16) {“le”, “det”} {“cadre”} {“de”, “case”} {“det”} {“nmod”}

This motif corresponds to realizations that vary only slightly around the sequence “dans le cadre de...” (‘as part of/within the framework of’). The position of this motif among the most contributive ones on the western side of the graph is, once again, particularly noteworthy, as it associates a fixed use of the common noun *cadre* within a prepositional phrase with the 2000s–2010s. This observation can be interpreted as an indicator of the increased grammaticalization of the noun *cadre* over the period, as the form loses its distributional freedom. Returning to the texts, one indeed observes uses of the form that increasingly erase its original semantics. Thus, one may contrast examples in (17-19) from the 1990s with examples (20-23) from the 2000s–2010s.

- 17) *dans le cadre de la loi* (1984-01-16)
 ‘under the law’
 18) *dans le cadre de la contractualisation* (1991-02-18)
 ‘as part of the contracting process’
 19) *dans le cadre d’un partenariat État, Région, Département* (1994-07-04)
 ‘within the framework of a partnership with the State, the Region and the Department’
 20) *dans le cadre des travaux du carré central* (2003-09-29)
 ‘as part of the works on the central yard’
 21) *dans le cadre de l’aménagement du secteur autour de l’université* (2005-06-27)
 ‘as part of the development of the area around the university’
 22) *dans le cadre des enseignements à distance* (2012-05-03)
 ‘as part of distance learning’
 23) *dans le cadre des élections présidentielles de 2017* (2016-07-11)
 ‘in the context of the 2017 presidential election’

Here, we observe another form of routinization, operating at an infra-predicative level. It makes it possible to consider the hypothesis that administrative professional genres such as minutes may have their own contribution to processes of linguistic change, through the grammaticalization of certain nouns into prepositions.

Thus, through its fully corpus-driven approach, MWB yields a set of results which, by their formal nature, call for interpretation. Nevertheless, the inductive nature of the method, combined with correspondence analysis, makes it possible to bring to light the different levels

of linguistic description at which routinization phenomena may emerge: textual, enunciative, and also infra-predicative.

6. Conclusion

By articulating sequential pattern mining with multidimensional statistics, the Motif Work Bench pipeline provides an integrated framework connecting pattern mining, textometry, and discursive interpretation. Thus, MWB seeks to bridge the gap between *corpus-driven* pattern mining approaches and the textometric framework. Its release under an open-source license is intended to foster adoption by the community and to open new perspectives for the contrastive analysis of discourse genres.³ Although the program is already functional, further development is on-going, including the creation of a cross-platform installer, a motif concordancer and a general GUI for all steps of the pipeline⁴.

References

- Agrawal R. and Srikant R. (1995) Mining sequential patterns. In *Proc. 19th int. conf. on data engineering, IEEE*, p. 3-14.
- Bakhtine M. (1984). *Esthétique de la création verbale*. Gallimard.
- Béchet N., Cellier P., Charnois T. and Crémilleux B. (2015). Sequence mining under multiple constraints. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 908-914. DOI: [10.1145/354756.354849](https://doi.org/10.1145/354756.354849)
- Béchet N., Cellier P., Charnois Th., Crémilleux, B. and Quiniou, S. (2013). SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes. In *Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13)*, Jan. 2013, Toulouse, France.
- Benzecri J.-P. (1973). *L'Analyse des données*. Dunod.
- Branca-Rosoff S. (1999). Des innovations et des fonctionnements de langue rapportés à des genres. *Langage & Société*, 87: 115–129.
- Evert S. and Hardie A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.
- Francis W. N. and Kučera H. (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Brown University.
- Heiden S., Magué J. P. and Pincemin B. (2010) TXM : Une plateforme logicielle open-source pour la textométrie-conception et développement. In *Actes des 10^e JADT*, Rome, 2(3): 1021–31.
- Kassambara A. and Mundt F. (2020) Factoextra – Extract and Visualize the Results of Multivariate Data Analyses, R package documentation.
- Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1: 127-165. DOI: [10.3406/mots.1980.1008](https://doi.org/10.3406/mots.1980.1008)
- Lê S., Josse J. and Husson F. (2008) FactoMineR: an R package for multivariate analysis, *Journal of statistical software*, 25: 1-18.
- Lebart L. and Salem A. (1988). *Analyse statistique des données textuelles*. Dunod.
- Legallois D., Longrée D., Novakova I. and Silvestre de Sacy, A. (2025). Introduction. Motifs, textométrie et stylistique des textes littéraires. *Corpus*, 27. DOI: [10.4000/13woh](https://doi.org/10.4000/13woh)

³ All scripts are available in the following repository: <https://github.com/HugoDumoulin0/MotifWorkBench>

⁴ This development step is supported by an intership funding from MoDyCo laboratory.

- Lethier V. (2009). Exploration textuelle du discours d'un quotidien régional au carrefour du XIX^e et du XX^e siècle : Le Petit Comtois (1883-1903). PhD thesis, Université Franche-Comté.
- Lethier V. and Wang. I. (2022). De l'exploration textométrique d'un corpus de comptes rendus de l'université de Nanterre aux pistes interprétatives sur l'évolution du genre en diachronie". *Cahiers de Praxématique*, 78. DOI: [10.4000/praxematique.8185](https://doi.org/10.4000/praxematique.8185)
- Loiseau S., Decorde M., Heiden S. and Vaudor, L. (2026) Textometry : Textual data analysis Package Used by the TXM Software, *R package documentation*.
- Longrée D. and Vanni L (2025). Identification des motifs textuels. Entre statistique et *deep learning*, *Corpus*, 27. DOI: [10.4000/13woj](https://doi.org/10.4000/13woj)
- Mekki J., Béchet N., Battistelli D. and Lecorvé G. (2020). Caractérisation de registres de langue par extraction de motifs séquentiels émergents. In *Actes des 15^e JADT*, Toulouse, pp. 1-12.
- Mekki J. (2022). *Caractérisation de registres de langue par extraction de motifs séquentiels émergents*. PhD thesis, Université de Rennes.
- Mellet S. and Longrée D. (2012). Légitimité d'une unité textométrique : le motif. In A. Dister, D. Longrée, G. Purnelle, editors, *Actes des Journée d'analyse des données textuelles 2012*, pp. 715-728.
- Nanni M. and Rigotti, C. (2007). Extracting trees of quantitative serial episodes. In *Proc. Of KDID '07*. DOI: [10.1007/978-3-540-75549-4_11](https://doi.org/10.1007/978-3-540-75549-4_11)
- Née É., Sitri F. and Vénier M. (2016). Les routines, une catégorie discursive pour catégoriser les genres ?. *Lidil*, 53. DOI: [10.4000/lidil.3939](https://doi.org/10.4000/lidil.3939)
- Qi P., Zhang Yuhao, Zhang Yuhui, Bolton J. and Manning Ch. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*.
- Silvestre de Sacy A. and Legallois D. (2025). Les bibliothèques *MotiveR* et *Pymotifs* pour identifier les patrons lexico-grammaticaux significatifs (ou motifs séquentiels), *Corpus*, 27. DOI: [10.4000/13wol](https://doi.org/10.4000/13wol)
- Saneifar H. et al. (2008). S2mp: Similarity measure for sequential patterns. *AusDM : Australasian Data Mining*, 87: p. 95-104
- Sitri F. (2022). La genèse d'un compte rendu de réunion : dialogisme et genre de discours. *Cahiers de Praxématique*, 78. DOI: [10.4000/praxematique.7299](https://doi.org/10.4000/praxematique.7299)
- Sitri F., Dumoulin H. and Mellet C. (2022). La Fabrication Discursive de l'université : Compte Rendus et Rapports Scientifiques En Diachronie. *Cahiers de Praxématique*, 78. DOI: [10.4000/praxematique.8135](https://doi.org/10.4000/praxematique.8135)
- Vanni L. (2024). Hyperbase Web. (Hyper)Bases, *Corpus*, Langage. *Corpus*, 25. DOI: [10.4000/corpus.8770](https://doi.org/10.4000/corpus.8770)

The Moving Border Model: Tuning a split point along sequential data. Application to linguistic periodization.

Bénédicte Pincemin¹, Céline Guillot-Barbance², Alexei Lavrentiev³

¹CNRS, ENS Lyon, IHRIM UMR5317 – benedicte.pincemin@ens-lyon.fr

²ENS Lyon, IHRIM UMR5317 – celine.guillot@ens-lyon.fr

³CNRS, ENS Lyon, IHRIM UMR5317 – alexei.lavrentev@ens-lyon.fr

Abstract

Drawing on Salem’s chronological characteristic elements (1988), the moving border model consists in testing every possible split of a chronological corpus into a first part and a last part, and finding the division that maximizes the contrast in the distribution of a given linguistic feature (measured with Lafon’s specificities (1980)). We tested this model on a large diachronic corpus of French texts spanning the 14th to 17th centuries, in order to refine the delineation of the *français préclassique* (Pre-Classical French) period.

Keywords: Textometry, Textual Data Analysis, Corpus Linguistics, Diachronic Linguistics, Periodization, Pre-classical French, Modern French, Textual Time Series, Chronological characteristic elements, Fisher’s exact test.

1. Introduction

Language changes smoothly over time; however, linguists identify successive periods to distinguish between language states. Can textometry help in the precise and data-based setting of time boundaries?

In a nutshell, textometry provides two main approaches in order to sequence diachronic data: either a multivariate process that considers the *overall* vocabulary in the corpus; or a *focused* analysis, that is based on the evolution of some selected words or features. The first approach typically opens with an overall correspondence analysis that shows a Guttman effect in the data. This asserts that the chronological evolution is the main structure in the corpus (Salem, 1988). Then, an unsupervised classification is applied, so as to group elementary texts or time steps into consistent periods that are induced from the data (de Sousa, 2021; Miao & Salem, 2021). A contiguity constraint may be applied to this classification, so as to force the classes to match time spans (Gries & Hilpert, 2008; Diwersy & Luxardo, 2016).

As regards the second approach, computing adjacent characteristic elements (‘spécificités chronologiques’, Salem, 1988, 2021; Lebart et al., 2019) is a means to find out and precisely adjust, for any given word or linguistic feature, the period that best fits its major use in the corpus. Another way to grasp a word’s evolution is to fit its temporal distribution with a parametric function: the parameters then characterize the temporal profile of the word (Feltgen, 2017). Afterwards, curve clustering techniques can identify temporal patterns (such as emerging words, words with a decreasing trend, or evergreen words) (Trevisani & Tuzzi, 2018). A non-parametric version of word temporal profiles involves computing its specificity distribution along the corpus (Guaresi et al., 2021).

In this paper, we present an original non-parametric focused approach derived from the adjacent characteristic element model, the moving border model, which proved successful in refining the boundaries of the *français préclassique* ('Pre-Classical French') linguistic stage. We explain how it can be conducted with open-source software (namely TXM and Calc) and show how this new analytical tool supplements existing textometric features dealing with time series. We also experimented with the overall approach towards periodization, which will be exposed in (Lavrentiev et al. forthcoming).

2. Method

2.1. Corpus

Pre-Classical French follows Middle French and precedes Modern French, or more precisely 'Classical French', which opens the Modern French era. In linguistic research, Pre-Classical French roughly extends from the middle of the 16th century to the middle of the 17th century (Marchello-Nizia & Combettes, 2020: 57-59) – these are the borders we aim to refine. To contrast the period with representative samples of the surrounding periods, we compiled 654 full texts spanning the 14th to 17th centuries (since ancient texts are rarer, we need to cover several centuries in the Middle French period to gather enough data).

Our corpus draws on open-access resources, mainly the open-access portion of the Frantext database, the Epistemon corpus from the Bibliothèques virtuelles humanistes, the Base de français médiéval. It also benefited from the PhraséoRoChe corpus and from the CreNum digital edition to supplement underrepresented decades. Considerable effort was made to harmonize the text markup and standardize the metadata from the various sources in order to create a consistent, usable corpus that we imported into TXM software (Heiden, 2010). In total, the corpus contains 27 million tokens annotated with part-of-speech tags and lemmas, including automatic tagging performed with the TreeTagger tool (Schmid, 1994) and the Presto parameter file (Diwersy et al., 2017) that we used for this study; and each text comes with metadata on date, discourse domain, genre, and type.

The available data did not allow for the creation of a corpus that would be both large and balanced, but the metadata make it possible to monitor variations within the corpus or to define more homogeneous subcorpora. For instance, the LitHist-NarrDram subcorpus selects the 291 texts in the literary or historical fields, restricted to the narrative and dramatic types, because these texts are rather evenly distributed throughout the entire time span of the corpus.

The corpus was first presented at the SIDF 2025 Conference (Guillot-Barbance et al., 2025). A data paper is currently being prepared (Lavrentiev et al., forthcoming) in connection with the publication of the corpus in accordance with the open science FAIR principles.

2.2. Selection of relevant words and paradigmatic modeling

Previous works have already identified several words and linguistic features that are excellent candidates for Pre-Classical French characterization (see Caron, in press, for a recent review). However, these works were based on the careful and expert reading of a few texts and on the bibliographic study of writings of 17th-century linguists (*remarqueurs*). A quantitative analysis of our large corpus could provide new insights about descriptive clues, and especially may reveal unnoticed markers.

To this end, we defined three time “poles” in our more homogeneous LitHist-NarrDram subcorpus, comprising dates that unquestionably fall within each of the three distinct periods: texts from the 1300-1479 pole are indisputably Middle French representatives (2.5 million tokens, 73 texts), the 1550-1619 pole stands for Pre-classical French (3.3 million tokens, 60 texts), then the 1650-1669 pole represents the Classical state of language (1.2 million tokens, 45 texts). Within this subcorpus, frequency variations can be attributed to time and are not affected by substantial variations in text domains and types. We then computed specificities contrasting Middle French pole with Pre-Classical French pole on the one hand (to explore changes characterizing the first border, the one that opens the Pre-Classical period), and Pre-Classical French pole with Classical French pole on the other hand (second border, closing the Pre-Classical period). Specificities measure a cumulative probability using a hypergeometric modeling approach (Lafon 1980). Direct formula implementation raises computational issues (Pincemin 2024); we therefore relied on TXM’s built-in implementation (R package *textometry*).

For each contrast, the output consists in a long table sorted by decreasing specificity score whose head and tail have to be analyzed by a linguist. Three improvements were introduced to help the linguist in her task: (i) computing specificities on lemmas proved more efficient than on word forms, because the list was more concise, linguistically relevant, and the detailed forms did not provide any decisive information; (ii) we separated results for grammatical words¹ and lexical words² in two distinct tables, to avoid mixing the two types of units; (iii) we highlighted the lemma lines with a high specificity score (greater than 10) and a low frequency in one of the two poles, that is, a typical emergence or obsolescence profile.³

In fact, the linguistic units we wanted to observe were not single features, but diachronic paradigmatic sets of features. The linguistic change has to be observed within a paradigm: language evolution is a change in the composition or balance in a paradigm, especially for grammatical words. Working with paradigmatic sets is not only a means to conform to language mechanisms, but also a way to address corpus heterogeneity. For instance, we studied the paradigmatic sets {DANS, EN} and {MAINT, MOULT, BEAUCOUP}: we observed how the frequency of one item of the paradigm evolves compared to the frequency of the whole set. If a text, or a text type, uses the paradigm much more or much less than other texts or text types, this has no effect on the balance of items within the paradigm. In other words, for each feature, its paradigmatic set defines the active corpus (Pincemin et al., 2022) for the textometric processing. Therefore, at this stage we can use the whole (heterogeneous) corpus; it is not necessary to restrict the analysis to a homogeneous subcorpus.

