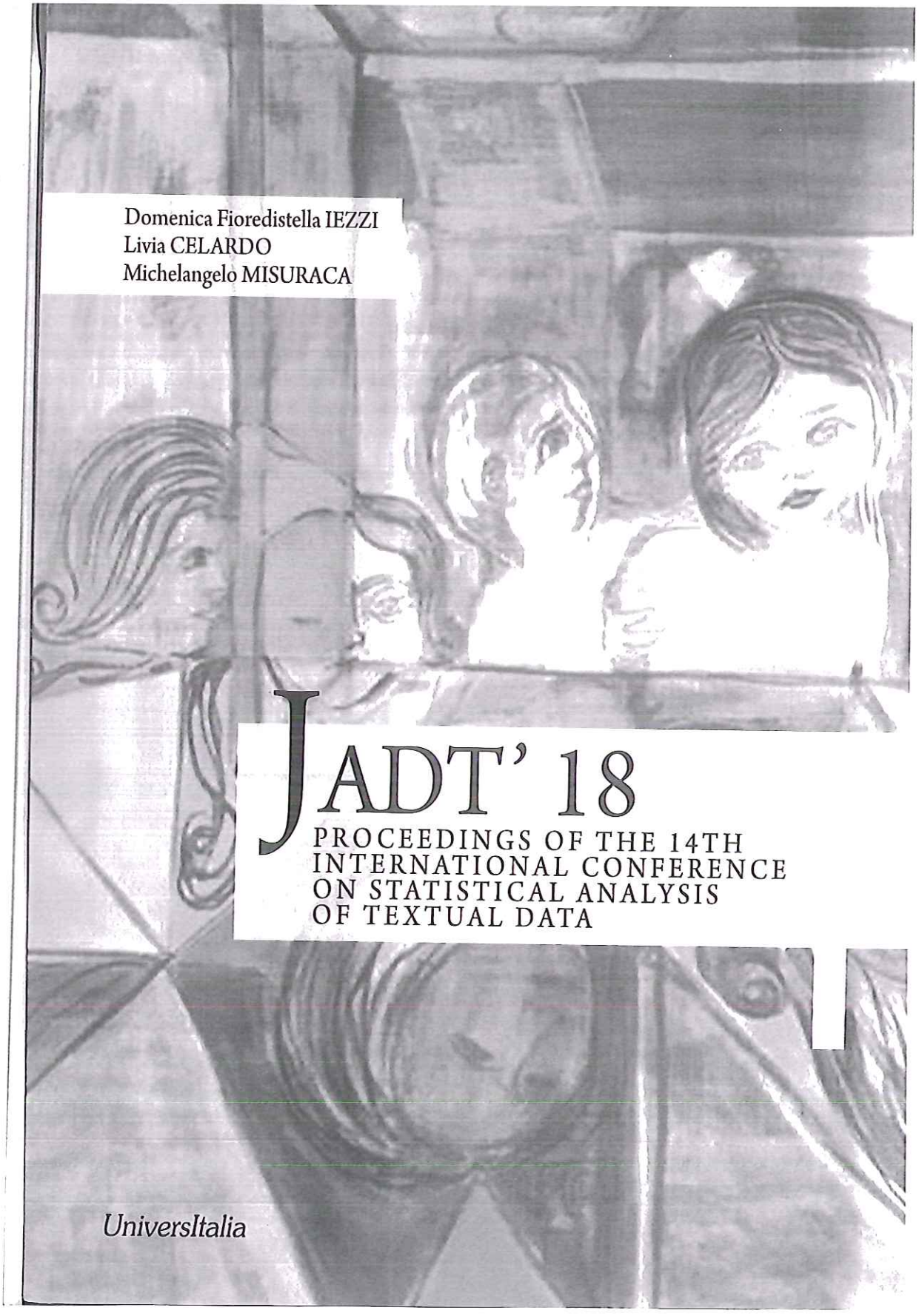


Domenica Fioredistella IEZZI
Livia CELARDO
Michelangelo MISURACA

# JADT' 18

## PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON STATISTICAL ANALYSIS OF TEXTUAL DATA

UniversItalia

quantitativa dei segmenti e dei contesti in cui il termine *vittima* compare esplicitamente o è richiamato in altro modo; con la disamina dei contesti e delle forme cui si ricorre per parlare di chi ha offeso, con l'attività social scaturita dalle cronache relative a momenti clou dell'anno in materia di violenza o di rivendicazione di genere, nello specifico nei confronti delle donne, quali la giornata contro la violenza sulle donne o l'8 marzo. Già attuata a campione, la raccolta e la successiva analisi di messaggi mostra una pervicace azione a ripetere impermeabilmente le proprie azioni comunicative, tanto nei contenuti tanto nella forma e nelle costellazioni di termini che accompagnano il focus di volta in volta oggetto di discussione. Segno inequivocabile della posizione che gli elementi da cui si irradia la costellazione stessa hanno nell'enciclopedia e nella coscienza e sensibilità della comunità linguistica italofona.



# Il *cosa* e il *come* del processo narrativo. L'uso combinato della Text Analysis e Network Text Analysis al servizio della precarietà lavorativa

Cristiano Felaco[1], Anna Parola[2]
Università degli Studi di Napoli Federico II – cristiano.felaco@unina.it; anna.parola@unina.it

**Abstract**
This paper shows the analytic procedures in order to use jointly Text Analysis and Network Text Analysis. Text Analysis allows to detect the main themes subjects in the narrations and hence the processes of signification, Network Text Analysis permits to track down the relations between linguistic expressions of text, identifying therefore the path of flow of thoughts. Using jointly the two methods is possible not only to explore the content of narrations, but, starting from the words and concepts with higher semantic strength, also to identify the processes of signification. To this purpose, we will present a research aiming to understand high school students' perception of employment precariousness in Italy. The lexical corpus was built by narrations collected from 2013 to 2016 in blog of Repubblica "Microfono Aperto".