We defined a dozen grammatical paradigms (Guillot-Barbance et al., 2025 and forthcoming) that could be captured with CQL expressions (Evert & Hardie, 2011). The CQL language supports advanced queries that can be tailored with a high degree of precision. For instance, we

¹ CQL query for grammatical words (Presto tagset): [frnpos="C.|D.|P.|R.|S.*|Xi" & frlemma!=".*ment"%c]

² CQL query for lexical words (Presto tagset): [frnpos="N.|Vv.|G.|I" | (frnpos="Rg" & frlemma!=".*ment"%c)]

³ A first and relative threshold was set to about 10 occurrences in 1 million tokens, that is, lower than 20 in Middle French, lower than 30 in Pre-Classical French, lower than 10 in Classical French; and a second (tougher) and absolute threshold was set to 10 occurrences in total in the pole.

could develop a query that covers and distinguishes positive and negative uses of the lemmas AUCUN, NUL, PERSONNE, QUELQU- and RIEN.⁴

2.3. The moving border implementation and the need for robustness toward major linguistic changes

The moving border is a simple case of Salem’s chronological characteristic elements (1988): the specificity is not computed for all possible time periods in the corpus (*two* variable borders, computational complexity of $O(n^2)$), but only for all possible positions for one border that splits the corpus into two parts (*only one* variable border, $O(n)$): one part before the border and one after. The result can be visualized with the plot of the specificity score of the contrast between the two parts as a function of the position of the boundary along time steps. The maximum of the specificity indicates the position of the border that best enhances the contrast between a “before” and an “after”. In summary, this method combines many advantages: it relies directly on the data without introducing external assumptions (non-parametric model); the Specificity measure provides a highly interpretable framework (Pincemin 2024); the calculation is less resource-intensive than computing full chronological specificities; and the resulting boundary is stable and reliable (the statistical approach accounts for all occurrences, unlike qualitative approaches based on first or last attestation).

The specificity computing iteration we needed is available as the TXM utility *r/Specificities File* that was implemented by Serge Heiden in the context of (Dégeilh et al., 2021). We applied it with a ten-year time step, that is, we tested every border position from 1310, 1320, 1330... to ...1650, 1660, for a total of 36 possible positions.

However, the linguistic change from one language state to another is often substantial, since the occurrences’ distribution is extremely far from random. In the 12 cases we studied, 9 could not be computed, because the corresponding probabilities were too low (lower than 10^{-300}), at least in the portion in which we had to look for the extremum. This problem was resolved by computing test-values instead of specificities (Lebart et al., 2019: 315). Test-values are another notation for the same (hypergeometric) probability, so that specificities and test-values reach their extrema for the same parameter values. The test-value calculation is much simpler and much more robust; it can be implemented in a simple spreadsheet.⁵ Test-values cannot be used for a feature with a total frequency F lower than about 30, but for these low frequencies the specificity computation is available; in our cases, all F frequencies were higher than a thousand. With this solution, we obtained complete results for all cases and could draw every plot.

2.4. Coupling moving border results with complementary textometric views

The moving border algorithm indicates where to place the boundary between two periods. The resulting plot looks like a temporal profile of the word’s usage over time, although it is not: for

⁴ Query (with the ‘longest’ strategy): `(([word="ne|n"%"c][word!="ne|n"%"c]{1,3})? [(frlemma="PERSONNE" & frnpos!="Nc") | frlemma="NUL|AUCUN|RIEN" | word="null?es?"%"c] | [(frlemma="PERSONNE" & frnpos!="Nc") | frlemma="NUL|AUCUN|RIEN" | word="null?es?"%"c] [word="ne|n"%"c]?) | [frlemma="QUELQU.*"]`

⁵ The formula we used in our French Calc spreadsheet, in which columns A to D contain f , F , t , T respectively, is (line 2): `= (A2-B2*C2/D2) / RACINE (B2* (D2-B2) *C2* (1-C2/D2) / (D2* (D2-1) ) )`. The RACINE() function is the square root function. Supplementary materials are available in the HAL repository for Guillot-Barbance et al. (2025): <https://hal.science/halshs-05539457>

each date the curve does not show what happens *locally* at that time, it reflects the *global* contrast between the whole period before and the whole period after. This is why the moving border view does not replace the existing textometric features that provide temporal profiles: these complementary views benefit from being considered together. In TXM software, both the progression graph (that plots the cumulative frequency along the corpus) and the specificity bar chart (on a chronological partition, for instance on decades) can thus complement the moving border view.

3. Results

3.1. Basic case

The replacement of the Middle French word *moult* with *beaucoup* provides a clear illustration of the moving border results (Figure 1). Time is on the x-axis of all four graphs. In the moving border graph (bottom right), we look for the maximum value (in absolute terms) of the specificity (or test-value): the corresponding decade, here 1530, is the time for which the contrast is the highest between the period before and the period after. We can then look at the specificity bar chart (above, top right), which confirms that, in comparison with *beaucoup* from 1300 to 1669, *moult* was overused before 1530, and underused after. The progression chart (top left) is crystal clear: *moult* was commonly used until the 1530s, when it was abruptly abandoned (a flat curve means no more occurrences). However, we also note that *beaucoup* appeared a few decades earlier, so that, without the moving border output, we could have hesitated about the precise time for the border.

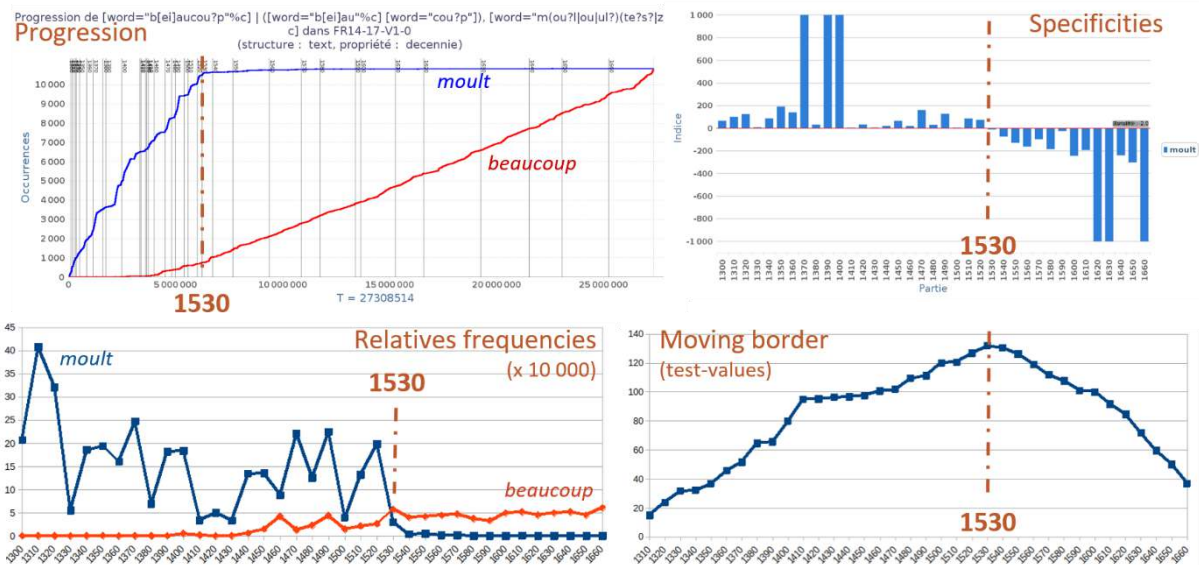


Figure 1. Analysis for the lemma *moult* in the paradigm {*beaucoup*, *moult*}

For discussion, we also include (Figure 1, bottom left) a temporal profile of the relative frequency in every decade. This measure is more erratic than the others: it overestimates the presence in decades with little data and underestimates it in decades with more data. The specificity measure was designed to be safer, and the progression view is closer to the original data, so ultimately, we do not need the relative frequency graph for our analyses.

3.2. More complex patterns

Nevertheless, the moving border graph interpretation may not be as clear: it may show a long flat extremum (Figure 3i), or several extrema. The *chaque* vs *chacun* analysis (Figure 2) illustrates this second case: the moving border graph (bottom right) shows two minima, one in 1580 and the other in 1630. The temporal profile views help us understand what happens at these times. The specificity bar graph (above, top right) displays that *chaque* really takes hold starting in 1630, but the 1580s already show, albeit more sporadically, a higher frequency of *chaque* compared to *chacun* when looking at the corpus average.

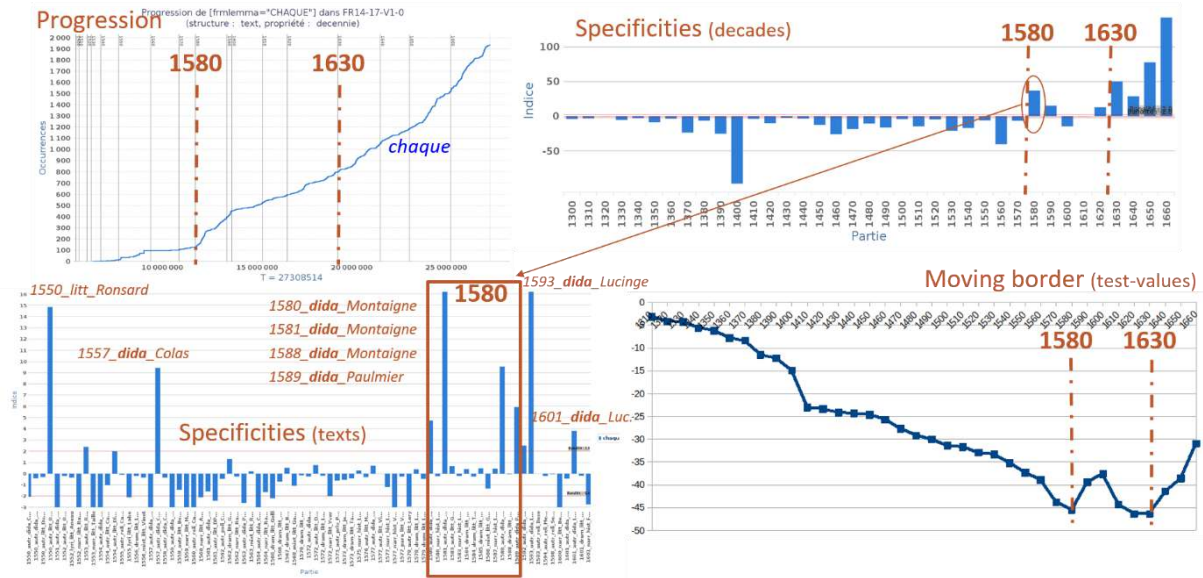


Figure 2. Analysis for the lemma *chaque* in the paradigm {*chacun*, *chaque*}

In the case of such isolated peaks, we verify whether this is due to a specific instance or whether it appears in multiple texts; we recalculate the specificities at the text-level detail (bottom left). We conclude that the rise of *chaque* in the 1580s is indeed due to multiple texts, and we note that almost all of the precursor texts are didactic treatises and essays. Thus, *chaque* begins to outpace *chacun* in the didactic field as early as 1580, and this trend becomes more widespread starting in 1630.

With regard to the progression curve, we note that the progression alone would not have allowed us to see the change occurring in 1630 as clearly; all the views do complement one another.

3.3. Overview and Summary

We applied the moving border model to 12 cases; Figure 3 gives an overview of the kind of curves we obtained. Most of them (3a-3g) exhibit a rather clear extremum, whereas in some cases (3h-3j), the extremum forms a plateau, with the result that the border cannot be clear-cut, as it rather manifests as a thick and fuzzy transition. Thus, the moving border may not automatically output a clear border for any given linguistic feature, yet it still provides new insight about the feature's evolution and its sensitivity to the periodization at stake.

The linguistic interpretation of the textometric outputs was communicated at SIDF (Guillot-Barbance et al., 2025) and will be developed in (Guillot-Barbance et al., forthcoming). On the whole, the dating proposed by the moving border model is consistent with the current state of

THE MOVING BORDER MODEL

the art in linguistic studies (Marchello-Nizia et al., 2020). Furthermore, each of the results is supported by our data, with which it can be clearly correlated and understood.

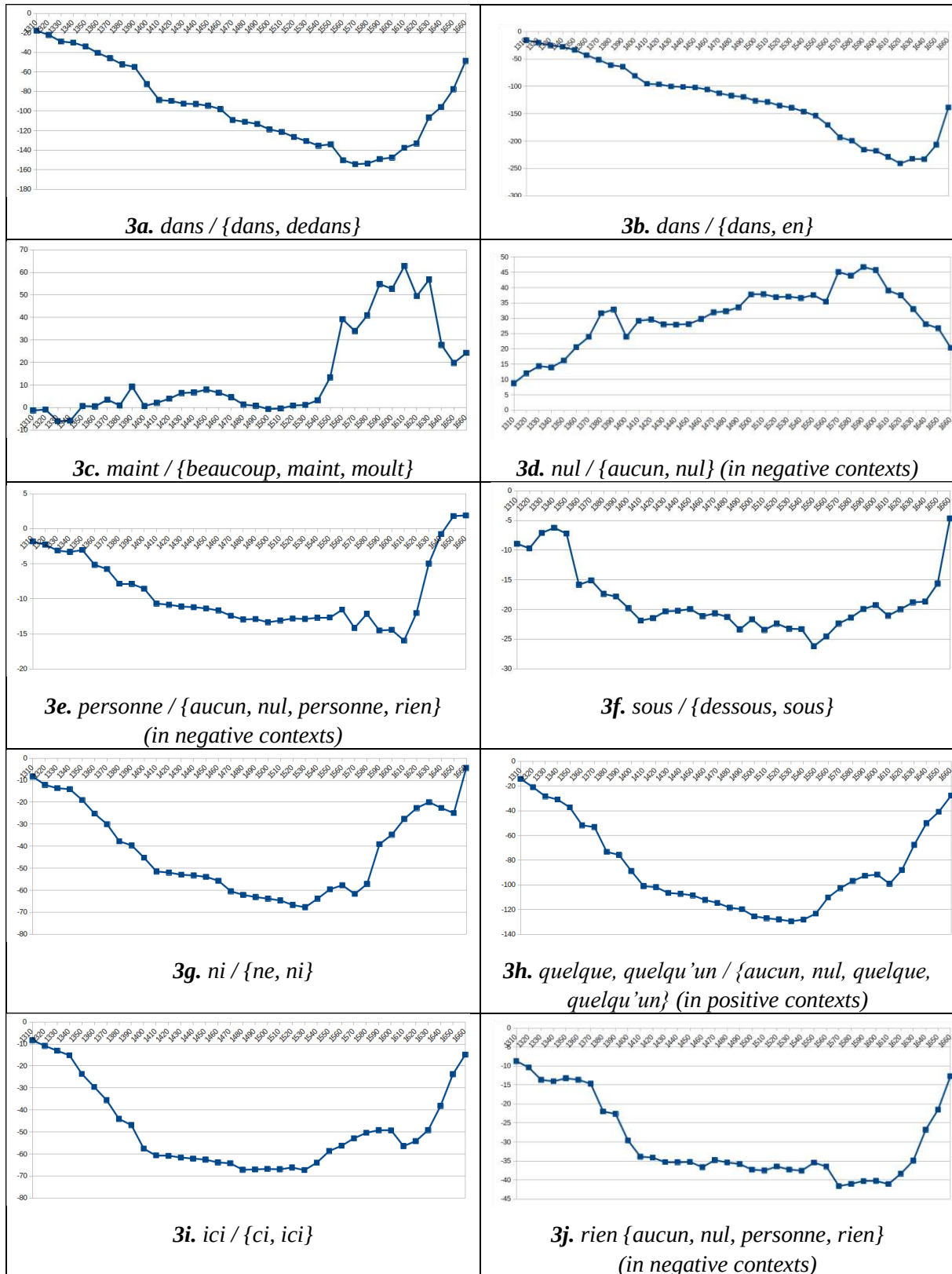


Figure 3. Moving border plots for the 10 other features we studied

4. Discussion

4.1. Choices in corpus modeling

The moving border model splits the corpus into two parts: a “before” and an “after”. This is less powerful than the adjacent characteristic elements model, which sets two borders and cuts the corpus into three parts: it finds the entire period with its beginning and its end. However, for linguistic change patterns, it seems relevant to focus on individual borders rather than on periods, because each border may be characterized by specific evolutions.

Another question that arises concerns the reference period to be considered. In our experiment, we always computed the contrast on the whole extent of the corpus, from 1300 to 1669. Should we have focused on more targeted subcorpora, for instance 1300-1619 for a search for the beginning of the Pre-Classical French period, and 1550-1669 for a search for its end? We conducted some tests that seem to indicate that this is not relevant in our corpus: considering the entire corpus provides a broader perspective and helps put certain variations into context. However, the question would likely need to be reconsidered for a corpus spanning a longer period.

4.2. Complementary strengths and weaknesses of textometric chronological visualizations

The three textometric diachronic views proved to complement each other clearly.

The progression view is not intuitive because of its cumulative nature; but once one knows how to read it, it is a key tool because it directly and accurately reflects what is happening at the data level. However, the main piece of information is given by the curve’s slope and its variations, which are not always easily visible on the graph (see point 1630 in Figure 2 for instance).

The specificity bar chart also provides a visualization for the temporal profile of a linguistic feature. Involving probabilities, the specificities are less close to the data but they are effective at highlighting times when there are significant variations. What might be tricky in this visualization is that the scores have a relative meaning, not an absolute one. Scores do not express the presence of the feature, scores actually correspond to the deviation from the mean: peaks do not denote high frequency but higher frequency (which may not be particularly high), and troughs do not necessarily denote absence, they just mean the feature is less present. Therefore, a specificity bar chart cannot be read in isolation either, complementary information about rough frequencies is needed.

Compared to these *analytic views of the current trends* of a feature over a breakdown of the 37 decades, the moving border provides a *synthetic view about the overall contrast* in the corpus depending on the time chosen for splitting it. It looks like a profile, but it is not a temporal profile for the feature occurrences; it shows the evolution of the contrast between two periods. As such, it is effective for delineating a border based on a clear and data-based criterion.

4.3. What linguistically defines a border: global contrast between periods, or local recorded use?

In all 12 of our cases, the connection to our data and the actual frequencies observed therein are clear: the result is justified in relation to the corpus and reflects concrete observations within the limits of our corpus. The moving border model provides a stable and often precise data-based dating according to the explicit criterion of contrast between periods.

This *global* contrastive criterion is somehow different from the traditional criterion that is *local* and based on recorded use (the *time spot* when a word first occurs, or last occurs, or becomes the most frequent). As such, the moving border tends to suggest slightly later start dates or earlier end dates than traditional studies. Interestingly, because the moving border computing is based on the overall evolution of the paradigm's equilibrium, the moving border's dates are more stable considering that the corpus is never exhaustive. Our probabilistic approach introduces another kind of delineation, with another linguistic meaning and another relevance that require further linguistic reflection, possibly in relation to the S-curve word evolution profile (Kroch, 1989).

5. Conclusion

The moving border algorithm rounds out the textometric toolkit by addressing the specific question of where to draw a boundary within a continuum. Derived from Salem's adjacent characteristic elements (1988), the moving border integrates a Fisher exact test and thus implements a very clear and reliable measure (Pincemin, 2024). It does not work as an automatic and independent tool, but it is fully in line with the textometric approach, which involves a semi-automated analysis that combines several complementary tools.

In the case of linguistic periodization, we were led to question the choice of the active corpus: linguistic features are not observed in relation to the language as a whole, but within paradigms in which evolution takes place. When targeted in this way, the observations are less affected by textual variations and can be applied to a larger corpus, thereby benefiting from a greater volume of data.

Moreover, the study of linguistic change involves describing major shifts in language, with the reconfiguration of systems articulating high-frequency words, and even the emergence or disappearance of terms. The statistical measure of specificity, based on deviation from randomness, has to account for deviations so large that they often pose challenges. The calculation of test-values provides a pragmatic solution that is simple to implement while remaining within the hypergeometric model.

The methodological approaches described here can be applied to other boundary-definition issues. Developed using open-source software, they are available to the academic community.

References

- Ayres-Bennett W. and Caron Ph. (2016). Periodization, Translation, Prescription and the Emergence of Classical French. *Transactions of the Philological Society*, 114/3: 339-390.
- Caron Ph. (in press). Le français pré-classique : un cas d'école pour la périodisation. In Skupien Dekens C., Ortner S., and Schmerbeck M., eds, *Le français préclassique, spécificités et tendances*. Garnier.
- Dégeilh F., Lecouvey G., Hirst W., Heiden S., Pincemin B., et al. (2021). Changes over 10 years in the retelling of the flashbulb memories of the attack of September 11, 2001. *Memory*, 29 (8): 1006-1016.
- Diwersy S., Falaise A., Lay M.-H., and Souvay G. (2017). Ressources et méthodes pour l'analyse diachronique. *Langages*, 206: 21-44.
- Diwersy S. and Luxardo G. (2016). Mettre en évidence le temps lexical dans un corpus de grandes dimensions : l'exemple des débats du Parlement européen. In Mayaffre, D., Poudat, C., Vanni, L., Magri, V., Follette, P., eds, *Statistical Analysis of Textual Data, Proc. JADT 2016*, pp. 169-180.
- Evert S. and Hardie A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proc. of the Corpus Linguistics 2011 conference*, Univ. of Birmingham

- Feltgen Q. (2017). *Statistical physics of language evolution: the grammaticalization phenomenon*. PhD thesis, PSL Research University.
- Guaresi M., Mayaffre D., and Vanni L. (2021). Entre rupture et continuité, le discours du PCF (1920-2020). *Histoire & Mesure*, 36(2): 125-162.
- Gries S. Th. and Hilpert M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1): 59-81.
- Guillot-Barbance C., Lavrentiev A., and Pincemin B. (2025). Périodisation du français préclassique : explorations quantitatives. *Colloque de la Soc. Internationale de Diachronie du Français*, Grenoble.
- Guillot-Barbance C., Lavrentiev A., and Pincemin B. (forthcoming). Périodisation du français préclassique : explorations quantitatives. Submitted to *Le Français préclassique*.
- Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro R. et al., ed., *Proc. PACLIC 2010*. Waseda, pp. 389-398.
- Kroch A. S. (1989). Reflexes of grammar in patterns of language change. *Language variation and change*, 1: 199-244.
- Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1: 127-165.
- Lavrentiev A., Pincemin B., and Guillot-Barbance C. (forthcoming). Le corpus FR1300-1669 : préparation, traitement et analyse de sources pour la périodisation du français préclassique. Submitted to *Corpus*, 28.
- Lebart L., Pincemin P., Poudat C. (2019). *Analyse des données textuelles*. Presses de l'univ. du Québec.
- Lebart L., Salem A., and Berry E. (1998). *Exploring Textual Data*. Kluwer Academic.
- Marchello-Nizia C. and Combettes B. (2020). Périodisation. In Marchello-Nizia C., Combettes B., Prévost S. and Scheer T. eds, *Grande grammaire historique du français*, De Gruyter Mouton, 54-61.
- Marchello-Nizia C., Combettes B., Prévost S., and Scheer T., editors (2020). *Grande Grammaire Historique du Français*. De Gruyter Mouton.
- Miao J. and Salem A. (2021). Des textes en mouvement... Analyse textométrique des rapports d'ouverture présentés aux congrès du Parti communiste chinois (1982-2017). *Histoire & Mesure*, 36(2): 91-124.
- Pincemin B. (2024). Specificities and other applications of the Fisher exact test to textual data: What's the matter with lexical frequencies? In Dister A. & Longrée D., eds, *JADT 2024 : Mots comptés, textes déchiffrés*. Presses univ. Louvain, 703-712, corrected ed.: <https://hal.science/hal-04874716v2>.
- Pincemin B., Heiden S., Mazuet F. (2022). The Textometric Concept of Active Corpus. Illustration by an Analysis Scenario based on Annotation then Projection. In Misuraca M., Scepi G., Spano M., eds, *Proc. of the 16th International Conf on Statistical Analysis of Textual Data*, Vadistat Press, 691-698.
- Salem A. (1988). Approches du temps lexical. Statistique textuelle et séries chronologiques. *Mots*, 17: 105-143.
- Salem A. (2021). Le temps lexical. Un bilan méthodologique sur l'analyse des séries textuelles chronologiques. *Histoire & Mesure*, 36(2): 21-56.
- Schmid H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- de Sousa S. (2021). Le discours de Fidel Castro : périodisation et évolution (1959-2008). *Histoire & mesure*, 36(2): 57-90.
- Trevisani M. and Tuzzi A. (2018). The recent history of statistics: Comparing temporal patterns of word clusters. In Tuzzi, A., ed, *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*. Springer, pp. 105-129.

Validation of Simultaneous Additive Trees in Textual Data Visualization

Ludovic Lebart

CNRS (R) – ludovic@lebart.org

Abstract

Simultaneous (or augmented) additive trees are used to visualize lexical similarities and thematic relationships within large corpora. They allow words and texts to be represented on the same graphical representation, along with external (or supplementary) categorical variables describing these texts. In poetry, for example, the most frequent categories are collections, periods of creation, or major themes recognized by experts. However, the validation of observed patterns remains an insufficiently addressed problem. Interpretations often rely on visual inspection and heuristic criteria, which can limit their analytical reliability. The categorical variables mentioned above can be replicated by bootstrapping (a bootstrap replication being a random sampling with replacement of their texts). Replications of size k (e.g., $k = 30$, a value generally acceptable in a multidimensional context that does not involve numerical estimates) allow the calculation of confidence ellipses (or convex confidence hulls) for the position of the categories. The methodology is illustrated by an application to a classic literary corpus: the 154 sonnets of William Shakespeare (with their classification into eight themes, recognized by many experts). This corpus will serve as a reference for defining and comparing certain characteristics of the proposed validation procedure (to complement the classical example, a more contemporary corpus would certainly be helpful for both the impact and dissemination of the method, but the presented visualizations would exceed the allowed space). A complete, transparent set of Python programs (with a user-friendly Tkinter interface) will be freely available from the websites: www.dtmvic.com and <https://ludovic.github.io/DTM-VIC>.

Keywords: Simultaneous Additive Trees, Bootstrap validation, Visualization. Poetry, Shakespeare Sonnets.

1. Introduction

Visualizing textual data here involves producing graphical representations that highlight proximities between textual units (words, graphic forms, lemmas) and proximities between texts within the same corpus (speeches, novels, poems, authors, etc.). Correspondence analysis (CA) is a widely used tool, particularly because of its ability to simultaneously represent textual units and texts. Another descriptive tool is additive trees (AT) (or phylogenetic trees), which attempts to represent calculated distances in spaces of more than two dimensions. The simultaneous representation of texts and units, and the positioning of additional variables, are also possible with AT, although they do not offer the optimal performance of CA. In this work, we show that in both AT and CA, non-parametric confidence zones for point positions can be plotted. This technique will complement CA, not replace it.

Correspondence analysis (CA) remains an essential step for initial work with lexical tables, but it does not guarantee a comprehensive overview. In contrast, additive trees, with their interpretation rules, allow for the reconstruction of distances in a plane (a tree is a planar graph) within spaces of dimensions far exceeding two. However, they lacked the simultaneous representation of rows and columns (i.e., words and texts) of the basic lexical table. A pragmatic method for the simultaneous representation and positioning of additional variables was proposed at JADT2024 in Brussels (Lebart, 2024). It is summarized in Section 2. A non-

parametric statistical validation procedure based on the bootstrap technique is presented in Section 3. Section 4 then describes the text corpus (Shakespeare Sonnets) used as an illustration (other examples of corpora involving more contemporary texts will accompany the software). Section 5 presents the application itself. Finally, Section 6 provides some pieces of information about the self-contained Python program chain.

2. Correspondence Analysis (CA) and Additive Trees (AT)

The calculation of additive trees, originally proposed by Buneman (1971), was improved by Sattath and Tversky (1977). The work of Barthélemy and Guénoche (1988) and Luong (1988) promoted the use of these methods (under the name of “*Analyse Arborée*”) in the field of text analysis. Additive trees were made easily computable by Saitou and Nei (1987) and subsequently became established as a synthesis of principal component analysis and clustering techniques (hierarchical classifications, k-means partitions, etc.). Bryant (2005), and then Huson and Bryant (2006), later justified the method and made corresponding software available. Theoretical justifications for the algorithm's effectiveness were presented by Mihaescu et al. (2009).

Let us briefly recall that the concept of hierarchy at the heart of hierarchical clustering amounts to approximating the initial distances by an ultrametric distance, which, in addition to the classical axioms of distances, satisfies the following inequality for any triplet (x, y, z) :

$$d(x, y) \leq \max(d(x, z), d(y, z))$$

Additive trees require, for any quadruplet (x, y, z, t) , that the following more complex and less intuitive, but less demanding, inequality be satisfied:

$$d(x, y) + d(z, t) \leq \max(\{d(x, z) + d(y, t)\}, \{d(x, t) + d(y, z)\})$$

A tree can then be drawn with the objects to be classified as terminal elements (or “leaves”). More flexible than the minimum spanning tree, which depends on $n-1$ parameters, the additive tree involves $2n-3$ parameters. It is therefore necessary to find an approximation of the initial distances that satisfies these conditions, which is made possible by the mentioned work of Saitou and Nei. The important result is that the distances between nodes (texts) of the graph (length of the shortest path on the graph) are approximations of the actual distances throughout the whole space, approximations often better than those provided by a first AC plane.

2.1 On the options for drawing additive trees

It remains to draw these trees. For a comprehensive review of general graph visualizations, see Di Battista *et al.* (1999), and more specifically, regarding methods using force-directed drawing algorithms, the article by Kobourov (2013), which analyzes more than 60 publications corresponding to several dozen algorithms. Tutte's graph drawing algorithm (1963) was one of the first drawing methods based on algorithms of this type. Then, the methods proposed by Eades (1984) and the Fruchterman and Reingold algorithm (1991) both rely on repulsive forces between all the nodes of the graph, but also on attractive forces between adjacent nodes (the edges are considered springs, and the goal is to find a balance between all the tensions, hence the name force-directed drawings).

Alternatively, the forces between nodes can be calculated based on concepts from graph theory. The distances between nodes are then the lengths of the shortest paths connecting them. For additive trees, these distances are precisely an approximation of the original distances (chi-square distances calculated from the original lexical table). The algorithm of Kamada and Kawai (1989) uses these "spring forces" proportional to these distances calculated from the graph. This plotting algorithm is therefore the most compatible with the properties of additive trees, and thus with the basic lexical distances. Experimentally, the good compatibility of these representations with the principal planes derived from the CA of the original lexical table is also observed. However, a validation technique remained necessary to confirm that observation.

The ability of correspondence analysis to describe undirected graphs of the planar grid or geographical map type from their associated matrices has been highlighted by Lebart *et al.* (1998). One might have thought of using correspondence analysis again to obtain a plot of the additive tree. However, trees (connected graphs without cycles) are poorly visualized by this method, and vice versa, as noted by Sattath and Tversky (1977): *"It is interesting to note that tree and spatial models are opposing in the sense that very simple configurations of one model are incompatible with the other model. For example, a square grid in the plane cannot be adequately described by an additive tree."*

Finally, we proceed with the representations provided by the Kamada-Kawai method, the most suitable for additive trees, enrich them with a simultaneous representation of the rows and columns of the lexical table, and complement them with a bootstrap validation procedure.

2.2 Simultaneous Representation and Characteristic Words

CA can easily be presented directly as the search for the best possible simultaneous representation of the proximities between rows and columns of a contingency table. One could indeed look for a simultaneous positioning of texts and words on an axis (to begin with) in order to obtain a doubly centroidal relationship: words at the centroid of the texts, and texts at the centroid of the words (the weights being respectively the lexical profiles in rows and columns calculated on the basic lexical table). This double relationship is impossible, because centroid computation is contractive: words must be within the interval covered by the texts and, simultaneously, the texts must be within the interval covered by the words. For the relationship to be possible, these centroids must be expanded (coefficient $\beta_\alpha > 1$, for axis α).

The optimal solution corresponds to values of β_α closest to 1, which gives us the positions of the words and texts on the first axes of the correspondence analysis. These doubly barycentric relations are known as "transition relationships".

Therefore, we cannot hope to find an optimal dual simultaneous representation, but a single simultaneous representation is sufficient (words as centroids of the texts, which are the terminal elements of the tree). We will enrich this simultaneous representation by also incorporating the notion of characteristic words (or specificities).

The simultaneous representation procedure includes the following steps: (columns and rows play similar roles and can be interchanged).

- a) Preliminary CA of the lexical table.
- b) Choice of the dimension nx of the space deemed significant (generally by bootstrapping) (12 axes, for example). Distances will be calculated from the first nx

principal axes of the CA (in our illustrative examples, $nx = 12$). This allows for the regularization of the initial distances, a well-known procedure in discriminant analysis and in certain Deep Learning procedures.

- c) Calculation of the additive tree (Neighbors-Joining method) on the distance matrix between the coordinates of the columns (texts) on the first nx axes.
- d) Plotting the tree (Kamada-Kawai procedure)
- e) The rows (words - forms, lemmas) are positioned barycentrically based on the coordinates of the column points (texts) (terminal elements of the tree) issued from Kamada-Kawai procedure and the textual profile of the rows (words, tokens, lemmas).

To limit the number of words on the graphs, the characteristic rows/words (with a fixed probabilistic threshold) are calculated directly from the lexical table for each column-text using test values (for test values, see, for example, Lebart *et al.*; 1998, 2019).

In practice, new edges (different in color and thickness from those of the edges of the additive tree) are drawn, joining each column point (text) on the graph to its characteristic rows (words).

The proposed methodology is illustrated by its application to a classic literary corpus: the 154 sonnets of William Shakespeare (with their classification into eight themes, recognized by experts). This corpus will serve as a reference for defining and comparing certain characteristics of the proposed validation procedure.

3. Validation and bootstrap

Computer intensive techniques allow us to go far beyond the criterion of “interpretability of the results” that was frequently used during the first phases of the upsurge of data analytic methods. To compute the precision of estimates, the classical analytical approach is both unrealistic and analytically complex. The bootstrap (see: Efron and Tibshirani, 1993), on the contrary, makes almost no assumption about the underlying distributions, and gives the possibility to master every statistical computation for each sample replication and therefore to deal with parameters computed through the most complex algorithms.

The nonparametric bootstrap consists of drawing with replacement k samples of size n out of n statistical units. Then, parameter estimates such as means, variances, coordinates on a specific plane (in our case) are computed on the k new obtained samples. Empirical evidence suggests that 30 is an acceptable value for k in the context of Principal Axes Methods. In this particular context, we have to assess only the stability of patterns, and not to estimate a unique numerical parameters. In any case, it is straightforward to test several values of k . At this stage k samples (the replicates) are drawn from a new theoretical population defined by the empirical distribution of the original data set, leading to k estimates of the set of parameters of interest.

In our case, we will use the Partial Bootstrap (see Greenacre, 1984; Chateau and Lebart, 1996): We do not question the validity of the coordinates of the active statistical units (words and texts) but we wish to appreciate the stability of the supplementary categories (themes/topics in our example). One replication then consists of a drawing with replacement of the text units (the 154 sonnets in the case of the example of section 4 below). This leads to a vector of bootstrap weights, such as 0, 1, 2, 3... the sum of the weight being 154. A replicated category is then a weighted barycenter of the text units. The process is the same for CA and AT. The cloud of replicates for one category allows for the drawing of a confidence ellipse containing 95 % of

the replicates. Examples of such confidence areas are given below for both CA and AT (Fig. 3 and 4).

4. Application: Corpus and context

4.1 An Overview of the Content of Shakespeare's Sonnets

William Shakespeare's 154 sonnets deal with themes or topics such as love, friendship, the effects of time, beauty, betrayal, lust, and death. Three contiguous sets of sonnets are generally recognized as corresponding to three dominant themes:

Sonnets 1-17: (Procreation). These sonnets celebrate the beauty of a young man whom the poet urges to marry in order to perpetuate that beauty.

Sonnets 18-126: (Young Man). This sequence, the longest, concerns the same young man (not identified), the destructive effect of time, and the power of love, friendship, and poetry.

Sonnets 127-154: (Dark Lady). These sonnets are primarily addressed to a dark-haired woman.

The themes Young Man and Dark Lady could themselves contain five sub-themes assigned to the five new categories below (Absence, Storm, Rivalry, Death, Eternal Poetry).

Table 1. List of eight a priori categories (topics) with corresponding sonnet numbers

Category 1: Procreation 1 - 17
Category 2: Young Man 20-25, 33-38, 40-42, 46, 47, 49, 53-55, 59-60, 62-70, 75-77, 88-106, 108-112, 115-125,
Category 3: Dark Lady 127-136, 139, 140, 143-146, 153, 154
Category 4: Absence 26-32, 39, 43-45, 48, 50-52, 56-58, 61, 113-114
Category 5: Storm 141, 142, 147-152
Category 6: Rivalry 78-87
Category 7: Death 71-74
Category 8: Eternal poetry 18, 19, 81, 107

The sonnet divisions in Table 1 are inspired by the work of Alden (1913) and Paterson (2010).

4.2 Specificity of poetic texts

The statistical analysis of poetry and song collections is hampered by the specific nature of these texts. They may contain choruses, deliberate repetitions, stylistic devices, or versification constraints that challenge the statistical significance of lexical frequencies. A pragmatic solution is to work with "bags of words" (the presence or absence of words, tokens, lemmas) which constitute only a tiny fraction of the texts' artistic aspect, but nevertheless prove useful for studying, for example, a work within its temporal context. This is not an in-depth study or an elucidation of poetic texts, but simply an examination of lexical traces, which will ultimately prove to be of interest.

4.3 Inadequacy of Principal Axes

With the presence-absence coding of words in text units, frequencies are somewhat constrained. Word bags are created from a certain size of text units. These units will be the poems themselves

in the case of poetic works. In our case, for Shakespeare, the text units will be the 154 Sonnets. The maximum frequency of a word cannot exceed the number of text groups, since it is simply a matter of "presence-absence" coding within each text unit. Under these conditions, the dimensionality and sphericity of the word cloud rarely allow for good visualizations in a low-dimensional space. In the case of CA, the eigenvalues will be poorly separated, a source of instability for the principal axes. The two-dimensional representation of words and texts, an undeniable advantage of CA, becomes insufficient, and additive trees become indispensable.

5. Implementation and examples of visualization

We will characterize and summarize our approach with the help of four graphical displays.

5.1 Figure 1, Aggregated lexical table

For Figure 1, the eight themes are considered known and indisputable. They were used to construct the lexical table (Themes x words). Figure 1 then shows the simultaneous additive tree (SAT) linking the eight themes, which are terminal elements of the tree (its "leaves"), with the positioning of the words (which are centroids of the themes). The words were linked to the themes they most characterize (criterion: test values). As with the following figures, we are dealing with vocabulary and not lexical frequency: each sonnet is a bag of words. The themes play an active role in the construction of figure 1. This figure reminds us of the principle of the SAT.

5.2 Figure 2, Direct Additive Tree

The additive tree in Figure 2, on the other hand, is derived from the AT of the lexical table (Sonnets x words) (the 154 sonnets being identified from "S01" to "S154"). It completely ignores the themes. These themes are then located *a posteriori*. As in CA, they are centroids of the implied sonnets, these centroids being expanded to equalize the dispersions of the sonnets and of the themes (the famous β coefficients mentioned earlier in section 2.2).

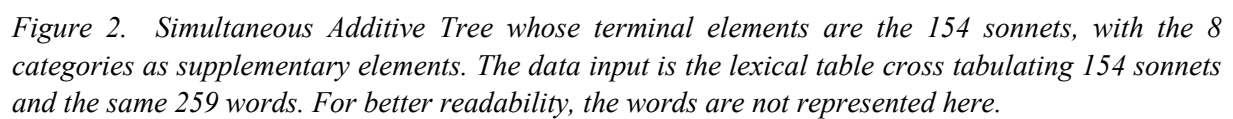
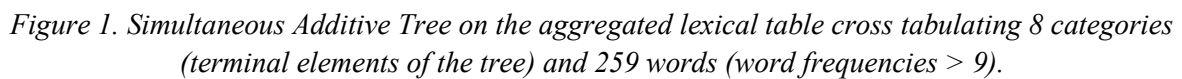
For example, we can see that the theme "Death" (at the top) is particularly far removed from the theme "Rivalry" (at the bottom right). Words could also be used to characterize the sonnets (with colors and a much larger figure format...).

5.3 Figure 3, Direct CA with confidence areas

Figure 3 shows the first principal plane of a CA of the same lexical table as Figure 2. The words are still omitted for readability, but the themes are shown as additional elements. Only two bootstrap confidence ellipses for the position of the themes are drawn (again for readability) for the two preceding themes (boxed), Death and Rivalry. We can see that the themes are not as far apart as in Figure 2, but also that their confidence ellipses have a fairly clear intersection, which casts doubt on the significance of their distinction.

5.4 Figure 4, Direct AT with confidence areas

Finally, Figure 4 shows us the same additive tree as in Figure 2, but now with bootstrap confidence ellipses for the two highlighted (and boxed) themes. It is clear that these themes are not only distinct but almost opposite, something the principal plane of the CA, limited to two dimensions, could not detect. One could argue that it is possible to observe such divergences by analyzing the sequence of principal axes beyond axis 2. However, practitioners know that this is difficult and tedious, as proximities in the (3,4) plane, for example, must be interpreted "all other things being equal" ("ceteris paribus") in the (1,2) plane...



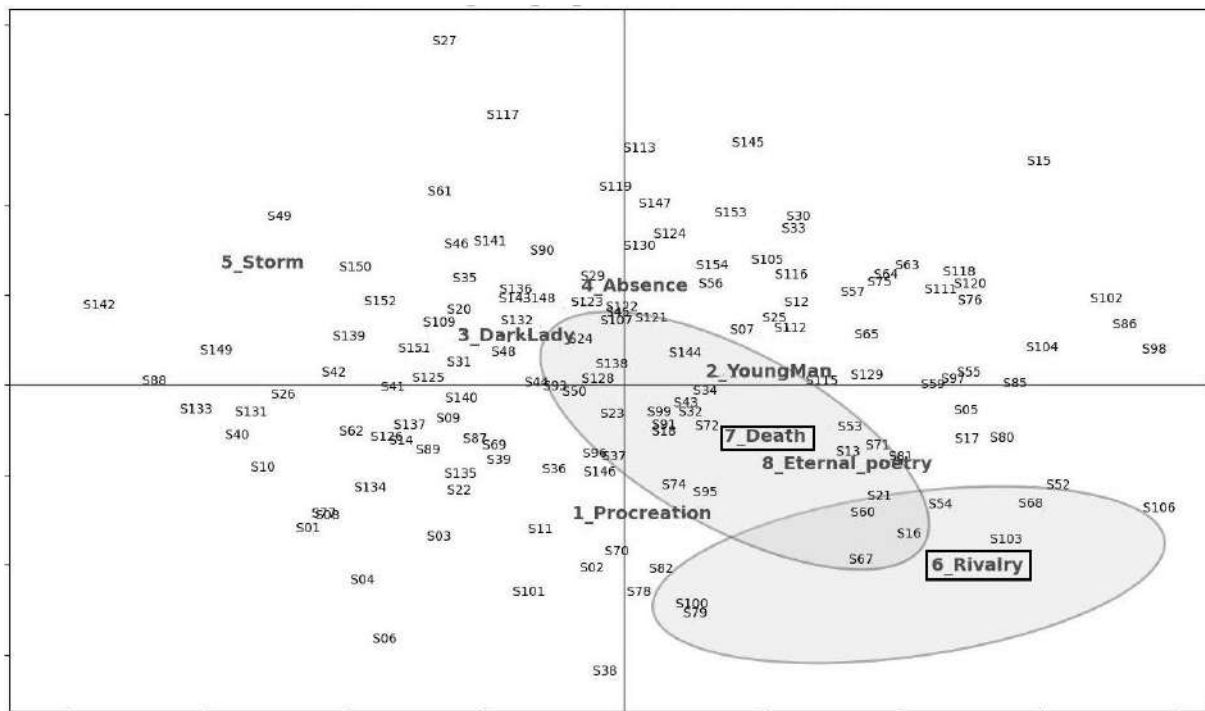


Figure 3. Plane (1, 2) from the CA of the same lexical table as in Figure 2. In this classical simultaneous representation, the 154 sonnets are identified by their traditional numbers in the sequence of sonnets. Two bootstrap confidence ellipses are drawn, for the two topics 7_Death and 6_Rivalry (squared identifiers). These ellipses are overlapping, inducing a disputable distinction between these two topics.

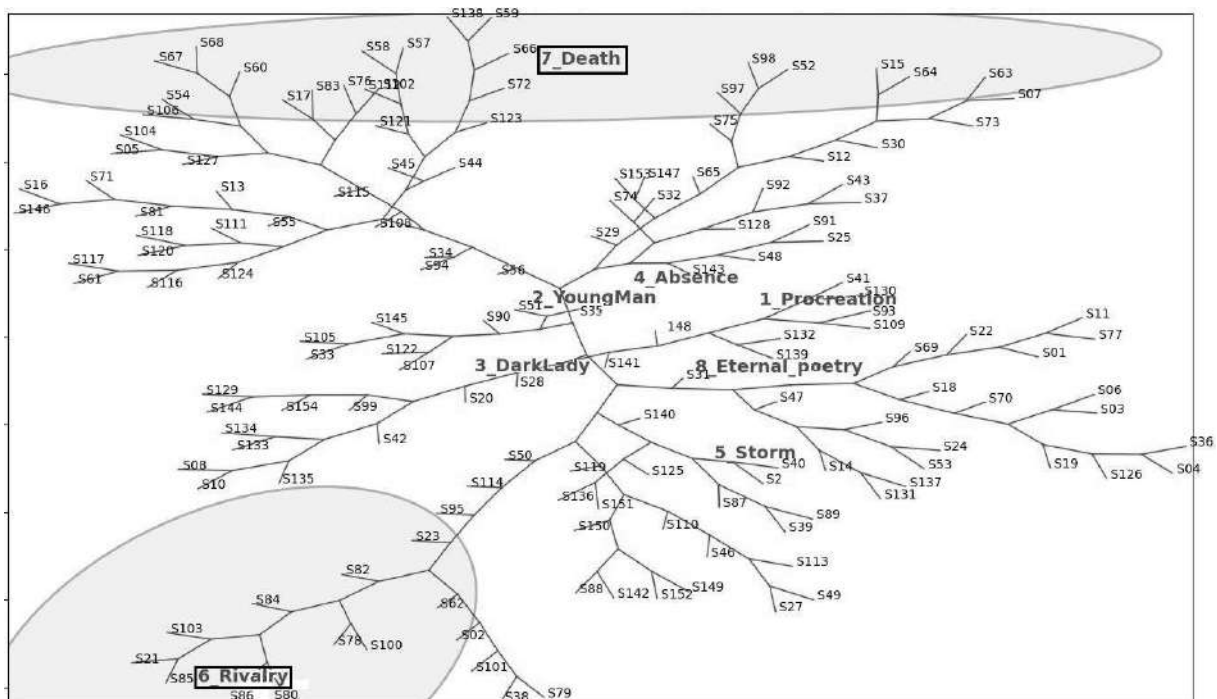


Figure 4. On the same tree as in figure 2, the confidence ellipses of the two topics Death and Rivalry are drawn, exhibiting now a marked difference between these topics, probably imputable to the axes of order > 2 .

6. Note about the Python Software

The Python code for the complete seven- or nine-step processing chain (from raw text to bootstrap confidence ellipses plotted on simultaneous visualizations of additive trees), will be freely available at both URL: www.dtmvic.com and <https://ludoriv.github.io/DTM-VIC>.

This chain uses the following standard Python modules: *Tkinter* (interface, graph framing), *os*, *Lib*, *numpy*, *pandas*, *matplotlib*, *igraph*, *importlib*, *textwrap*, *re* (regular expression), *sklearn*, and *scipy*.

Briefly describing the user interface (Table 1) allows us to sketch both the scope and the content of the software.

Table 1: Folders, files and parameters

The first section entitled: <i>Select folder</i> looks for the basic folder containing data files and modules.
A second section entitled: <i>Select files</i> specifies the three input files.
- " Select basic_text ": (raw input text : tokens, lemmas, etc.)
- " Select categories ": (XL file (.csv) file describing categories of texts)
- " Select identifiers ": (XL file (.csv) file describing the identifiers of categories)
A third section entitled: <i>Parameters</i> involves seven parameters and describes the various options.
" wordbag ": If wordbag = 1: Each raw text is replaced with a bag of words; if wordbag = 0: the analyses are performed on raw texts,
" anagreg ": If anagreg = 1: Poems/texts are aggregated according to the categories described in category file; if anagreg = 0: Direct analysis of all poems/texts, leading to validation procedures.
" freqmin ": freqmin: Minimum frequency for kept words.,
" nax ": nax: Maximum number of axes to keep in the initial correspondence analysis.,
" nvt ": nvt: Number of kept characteristic words (criterion: test values)
" Detail ": For anagreg = 0: Detail = 1.; detailed outputs (bootstrap replications on graphical displays as a check, various visualizations); Detail = 0: basic outputs.,
" seuil_vt ": seuil_vt: Significance threshold for the test-values (for selection of characteristic words).

7. Conclusion

Despite several similarities with artificial intelligence as regards some algebraic or algorithmic tools, data visualization methods occupy a particular place in a knowledge process. Visualization is neither a simple task to perform nor a decision to make. We submit our texts to understand and reflect. The approach is mainly unsupervised, a phase of work that Deep Learning will increasingly need, according to the predictions of the seminal text by LeCun et al. (2015). As in correspondence analysis with its simultaneous representations, the observable pattern of column points (texts, collections) tells us how these points are organized, and the presence of row points (words) and additional variables (themes, in that example) tells us why they are organized in this way: texts are similar because they often use the same words or deal with the same themes. It has been said that words and categories dress the skeleton of the additive tree. But this dressing would be merely a disguise in the absence of validation procedures. The non-parametric validation proposed here hopes to give these visualizations a scientific status.

References

- Alden, R. M. (1913). *Sonnets and a Lover's Complaint*. New York: Macmillan.
- Barthélémy J.-P. and Guénoche A. (1988). *Les arbres et les représentations de proximité*. Paris: Masson.
- Bryant D. (2005). On the uniqueness of the selection criterion in Neighbor-Joining. *Journal of Classification*, vol. (22), 1: 3-16.
- Buneman P. (1971). The recovery of trees from measurements of dissimilarity. In: Hodson F. R. D. Kendall G., and Tautu P., (Editors). *Mathematics in the archeological and historical sciences*. Edinburgh University Press, Edinburgh: 387-395.
- Chateau, F. and Lebart, L. (1996). Assessing sample variability in visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. In : A. Prats (Ed): *COMPSTAT96*, Physica Verlag, Heidelberg, 205-210.
- Di Battista, G. Eades, P., Tamassia R., and Tollis, I.G. (1999). *Graph Drawing: Algorithms for the Visualization of Graphs*, Englewood Cliffs: Prentice. Hall.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall: New York.
- Eades P. (1984). A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160.
- Fruchterman T. and Reingold E. (1991). Graph drawing by force-directed placement. *Softw. – Pract. Exp.*, 21(11):1129–1164.
- Greenacre, M. (1984): *Theory and Applications of Correspondence Analysis*. Academic Press, London.
- Huson D. H. and Bryant D. (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, 23 (2): 254 - 267. Software available from www.splitsree.org.
- Kamada T. and Kawai S. (1989). An algorithm for drawing general undirected graphs. *Inform. Process. Lett.*, 31:7–15.
- Kobourov, S. G. (2013). Force-Directed Drawing Algorithms. in: *Handbook on Graph Drawing and Visualization*. Chapman and Hall/CRC.
- Lebart L. (2007). Which bootstrap for principal axes methods? In: *Selected Contributions in Data Analysis and Classification*. P. Brito *et al.* (eds), Springer: 581-588.
- Lebart L. (2024). Des outils pour décrire certains corpus de poèmes et de chansons : Les arbres additifs simultanés. (2024). In: JADT 2024, "Mots comptés, Textes déchiffrés" Anne DISTER et Dominique LONGREE (eds) Vol. 2, PUL, (Presses Universitaires de Louvain) : 553- 562.
- Lebart L., Pincemin B. and Poudat C. (2019). *Analyse des Données Textuelles*. Presses de l'Université du Québec, Québec.
- Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*, Kluwer, Dordrecht.
- LeCun Y, Bengio Y. and Hinton G. (2015). Deep learning. *Nature*. May 28;521(7553):436-44.
- Luong X. (1988). *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse pour le doctorat ès sciences. Université Paris V.
- Mihaescu R., Levy D. and Pachter L. (2009). Why Neighbor-Joining works? *Algorithmica*, vol. (54): 1-24.
- Paterson D. (2010). *Reading Shakespeare Sonnets*. Faber & Faber Ltd. London.
- Saitou N. and Nei M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, vol. (4), 4: 406-425.
- Sattath S. and Tversky A. (1977). Additive similarity trees. *Psychometrika*, vol. (42), 3: 319-345.
- Shakespeare, W. (1901). *Poems and sonnets: Booklover's Edition*. Ed. The University Society and Israel Gollancz. New York: University Society Press. Shakespeare Online. Dec. 2017.
- Tutte, W. T. (1963). How to draw a graph. *Proc. London Math. Society*, 13(52):743–768.

Pulling strings: scalable symbolic methods for Multi-Word Units detection in text corpora

Antonio Balvet¹

¹ UMR 8163 – STL “Savoirs Textes Langage” – F-59000 , Lille University, France

Abstract

Automatically detecting multi-word units (MWUs) in text corpora remains a fundamental challenge in corpus linguistics and textometry. Traditional approaches, based on n-gram enumeration for lexical bundles Biber et al. (1999) or repeated segments Lebart and Salem (1994), are inherently limited to contiguous sequences. Approaches such as quasi-segments detection Bécue (1993) address discontinuity but remain computationally constrained. We present MWU-MINER, a scalable toolkit for extracting both contiguous and discontinuous recurring patterns from corpora that are too small for classical statistical procedures while remaining much too large for manual extraction. Our approach assumes that if two corpus segments are similar, they exhibit a certain level of repetition, and their common subsequence likely constitutes a MWU. However, a naive pairwise comparison yields $O(n^2)$ complexity, which quickly makes the extraction of Longest Common Subsequences (LCS) intractable. Our key contribution is a two-stage pipeline that circumvents this bottleneck. First, Locality-Sensitive Hashing (LSH) with minHash signatures clusters similar segments in near-linear time, dramatically reducing the search space. Second, within each cluster, LCS extraction identifies shared patterns, including those with variable slots (e.g. “chain acyl” / “chain fatty acyl” → “chain _ acyl”). The toolkit operates at the token or character level. MWU-MINER is designed for a “human-in-the-loop” pipeline: a final filtering phase, based on a range of metrics (Pointwise Mutual Information, log-likelihood, Levenshtein distance, among others) ranks candidates for further evaluation by a linguist. The extracted patterns, with their frequency and association scores, thus provide empirical material for further qualitative analysis, bridging quantitative corpus exploration and fine-grained linguistic description. Beyond individual MWU extraction, both modes yield a complementary output: pair mining produces a ranked TSV (Tab-Separated Values) file of annotated segment pairs, while DAG (Directed Acyclic Graph) mining produces a JSON (JavaScript Object Notation) summary of per-cluster recurring patterns.

Keywords: multi-word units, phraseology, minHash, locality-sensitive hashing, LCS, corpus linguistics, pattern extraction.

1. Introduction

Automatically detecting Multi-Word Units (MWUs), from fixed collocations to semi-frozen syntactic patterns, remains a central task in corpus linguistics applications such as terminology, lexicography, translation, Natural Language Processing, and associated domains such as textometry or forensic linguistics.¹ Traditional approaches based on n-gram enumeration (Biber et al., 1999; Lebart and Salem, 1994) are inherently limited to contiguous sequences, while approaches addressing discontinuity, such as quasi-segments (Bécue, 1993), face severe computational constraints that restrict their applicability to small corpora. Repeated segments detection methods, with or without gaps, can hint at discontinuous sequences in specialized corpora, such as:²

¹ Plagiarism detection, author attribution, computer-aided stylistics.

² Occurrences were taken from a multilingual corpus of medical reports; processed segments have been lemmatized.

1 le examen cardiovasculaire et pleuro-pulmonaire être sans particularité
 2 le examen clinique être sans particularité

Based on these segments, one can identify the underlying pattern DET + examen + être + EVALUATION.³ Sifting through different fixed-sized n -gram lists is a time-consuming—although necessary—task, provided linguists have at least a faint idea of what they are looking for. But this conventional approach does not scale well. Moreover, it relies on a linguist’s prior expectations about the phraseological constraints of the domain under study. What if this process could be, at least in part, automated by using well-established symbolic methods such as Longest Common Subsequence (LCS)? Given a list of similar (or quasi-repeated) segments such as those listed above, LCS extraction would yield le + examen + - + - + - + être + sans + particularité.⁴ By examining all variations of the pattern and by generalizing over bare tokens, or—if annotations are available—lemmas, parts-of-speech, universal dependencies, or other annotation layers (e.g. Named Entities), a pattern such as DET + examen + être + EVALUATION (DET + examination + be + EVALUATION) could be identified semi-automatically. This example shows how linguists can leverage well-established and verifiable string processing methods to promote linguistic patterns based not only on their own expectations, but also on corpus-derived evidence.⁵

Although promising for corpus-driven linguistics, the standard LCS extraction algorithm faces a serious limitations, since it examines all possible pairs in order to extract a LCS for each pair. Unfortunately, this approach makes processing a large list of similar segments intractable in practice, as this pairwise comparison on a corpus of size n amounts to $\frac{n!}{2!(n-2)!}$ combinations, inevitably leading to $O(n^2)$ complexity, or worse, with a naïve implementation. Under this assumption, a corpus of moderate size (ca. 20 000 segments) would yield well over 220 million pairs to process. Since complexity is (at least) quadratic, usual optimization strategies such as parallel processing would have only a limited impact on this bottleneck.

MWU-MINER addresses this bottleneck through the use of advanced string processing techniques borrowed from the fields of data mining and information retrieval. Instead of indexing each individual segment (e.g. a syntactic chunk, sentence, or basic utterance) by using standard hashing techniques, we resort to minHash signatures (Broder, 1997) combined with Locality-Sensitive Hashing (Indyk and Motwani, 1998), in order to approach linear-time complexity. Standard hashing techniques aim at associating each item to a unique key, so that collisions are avoided at all costs. minHash and LSH work in the opposite direction: they provide scalable methods to assign the same key to *similar* segments, grouping them into common “buckets”. This transforms pairwise comparison into a tractable problem. The slight cost in precision (false positives) and recall (false negatives) is largely compensated by the drastic gain in processing time.⁶

Based on this foundation, MWU-MINER implements two complementary approaches:

1. a *pair mining* mode that performs explicit LCS extraction with positional displacement

³ An English equivalent would be DET + examination + be + EVALUATION: “the (rest of the) examination (is/remains) (within normal limits/nominal)”.

⁴ Alternatives are shown as dashes.

⁵ The general strategy outlined here is inspired by Z. Harris’ distributional approach (Harris, 1951), as well as M. Gross’ “Lexique-Grammaire” Gross (1975), and S. Hunston and G. Francis’ “pattern-grammar” (Hunston and Francis, 2000).

⁶ The current version is available at <https://github.com/abalvet/MWUMiner>.

detection, preserving full analytical transparency;

2. a *Directed Acyclic Graph (DAG) mining* mode aimed at encoding the implicit common subsequence as a weighted DAG, i.e. a directed network with no cycles, summarizing the recurring patterns of a cluster of similar segments.

Both modes support multiple annotation levels, i.e. raw tokens, lemmas, POS tags, and mixed lemma+POS pairs, and are designed for a *human-in-the-loop* pipeline where extracted candidates serve as empirical material for qualitative linguistic analysis.⁷

2. Related Work

MWU extraction methods can be categorized into four complementary families.

2.1. Statistical association measures

The most widely used approaches identify MWU candidates through lexical association measures over co-occurrence windows: Pointwise Mutual Information (PMI) (Church and Hanks, 1990), log-likelihood (Dunning, 1993), and their derivatives. These methods are efficient and language-agnostic but operate on unordered pairs and do not model the internal sequential structure of candidate expressions. Syntactically-filtered co-occurrence (Seretan, 2011) constitutes a direct predecessor of our work in spirit, though it does not address the scalability of pairwise comparison.

2.2. Sequence and tree mining

Closer to our approach in spirit, sequence mining methods (Béchet et al., 2012) apply frequent-pattern detection algorithms to tokenized corpora for linguistic pattern discovery, explicitly modeling the sequential structure of recurring units. Tutin and Kraif (2017) propose recurring lexico-syntactic trees (RLTs) as an alternative to n-gram techniques for extended phraseology extraction, demonstrating that syntactically-grounded recurring patterns capture phraseological units more precisely than surface n-grams alone. MWU-MINER shares these goals but addresses scalability via minHash/LSH preprocessing, enabling application to corpora considerably larger than those typically handled by sequence mining methods.

2.3. Graph-based ranking

TextRank (Mihalcea and Tarau, 2004) applies the PageRank algorithm to word co-occurrence graphs for unsupervised keyphrase extraction. Tarau (2019) extends this approach to dependency parse graphs, extracting key-phrases from the largest strongly connected component. These approaches excel at identifying locally salient units within a document, but their notion of relevance—centrality in a local graph—diverges from the MWU criterion of cross-document *formal stability* that motivates MWU-MINER’s DAG construction.

2.4. Approximate similarity search

minHash (Broder, 1997) and Locality-Sensitive Hashing (Indyk and Motwani, 1998) reduce the $O(n^2)$ comparison to near-linear complexity by mapping sets of surface “shingles” to compact signatures, grouping candidate pairs into buckets via banding. Embedding-based similarity methods (Johnson et al., 2019), by contrast, operate over a distributionally-defined similarity

⁷ Universal Dependencies annotations (Nivre et al., 2016) will be supported in future versions.

space, capturing semantic proximity (e.g. “be on the lookout” $\approx$ “monitor”) rather than formal surface recurrence. For MWU identification, formal recurrence is the defining criterion: the two approaches are therefore *complementary* rather than competing.

3. Architecture

MWU-MINER takes a plain text, or a CoNLL-U formatted file as input⁸ and produces either a ranked TSV (Tab-Separated Values) file of annotated segment pairs (pair mining), or a JSON (JavaScript Object Notation) summary with per-cluster pattern DAG candidates (DAG mining).

3.1. Shared preprocessing

All input segments are first tokenized at the chosen annotation level. In lemma+upos mode, each token is represented as a lemma/pos string (e.g. payer/VERB). Bigram shingles are then extracted from each tokenized segment to constitute the set representation used for Jaccard similarity estimation. Named entities and numerals are optionally normalized to placeholder tokens (<NAM>, @card@), reducing lexical sparsity in domain-specific corpora.

3.2. Pair mining mode

The intuition behind pair mining is straightforward: for each pair of segments exhibiting similarity above a given threshold, the system extracts their longest common subsequence and records which tokens are shared, whether they appear at the same position, or whether sub-elements have moved. Two candidate-generation strategies are available at runtime.

Exhaustive strategy. All $n(n-1)/2$ pairs are examined. For each pair, a bounded relative Levenshtein distance is computed; pairs below a threshold proceed to LCS extraction. This mode is limited by the $O(n^2)$ asymptotic complexity described above.

LSH strategy. minHash signatures of configurable length (default: 64 permutations) are computed per segment. The signature matrix is partitioned into b bands of r rows ($b \cdot r = \text{num\_perm}$); segments sharing at least one identical band are promoted as candidate pairs. The banding parameters control the effective similarity threshold. After this step, candidate pairs are scored for estimated Jaccard similarity (via signature agreement) and exact Jaccard (via shingle set intersection), then subjected to LCS extraction. For each processed pair, the following metrics are exported: relative Levenshtein distance, estimated and exact Jaccard similarity, shared LSH band count, LCS length and coverage ratios (`lcs_cov_seg1`, `lcs_cov_seg2`), positionally stable token count (`same_position_hits`), and displaced token count (`moved_hits`). The displacement metric annotates each moved token with its source and target positions (e.g. <MOV UNIT="ne" ORIG="1" DEST="2">), providing a direct assessment of MWU positional flexibility.

3.3. DAG mining mode

Rather than examining individual pairs, the DAG mining mode summarizes all similar segments within a cluster as a single compact directed graph. This “big-picture” representation makes it easier to detect the stable core of a phraseological pattern and its possible variations. In DAG mining mode, nodes represent token types at given positions within the cluster; edges represent sequential adjacency, weighted by co-occurrence frequency. Additional *skip edges* (with

⁸ CoNLL-U files encode tokenization, lemmas, part-of-speech tags, and dependency relations in a tab-separated ten-column structure.

reduced weight) encode non-contiguous transitions, making it possible to represent discontinuous patterns.

The DAG mining mode uses the Python `datasketch` `minHash/LSH` implementation to cluster segments into buckets, then builds a weighted directed acyclic graph (*PatternDAG*) over each bucket. Pattern extraction applies a priority-queue search over edges above a minimum frequency threshold. The maximum-weight path from source to sink constitutes the implicit LCS of the cluster; skip edges encode positional variation. In other words, heavier edges correspond to the most frequent transitions observed across the cluster’s segments.

Buckets and source segments are indexed through an inverted structure enabling post-hoc querying: given a lemma, POS tag, or lemma+POS pair, the system returns all clusters in which the element participates with co-occurrence frequencies, effectively computing the element’s *local grammar* (Sinclair, 2004) across the corpus.⁹

4. Experiments

4.1. Corpora and settings

We report results on two corpora: (1) a corpus of **French football match commentary** (raw tokens with named entity normalization), used to illustrate domain-specific MWU extraction;¹⁰ (2) the text layer of the **ParseME manually annotated corpus** (Web-fr subcorpus; 4 063 sentences, 108 244 tokens, annotation level: lemma+upos).

4.2. Pair mining: sports commentary

Figure 1 below shows a sample of similar segments detected by the LSH strategy on the raw text version of the football commentary corpus. The extracted segments show unexpected phraseological pattern candidates revolving around *cuir* (leather), as a conventionalized meronym of *ballon* (ball), and exhibit little internal variation. Since leather footballs have fallen out of use decades ago, the mention of *cuir* can only be attributed to idiomatic and stylistic constraints.

The MWU extraction strategy identifies pairs such as:

1. Ils combinent bien à l’ entrée de la surface mais Merville peut se saisir de le ballon .
2. C’ est un peu trop profond et Elana peut se saisir de le cuir .

Based on these segments, the LCS extraction phase yields the following pattern:

...	peut	se	saisir	de	le	ballon
...	peut	se	saisir	de	le	cuir
LCS:	peut	se	saisir	de	le	(cuir ballon)

This example shows how unexpected phraseological patterns can be spotted thanks to our MWU detection and extraction strategy. In this instance, a linguist can easily infer a synonymy relation

⁹ The notion of local grammar belongs to the Neo-Firthian tradition of corpus linguistics (Sinclair, 1991), which searches for patterns that emerge from observed co-occurrence constraints over predefined grammatical categories. It is also related to M. Gross’ “Lexique-Grammaire”, and Hunston and Francis’ “pattern grammar”.

¹⁰ This corpus was set up by N. Gasiglia (UMR STL 8163), but due to copyright constraints, it is not publicly available.

pair_id	i	j	segmen t_id_1	segmen t_id_2	segment_1	segment_2
41807-45527	41807	45527	50648	55373	le cuir filer directement dans le tribune	puisque le cuir filer directement dans le tribune
55085-61920	55085	61920	67537	75963	<NAM> placer le cuir sur le tête de <NAM> sur corner	<NAM> placer le cuir sur le tête de <NAM>
25939-39370	25939	39370	30760	47548	<NAM> hériter du cuir dans le surface lyonnais	<NAM> hériter du cuir dans le surface
62730-80220	62730	80220	76938	100262	<NAM> se battre pour conserver le cuir	se battre pour conserver le cuir
2185-13428	2185	13428	2381	15611	faire tourner le cuir	monaco faire tourner le cuir
16789-59586	16789	59586	19654	73081	<NAM> hériter du cuir sur le gauche de le surface	<NAM> hériter du cuir sur le gauche de le surface et frappe
30654-47795	30654	47795	36671	58317	<NAM> se saisir du cuir et soulager son équipe	<NAM> se saisir du cuir et soulager son défense
18653-28978	18653	28978	21861	34573	il capter le cuir	il capter le cuir aisément
49544-70920	49544	70920	60643	87706	il monopoliser le cuir sans se montrer dangereux	<NAM> monopoliser le cuir sans se montrer dangereux
13703-87302	13703	87302	15955	110317	<NAM> sortir de son but et capter le cuir	<NAM> sortir de son but et capter le ballon
4507-18653	4507	18653	5014	21861	il capter le cuir impeccablement	il capter le cuir
30166-79097	30166	79097	36056	98683	et <NAM> bloquer le cuir	<NAM> bloquer le cuir

Figure 1: Patterns based on “cuir” (French football commentary corpus)

between *cuir* and *ballon*, which is valid only for the domain of football commentaries, even though instances of such pairs are extremely rare—which would place this phenomenon outside the scope of purely statistical approaches.

4.3. DAG mining

As shown above, the pair mining approach provides solid, verifiable evidence to promote candidates to the status of actual collocational or phraseological units, but it lacks a “big picture” feature. The complementary DAG-centered approach makes it easier to detect stable elements in phraseological patterns, and paves the way to a corpus-driven pattern-grammar¹¹ approach, possibly beyond mere MWU detection and extraction. DAG mining relies on the same minHash/LSH foundation as pair mining, but it aims at extracting “local grammars” from the clusters identified during preprocessing. The extracted DAGs can easily be converted into finite-state automata and transducers, respectively for searching and annotating or normalizing detected patterns.

To illustrate the DAG mining strategy, we provide a local grammar extracted from the ParseME Web-fr corpus. This corpus was developed in the context of the ParseME Cost action, for the development of resources and methods for a proper treatment of Multi-Word Entities. Figure 2 shows how our DAG mining approach identifies a cluster centered on *ne payer pas de mine*.¹² The backbone path (ne→payer→pas→mine) corresponds to the core MWU; branching nodes encode left-context variability (*de l’extérieur, si, l’hôtel, la façade*) and right-context extensions. Querying the cluster index with lemma=mine, upos=NOUN separates this idiomatic cluster from those where *mine* occurs as a free nominal head (*mine d’or, mine de charbon*). In the extracted DAG, edge weights indicate co-occurrence frequency; heavy edges (≥ 4) form the backbone path.

¹¹Or even word-grammar.

¹²An English equivalent to the frozen expression “ça ne paye pas de mine” would be “it doesn’t look like much”.

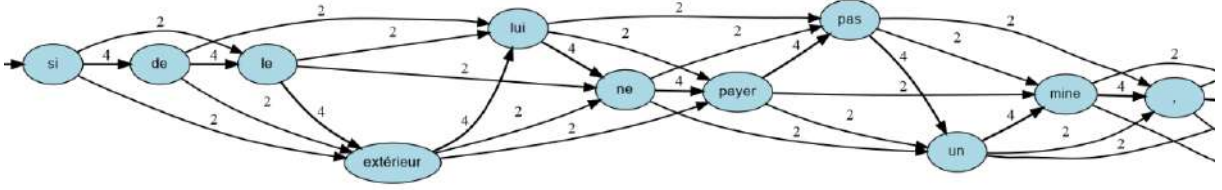


Figure 2: Pattern DAG for the cluster centered on *ne+payer+pas+de+mine*.

Although the DAG mining strategy outlined above appears as a powerful tool for corpus linguists, it should be underlined that, in its current version, MWU-MINER does not perform any grammar induction steps, therefore the extracted finite-state automata are mere “local grammar” candidates. A proper grammatical induction phase, based on inductive logic programming and probabilistic context-free grammars, is nonetheless an appealing perspective for future developments.

5. Discussion

MWU-MINER provides a scalable, language-agnostic, annotation-aware toolkit for Multi-Word Units detection and extraction, combining the analytical transparency of pairwise LCS mining with the throughput of minHash/LSH-based cluster analysis. The two modes are complementary: DAG mining identifies candidate clusters efficiently at scale, while pair mining characterizes the flexibility of candidates within clusters through explicit displacement metrics. When applied to corpora spanning parliamentary debates, web text, medical reports, and sports commentary, the system demonstrates consistent extraction of domain-specific formulaic expressions across annotation levels. Both strategies instantiate different trade-offs along three dimensions.

1) Analytical granularity vs. throughput. The pair mining mode provides explicit LCS strings, per-token displacement annotations, and multiple similarity scores amenable to statistical filtering. This analytical depth comes at the cost of throughput: even with LSH pre-filtering, scoring all candidate pairs requires $O(|C| \cdot L^2)$ time, where $|C|$ is the candidate set size and L the average segment length. The DAG mining mode sacrifices explicit LCS strings for cluster-level pattern generalisation: the implicit backbone is recoverable as the maximum-weight path in the DAG, but displacement is only encoded structurally via skip edges.

2) Representation and similarity. Both modes use Jaccard similarity over surface shingle sets as their proximity measure, which is epistemologically appropriate for MWU detection: the target is formal surface recurrence, not distributional semantic proximity. An embedding-based approach such as (Johnson et al., 2019) would capture the complementary task of MWU semantic clustering, i.e. grouping expressions that are semantically equivalent rather than formally recurrent. Since alternative efficient similarity-based approaches, such as simHash (Sadowski and Levin, 2007; Sood and Loguinov, 2011) have been developed, these methods should be assessed in the context of MWU detection, in replacement of or in conjunction with minHash (Shrivastava and Li, 2014).

3) Annotation level. The lemma+upos level provides the richest representation for syntactic pattern discovery; the token level, while noisier, handles corpus artifacts (web text, tokenisation errors) more robustly. Both levels are supported by both modes.

A key limitation of the current system is its treatment of variable slots: the pair mining mode identifies displaced tokens but does not yet generalise across multiple pairs to produce slot-

filling patterns (e.g. *chain _ acyl*). The DAG mining mode encodes slot variability structurally but does not output formal slot specifications. Explicit slot induction constitutes the primary direction for future development, alongside a systematic quantitative evaluation against gold standards such as the ParseME corpora.

6. Perspectives

6.1. Phraseological density as a corpus metric

Beyond individual MWU extraction, both the pair mining and DAG mining modes yield aggregate statistics that may serve as corpus-level metrics of *phraseological density*. Intuitively, a corpus strongly constrained by specialized terminology or formulaic language will produce few, heavily populated clusters, whereas a stylistically creative corpus will generate many sparsely populated ones.

Formally, given n segments clustered into K buckets of sizes $s_1 \dots, s_K$ at LSH threshold θ , with C representing the candidate set size, the following metrics can be defined:

1. **phraseological density:** $\Delta(C, \theta) = \frac{\sum_k s_k}{n \cdot K}$
2. **phraseological fragmentation:** $\Phi(C, \theta) = \frac{K}{n}$
3. **pattern coverage:** $\Gamma(C, \theta) = \frac{\sum_k s_k}{n}$

Plotting (Δ, Φ, Γ) as a function of θ produces a *phraseological signature* of the corpus: a profile that characterizes the distribution of sequential repetition at multiple granularity levels. Unlike classical lexicometry or textometry indices (Text Token Ratio, lexical specificity index), these metrics capture *sequential* structural repetition; moreover, they are computable in $O(n \cdot k)$ time via minHash signatures, which allows for the application of these metrics on large corpora, or text archives comprising multiple text files. Systematic comparison of such signatures across corpora of contrasting registers (specialized vs. literary, oral vs. written) constitutes a promising direction for applications in textometry.

6.2. Stylistics, authorship, and philology

The identification of highly similar segment pairs, regardless of the mining mode, opens further application domains.

Authorship attribution and stylistics. We posit that different authors may be associated with different phraseological density profiles, capturing habitual patterns of sequential repetition beyond the lexical level. Such profiles can complement classical character n-gram or embedding-based authorship attribution methods.

Plagiarism and near-duplicate detection. Plagiarism detection strategies often rely on the exact repetition of a significant subset of n-grams. Our approach to similarity detection opens new perspectives for exact and near-duplicate detection: high-Jaccard pairs ($\theta > 0.7$) with low `moved_hits` may signal verbatim copying, while high-Jaccard pairs with `moved_hits = lcs_length` may be interpreted as a paraphrased segment. The displacement annotation (`<MOV UNIT=... ORIG=... DEST=...>`) provides a structured representation of transformational distance between near-duplicate passages.

Digital philology. The LCS-with-displacement representation is directly applicable to textual variant collation: two versions of the same text will undergo token displacements, lexical sub-

stitutions, and omissions, which are precisely the transformations encoded by `moved_hits`, `same_position_hits`, and relative Levenshtein distance. This positions MWU-MINER as a lightweight alternative or complement to dedicated collation tools such as CollateX, applicable without standoff markup or TEI encoding.

References

- Biber D., Johansson S., Leech G., Conrad S., and Finegan E. (1999). *Longman Grammar of Spoken and Written English*. Longman, London.
- Broder A. Z. (1997). On the Resemblance and Containment of Documents. In *Proceedings of Compression and Complexity of Sequences 1997*, pp. 21–29. IEEE.
- Béchet N., Cellier P., Charnois T., and Crémilleux B. (2012). Discovering linguistic patterns using sequence mining. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 154–165. Springer.
- Bécue M. (1993). Une méthode d’analyse des données textuelles. In *Actes des 2^{es} JADT*, pp. 23–32.
- Church K. W. and Hanks P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.
- Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.
- Gross M. (1975). *Méthodes en syntaxe*. Hermann, Paris.
- Harris Z. (1951). *Methods in structural linguistics*. The University of Chicago Press, Chicago.
- Hunston S. and Francis G. (2000). *Pattern Grammar, A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing Company, Oxford.
- Indyk P. and Motwani R. (1998). Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC 1998)*, pp. 604–613, Dallas, TX. ACM.
- Johnson J., Douze M., and Jégou H. (2019). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Lebart L. and Salem A. (1994). *Statistique textuelle*. Dunod, Paris.
- Mihalcea R. and Tarau P. (2004). TextRank: Bringing Order into Texts. In *Proceedings of EMNLP 2004*, pp. 404–411, Barcelona. ACL.
- Nivre J. et al. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC 2016*, Portorož.
- Sadowski C. and Levin G. (2007). simHash: Hash-based Similarity Detection. Technical report, Technical report, Google.
- Seretan V. (2011). *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer, Dordrecht.
- Shrivastava A. and Li P. (2014). In defense of minHash over simHash. In *Artificial intelligence and statistics*, pp. 886–894. PMLR.
- Sinclair J. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford.
- Sinclair J. (2004). *Trust the Text: Language, Corpus and Discourse*. Routledge, London.
- Sood S. and Loguinov D. (2011). Probabilistic Near-duplicate Detection Using simHash. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1117–1126.
- Tarau P. (2019). Dependency-Based Text Graphs for Keyphrase and Summary Extraction with Applications to Interactive Content Retrieval. arXiv preprint arXiv:1909.09742.

- Tutin A. and Kraif O. (2017). Comparing Recurring Lexico-Syntactic Trees (RLTs) and Ngram Techniques for Extended Phraseology Extraction. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pp. 176–180.

Estimating the typical-source distribution in imbalanced corpora

Aris Xanthos¹

¹University of Lausanne — aris.xanthos@unil.ch

Abstract

When a corpus is assembled as a sample of sources from a population of interest, a natural research target is the *typical-source* distribution – the average linguistic profile across contributors, treating each as an equally valid informant. Token-pooled MLE (POOL) is biased for this target whenever source size and content are coupled; uniform source averaging (UNIF) eliminates this bias but injects noise from sparse sources at full weight. We compare these two estimators and capped-mass weighting (CAP) across a parametric simulation varying size–content coupling and the degree of size imbalance. The right choice depends on both: POOL is competitive when coupling is absent; UNIF is appropriate when coupling is strong and sources are well-sampled; CAP offers a robust option when the coupling regime is unknown. A WhatsApp emoji case study, complemented by a lexical sensitivity check, illustrates that the choice is consequential in practice.

Keywords: Source imbalance, population estimation, corpus weighting, simulation study, WhatsApp data

1. Introduction

When a corpus is assembled as a sample of sources from some population of interest, a legitimate research question concerns what a typical source looks like in this population. We call this target estimand the *typical-source* distribution – the average linguistic profile across sources, treating each as an equally valid informant. We use “source” mainly for individual speakers or users, but the same issue arises whenever observations are grouped into units over which one wants a population-level estimate: authors in literary corpora, speakers in parliamentary corpora, interviewees in oral-history corpora, users in social-media corpora, or non-personal categories such as genres, topics, periods, or registers.

Token-pooled MLE (POOL), the standard baseline, is optimal for this target only in special cases: when all sources are homogeneous in content, or when they contribute equal token mass; the latter is usually the result of an explicit design decision, like the 2,000-word samples of the Brown corpus (Kučera and Francis, 1967), and requires discarding data from larger sources or excluding sparser ones. Outside these cases, POOL induces a distortion that grows with *size–content coupling*: the situation in which the amount of material contributed by a source is not independent of its content profile. Larger sources are then not merely larger random samples from the same population; they may be associated with particular lexical, stylistic, or topical preferences, while smaller sources may be associated with different ones. Such coupling is plausible in corpora assembled from voluntary participation, where inclusion probability can be correlated with the communicative practices the corpus is meant to document (Groves and Peytcheva, 2008).

The direct estimator of the typical-source distribution, uniform source averaging (UNIF), eliminates this bias entirely – but at a cost. When sources contribute few tokens, their empirical

distributions are noisy, and UNIF injects that noise at full weight. The result is a reliability–representativeness trade-off: POOL is reliable but potentially biased; UNIF is unbiased but potentially noisy. The right balance depends on what the analyst knows, or is willing to assume, about the reliability of per-source estimates given the available token counts and about the degree of coupling between source size and content. In real corpora, the existence of such coupling is not directly observable, because we do not know what the corpus would have looked like if all sources had contributed the same amount of material. What we can observe is whether the estimate changes when source size is prevented from determining source influence. Large differences between POOL and UNIF therefore provide a sensitivity signal: they show that the population profile being reported depends on how the analyst handles source imbalance.

This paper compares POOL and UNIF as the two endpoints of the reliability–representativeness trade-off, and capped-mass weighting (CAP) as a practical compromise between them. We evaluate all three using total variation distance to the true typical-source distribution, across a parametric simulation that varies size–content coupling and the degree of size imbalance. The evaluation is designed to characterize each estimator’s behavior across these two dimensions and to identify where the choice between them is consequential. A WhatsApp corpus serves as a secondary case study: the main illustration concerns emoji distributions, while a compact lexical sensitivity check shows that the same estimator issue also arises for word distributions. We conclude by discussing estimator choice as a function of what can be known or assumed about the degree of size–content coupling.

2. Estimand and estimators

Let S sources be indexed by $s \in \{1, \dots, S\}$, with a vocabulary of V types. We observe counts $c_s(t)$ for each type t and source s , with $T_s := \sum_t c_s(t)$ the token mass of source s and $N := \sum_s T_s$ the total corpus mass. The empirical per-source distribution is $\hat{\theta}_s(t) := c_s(t)/T_s$ for $T_s > 0$.

2.1. The typical-source distribution

We take as our estimand the *typical-source distribution*, defined as the unweighted average of the true per-source distributions:

$$q^*(t) = \frac{1}{S} \sum_s \theta_s(t).$$

This is the distribution one would obtain by drawing a source uniformly at random from the population and then drawing a token from it. The choice of q^* as the target is a prior decision about what the corpus is meant to represent: not the aggregate token stream, but an average over the population of sources, each treated as an equally valid informant. “Source” thus denotes the unit over which the analyst wants population balance. In our WhatsApp case study this unit is a user, but in other corpora it may be an author, speaker, genre, topic, period, or other grouping variable.

2.2. Three estimators

All three estimators can be written as a weighted mixture

$$\hat{q}(t) = \sum_s w_s \hat{\theta}_s(t),$$

where the source weights $w_s \geq 0$ sum to one. They differ only in how those weights are assigned.

Pooled MLE (POOL).

$$w_s^{\text{pool}} := \frac{T_s}{N} \quad \Rightarrow \quad \hat{q}_{\text{pool}}(t) = \frac{1}{N} \sum_s c_s(t).$$

This is the natural baseline when tokens are the sampling unit. It is unbiased for q^* when all sources share the same true distribution, or when all sources have equal token mass. It concentrates influence on the largest sources and, when sources also differ in content, induces a form of bias whose severity increases with size–content coupling.

Uniform source average (UNIF).

$$w_s^{\text{unif}} := \frac{1}{S} \quad \Rightarrow \quad \hat{q}_{\text{unif}}(t) = \frac{1}{S} \sum_s \hat{\theta}_s(t).$$

This is the direct estimator of q^* and eliminates size bias entirely. Its weakness is the converse of POOL’s: when T_s is small, $\hat{\theta}_s$ is noisy, and that noise enters $\hat{q}_{\text{unif}}$ at full weight. POOL and UNIF define the two endpoints of a reliability–representativeness spectrum; CAP occupies the interior.

Capped-mass weighting (CAP).

$$w_s^{\text{cap}} := \frac{\min(T_s, M)}{\sum_{s'} \min(T_{s'}, M)} \quad \Rightarrow \quad \hat{q}_{\text{cap}}(t) = \frac{\sum_s \min(T_s, M) \hat{\theta}_s(t)}{\sum_s \min(T_s, M)}.$$

Sources with $T_s \leq M$ contribute their full mass; sources with $T_s > M$ are scaled down deterministically by M/T_s . At $M \geq \max_s T_s$, no source is capped and the estimator recovers POOL; at $M \leq \min_s T_s$, all sources are capped to the same effective mass M , so the estimator recovers UNIF. CAP is the source-level analogue of weight trimming in survey estimation: just as extreme sampling weights are truncated to prevent any single unit from dominating an estimate (Haziza and Beaumont, 2017), capping prevents any single source from dominating by sheer token volume. The threshold M is set as $M = \alpha \cdot M_0$, where $M_0 := N/S$ is the equal per-source share of corpus mass and $\alpha > 0$ is a dimensionless parameter. At $\alpha = 1$ the cap equals the equal share; $\alpha < 1$ pushes CAP toward UNIF, $\alpha > 1$ toward POOL. We evaluate CAP across $\alpha \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$; default results use $\alpha = 1$.

3. Method and data**3.1. Metric**

We evaluate on synthetic data where q^* is known exactly. The primary metric is total variation distance,

$$\text{TV}(\hat{q}, q^*) := \frac{1}{2} \sum_t |\hat{q}(t) - q^*(t)|,$$

a standard measure of distributional discrepancy bounded in $[0, 1]$ and sensitive to differences across the full vocabulary. All results report the median TV across 100 fixed random seeds; the interquartile range (IQR) serves as a measure of stability across replications.

On real data, where q^* is unknown, TV is used only to quantify disagreement between estimators, not error with respect to a ground truth.

3.2. Population generator

We use a single parametric generator with fixed global settings: $S = 100$ sources, $N = 100,000$ tokens, a vocabulary of $V = 1,000$ types whose base distribution follows Zipf’s law ($q_{\text{base}}(r) \propto r^{-1.1}$), and $K = 3$ content regimes. Regime centroids are constructed by boosting a disjoint 10% subset of vocabulary types per regime until the mean pairwise TV between centroids reaches 0.2.

Two factors are varied systematically. The first is the *degree of size imbalance*, with three levels: *no imbalance* (all sources receive equal token mass), *mild imbalance* (80/20 allocation: the top 20% of sources receive 80% of total tokens, approximately 4,000 each, the remainder approximately 250 each), and *strong imbalance* (95/5 allocation: the top 5% of sources receive 95% of tokens, approximately 19,000 each, the remainder approximately 53 each). The no-imbalance condition serves as a theoretical baseline: without size imbalance, q^* coincides with the pooled population distribution regardless of coupling, and estimator choice is inconsequential. The strong imbalance condition is motivated by the imbalance profiles observed in voluntary contribution corpora, where a small number of highly active contributors can account for the large majority of tokens while the bottom half contribute only a few percent.

The second factor is the *coupling parameter* $c \in \{0, 0.33, 0.67, 1\}$, which controls the fraction of large sources forced into a single content regime (“regime 0”), as opposed to being distributed uniformly across all regimes like the remaining sources. The index “0” is arbitrary; in the simulation design, regime 0 is the regime made increasingly over-represented among large sources as the coupling parameter increases. At $c = 0$, all sources are distributed uniformly across all regimes: size imbalance and content heterogeneity are both present, but without structural alignment. At $c = 1$, all large sources are in regime 0 and all small sources in the remaining regimes: size and content are maximally coupled. Intermediate values produce partial alignment between source size and content regime.

3.3. Real-data case study

As a secondary illustration, we apply POOL, UNIF, and CAP to data from the *What’s New, Switzerland?* corpus (WNS; Xanthos et al., 2025, 2024), a collection of donated WhatsApp chats between French-speaking users in Switzerland comprising 72 chat groups and 118 unique users. Sources are defined at the user level. After excluding system-authored posts, extracting contiguous emoji runs, removing variation selectors and skin-tone modifiers, and retaining gendered emoji sequences, the emoji subcorpus contains 122,828 emoji tokens and 1,087 distinct types produced by 117 users. It is strongly imbalanced: the top 10% of sources account for 67.7% of emoji tokens, while the bottom 50% contribute only 3.0% (median: 176 tokens/source; fifth percentile: 10; maximum: 27,900). Because no ground truth is available, the case study is illustrative rather than evaluative: it shows the magnitude of estimator disagreement on a realistic corpus and makes concrete the stakes of the weighting choice. We compare the three estimators by their top- K emoji distributions (Table 1); the TV between POOL and UNIF serves as a reference for the maximum possible disagreement in this dataset.

To assess whether estimator sensitivity is specific to emoji tokens, we also run a lexical sensitivity check on the text content of messages from the same source-partitioned corpus. The “system” pseudo-source is excluded, text is lowercased and tokenized with a simple word-token regular expression, and the lexical vocabulary excludes documented WNS de-identification placeholders, media/notification/system-message placeholders, and simple technical artifacts such as numeric, URL-like, and email-like tokens. The resulting data contains 3,191,868 word to-

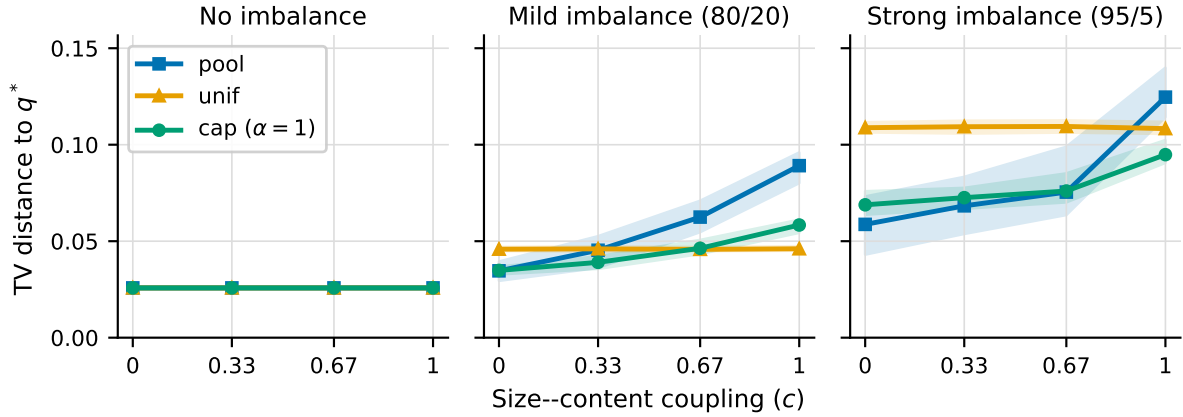


Figure 1: Median TV distance to q^* as a function of coupling c for three estimators (POOL, UNIF, CAP at $\alpha = 1$) across three size imbalance conditions. Shaded bands show the IQR across 100 fixed seeds. $S = 100$, $V = 1,000$, $N = 100,000$.

kens and 60,024 types, with size imbalance comparable to the emoji subcorpus: the top 10% of sources account for 68.2% of word tokens, while the bottom 50% contribute only 3.2% (median: 4,981 tokens/source; fifth percentile: 243; maximum: 522,083). This analysis is not intended as a substantive lexical study of the corpus; it is used only to test whether estimator choice also affects word distributions. We report full-vocabulary TV distances between estimators and the word tokens most shifted between POOL and CAP at $\alpha = 1$.

4. Results

4.1. Estimator behavior across coupling and imbalance regimes

Figure 1 shows median TV to q^* as a function of coupling c for each combination of imbalance level and estimator.

The no-imbalance panel confirms the theoretical prediction: when all sources have the same token mass, all three estimators are identical regardless of coupling strength, with $TV \approx 0.026$ throughout. This result serves as a sanity check and establishes the baseline: estimator choice is consequential only when size imbalance is present.

Under mild imbalance (80/20), POOL and CAP are nearly identical at $c = 0$ ($TV \approx 0.035$) while UNIF is already noticeably worse ($TV = 0.046$). This reflects the variance cost of giving full weight to small sources in the absence of coupling: with approximately 250 tokens each, the bottom 80% of sources contribute noisy per-source estimates that inflate UNIF’s error. As coupling increases, POOL rises steeply ($TV = 0.089$ at $c = 1$) while UNIF remains flat, and CAP occupies a stable intermediate position ($TV = 0.058$ at $c = 1$). The crossing point where POOL overtakes UNIF occurs around $c = 0.33$.

Under strong imbalance (95/5), the picture changes substantially. UNIF is now the worst estimator throughout most of the coupling range: with approximately 53 tokens per source for the bottom 95%, the per-source estimates are far too noisy to enter the mixture at full weight, and UNIF incurs $TV \approx 0.109$ regardless of c . POOL is competitive at low coupling ($TV = 0.059$ at $c = 0$) but fails badly as coupling strengthens ($TV = 0.125$ at $c = 1$). CAP at default α achieves a low TV across the coupling range: it sits close to POOL at low coupling ($TV = 0.069$ at $c = 0$), tracks it closely as coupling increases, and overtakes both endpoints at full coupling

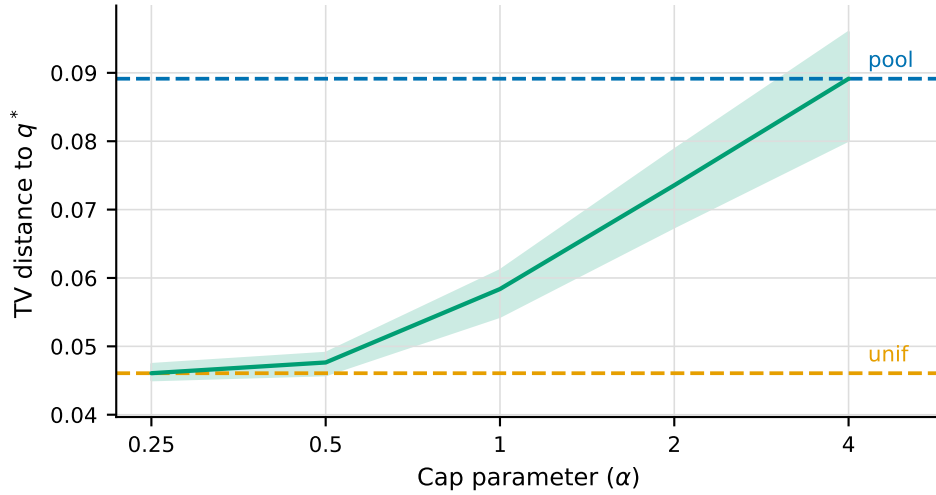


Figure 2: Median TV of CAP as a function of α at mild imbalance (80/20), full coupling ($c = 1$). The x-axis is log-scaled; consecutive ticks correspond to doubling the cap. Shaded band shows the IQR across 100 fixed seeds. Dashed lines show POOL and UNIF as reference.

(TV = 0.095 at $c = 1$, vs 0.125 for POOL and 0.108 for UNIF). In the present evaluation, this is the most favorable condition for CAP: strong size imbalance combined with strong size–content coupling.

4.2. Parameter sensitivity

Figure 2 shows the full α sweep for CAP at mild imbalance and full coupling ($c = 1$), with POOL and UNIF as reference lines.

The sweep is monotone and spans the full estimator spectrum: at $\alpha = 0.25$, CAP matches UNIF exactly (TV = 0.046); at $\alpha = 4$, it matches POOL (TV = 0.089). In this setting, α therefore controls how much source-mass information is retained, providing a practical way to tune the balance between representativeness and reliability without implying a simple token-wise interpolation between POOL and UNIF.

4.3. Real-data illustration

Applying the estimators to the WNS emoji corpus gives $\text{TV}(\text{POOL}, \text{UNIF}) = 0.304$, confirming that the weighting choice has substantial practical stakes in this dataset. In total variation distance, CAP occupies an intermediate position between the two endpoints: at $\alpha = 0.5$ it falls closer to UNIF (TV = 0.153 from UNIF, 0.221 from POOL); at $\alpha = 1$ its distances to the two endpoints are similar (0.181 from UNIF and 0.199 from POOL); at $\alpha = 2$ it moves toward POOL (0.208 from UNIF, 0.175 from POOL). Without ground truth or an external coupling diagnostic, the case study cannot identify the underlying coupling regime. The divergence between estimators should therefore be read as a diagnostic of estimator consequentiality: the estimated population profile depends strongly on whether prolific users are allowed to dominate.

Table 1 shows the top-20 emoji distributions. The face with tears of joy 🥳 receives 25.6% of mass under POOL but only 12.1% under UNIF, showing that conclusions about the dominant emoji depend strongly on the weighting scheme. At $\alpha = 1$, CAP gives an intermediate estimate of 16.3%. The composition of the top-20 list also changes. POOL’s top 20 includes types such as 🍷, 🍷, and 🍷 that are absent from UNIF; 🍷 reappears only under the most permissive cap

Rank	UNIF	CAP			POOL
		$\alpha=0.5$	$\alpha=1.0$	$\alpha=2.0$	
1	😂 0.121	😂 0.146	😂 0.163	😂 0.182	😂 0.256
2	😂 0.063	❤️ 0.075	❤️ 0.078	❤️ 0.079	❤️ 0.091
3	❤️ 0.052	😂 0.072	😂 0.072	😂 0.077	😂 0.054
4	👍 0.042	😂 0.063	😂 0.068	😂 0.064	😂 0.050
5	😊 0.042	👍 0.035	👍 0.031	😂 0.029	📧 0.032
6	😂 0.042	😊 0.030	😊 0.029	👍 0.028	😂 0.027
7	😂 0.039	😂 0.028	😂 0.029	😊 0.026	😂 0.024
8	😊 0.033	😊 0.027	😊 0.026	😊 0.024	😂 0.020
9	😊 0.030	😊 0.026	😊 0.024	😊 0.024	🌟 0.019
10	😂 0.023	😂 0.024	😂 0.023	😂 0.023	😊 0.018
11	😂 0.022	😊 0.019	😊 0.020	😊 0.022	👍 0.017
12	😂 0.020	😂 0.019	😂 0.017	😂 0.015	😊 0.015
13	😂 0.018	💜 0.013	😊 0.012	😊 0.012	👍 0.015
14	👍 0.017	😊 0.012	🙏 0.011	🙏 0.012	🙏 0.014
15	😂 0.014	🙏 0.012	🏠 0.011	📧 0.010	😊 0.013
16	🌟 0.014	🏠 0.011	👋 0.011	👋 0.010	👋 0.013
17	😊 0.013	👋 0.011	🙏 0.010	🙏 0.010	😊 0.013
18	🙏 0.012	😂 0.011	😂 0.010	🏠 0.010	🏠 0.011
19	😂 0.011	😂 0.011	💜 0.010	😂 0.009	🙏 0.011
20	👍 0.010	👍 0.010	😂 0.010	👍 0.009	😊 0.010
Head mass	0.636	0.655	0.665	0.676	0.722
TV to POOL	0.304	0.221	0.199	0.175	–
TV to UNIF	–	0.153	0.181	0.208	0.304

Table 1: Top-20 emoji by estimator on the WNS corpus (probabilities rounded to 3 d.p.). Three values of α show the effect of relaxing the cap: smaller values give more source-balanced weights, while larger values allow high-mass sources to retain more influence. Head mass is the total probability assigned to the top-20 types. TV distances use the full distribution over all emoji types.

shown here ($\alpha = 2$). Conversely, 😂 and 🙏 are present only in UNIF. Other types, such as 💜 and 🙏, appear only under capped weighting for some values of α , while 🌟 is present at both endpoints but not under CAP. This reflects a rebalancing of source weights rather than a simple interpolation of the endpoint distributions.

Because no ground truth is available, this case study is illustrative rather than evaluative, but it establishes that estimator disagreement at this scale is large enough to affect substantive conclusions about emoji usage in this community. The disagreement is not by itself proof of a particular coupling mechanism; it shows that source weighting changes the estimated population profile.

The lexical sensitivity check leads to the same methodological conclusion. While estimator disagreement is lesser for words than for emojis, it remains substantial (table 2): $TV(\text{POOL}, \text{UNIF}) = 0.138$, $TV(\text{POOL}, \text{CAP}) = 0.089$, and $TV(\text{UNIF}, \text{CAP}) = 0.086$. The most shifted items (table 3) are mainly ordinary conversational forms – pronouns, acknowledgements, laughter tokens, and short grammatical items – rather than an obvious set of topic-specific lexical markers. This suggests that source imbalance affects not only rare or domain-specific vocabulary, but also the estimated profile of everyday conversational usage. This pattern is compatible with lexical profile differences associated with source size in the corpus, but is used here only as a sensitivity check, not as a substantive lexical analysis.

	POOL	CAP ($\alpha = 1$)	UNIF
POOL	–	0.089	0.138
CAP ($\alpha = 1$)	0.089	–	0.086
UNIF	0.138	0.086	–

Table 2: Total variation distances between lexical estimates after excluding documented WNS placeholders and simple technical artifacts.

Higher under CAP	n_{src}	$\Delta/10\text{k}$	Higher under POOL	n_{src}	$\Delta/10\text{k}$
et	118	+23.26	oui	114	-71.27
je	117	+20.27	haha	50	-40.68
on	115	+17.61	c	115	-29.92
pour	116	+16.53	est	117	-28.35
vous	107	+14.06	il	113	-26.60
à	117	+12.08	ça	116	-22.74
merci	113	+10.77	ils	94	-15.44
si	114	+10.18	hahaha	50	-10.71
coucou	75	+10.05	un	116	-8.41
moi	113	+8.69	j	118	-8.25

Table 3: Largest signed lexical shifts between the capped ($\alpha = 1$) and pooled estimates. Values are differences per 10,000 tokens; n_{src} gives the number of sources in which the token occurs.

5. Discussion and conclusion

The estimator choice studied here is fundamentally a question of what the analyst knows – or is willing to assume – about the corpus and its collection process. POOL is not simply wrong: it is the optimal estimator when sources are homogeneous in content, and remains competitive under weak size–content coupling. The problem arises when coupling is strong, a situation in which POOL systematically over-represents the largest sources. UNIF eliminates this bias entirely but is appropriate only when sufficient per-source evidence makes the individual estimates reliable. When some sources are very sparse – as can occur in voluntary contribution corpora – giving full weight to noisy per-source distributions can increase TV more than the bias reduction is worth, and UNIF performs worse than POOL.

CAP is the most robust choice across the range of conditions studied in this simulation study. Under mild imbalance, it behaves as a useful compromise: it remains close to POOL when coupling is absent and moves toward UNIF as coupling strengthens. It is the best-performing estimator under strong imbalance and strong coupling in the present grid. The reason follows from the weighting rule: by capping each source’s token contribution at a fixed ceiling, CAP directly limits the influence of large, potentially distinctive sources – the mechanism that drives POOL’s bias under coupling. It does not reduce the noise in small sources’ estimates, but at intermediate imbalance levels this cost is outweighed by the bias reduction.

The α parameter serves two distinct purposes that should not be conflated. As a sensitivity tool, sweeping α and reporting estimates across its range documents how conclusions depend on assumptions about source representativeness – this use is encouraged precisely because it makes those assumptions explicit. As a tuning device, selecting a single α to optimize performance requires assumptions about the coupling regime and can be as hard to justify as committing to one of the endpoints.

The results have clear scope conditions. The generator is a controlled abstraction; real corpora exhibit coupling patterns that are more gradual and multi-dimensional than the parametric design studied here. The WNS case study illustrates that estimator disagreement can be substantively large – $TV = 0.304$ between POOL and UNIF for emoji, and $TV = 0.138$ for filtered word tokens – but cannot adjudicate which estimator is closer to the target distribution. The unigram setting studied here is a natural starting point, but extending the same source-weighting logic to downstream sequence statistics and association measures is an important next step. The simulation also holds fixed several parameters – the number of sources S , total token count N , vocabulary size V , Zipf exponent, number of content regimes K , and degree of content heterogeneity – whose influence on estimator behavior warrants further study. Finally, CAP should be viewed as a simple baseline rather than a final estimator. Future work should explore more adaptive ways of navigating the reliability–representativeness trade-off while preserving the interpretability that makes capped weighting useful.

Data and code availability. The reproduction package for this paper is available as a GitHub repository and archived on Zenodo as version 1.0.0: <https://doi.org/10.5281/zenodo.20314185>. It contains the estimator implementations, toy data, tests, WNS preprocessing scripts for authorized local use, and the scripts used to regenerate the reported tables and simulation figures. The WNS corpus itself is not redistributed. The dataset is available on demand for research purposes, under a restricted license contract, from the SWISSUbase repository (<https://www.swissubase.ch>).

Use of AI tools. Portions of this work used AI language models (ChatGPT 5.2 to 5.5 Thinking and Claude Sonnet 4.6) for exploratory brainstorming, code development, bibliographic research, and drafting assistance. All research decisions, experimental design, result interpretation, and intellectual responsibility for the claims made in this paper remain with the author.

References

- Groves R. M. and Peytcheva E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2):167–189.
- Haziza D. and Beaumont J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2):206–226.
- Kučera H. and Francis W. N. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.
- Xanthos A., Doudot L., and Gupta P. (2025). *What’s New, Switzerland?* Collecting and sharing half a million WhatsApp messages in French. In Cotgrove L., Herzberg L., and Lungen H., editors, *Exploring digitally-mediated communication with corpora. Methods, analyses, and corpus construction*, pp. 187–206. de Gruyter, Berlin/Boston.
- Xanthos A., Gupta P., Benkais L., Doudot L., and Grütter A. (2024). *What’s New, Switzerland?* Corpus (version 1.0.0). LaRS – Language Repository of Switzerland. Data set. <https://doi.org/10.48656/pa3t-xh52>.