**Riassunto**
Il lavoro presenta le procedure analitiche per un uso congiunto delle tecniche di Text Analysis e Network Text Analysis. La prima permette di cogliere i temi principali affrontati nelle narrazioni e quindi i processi di significazione, la seconda di rintracciare le relazioni tra le espressioni linguistiche di un testo, individuando i percorsi dei flussi di pensiero. L'uso combinato delle due tecniche permette, dunque, non solo di esplorare i contenuti delle narrazioni, ma, lavorando su parole e concetti con una maggiore carica semantica, anche di ricostruire i percorsi attraverso i quali si costruisce il significato. A tale scopo sarà presentata una ricerca volta a comprendere la percezione degli studenti delle scuole secondarie superiori sulla precarietà lavorativa in Italia. Il corpus testuale è stato creato a partire dalle narrazioni raccolte dal 2013 al 2016 nel blog di Repubblica "Microfono Aperto".

**Keywords:** Thematic Analysis of Elementary Contexts; Network Text Analysis; Employment Precariousness; Students.

## 1. Introduzione

La narrazione, e più nello specifico il narrare, è un processo di costituzione di una tessitura testuale dotata di senso e veicolante significati. Analizzare i testi permette di cogliere da un lato la percezione di chi narra su un dato argomento e il processo di significazione attribuita all'esperienza narrata, ma dall'altro di comprendere i flussi di pensiero, entrando nello specifico delle parole utilizzate e della loro sequenzialità. L'uso della statistica testuale al servizio delle narrazioni permette, perciò, il riconoscimento in profondità del significato delle parole e del senso ivi presente (Bolasco, 2005). Tra le tecniche di analisi del contenuto, l'uso combinato della Text Analysis (TA) e Network Text Analysis (NTA) si presta bene a questi scopi. Se la TA permette di cogliere i temi affrontati, le parole scelte e utilizzate e le dimensioni di senso attribuite (Lebart et al., 1998), il *cosa* si narra, l'uso della TNA offre un ulteriore approfondimento sul *come* si narra. Analizzando, infatti, la posizione delle parole all'interno della rete testuale è possibile rintracciare le parole con una maggiore carica semantica, individuando in questo modo i diversi percorsi e contesti di significato (Hunter, 2014) mediante lo studio della natura delle relazioni tra i vari termini. Partendo dall'assunto che la struttura di relazioni tra le parole di un testo possa corrispondere ai modelli mentali e alle mappe cognitive messe in atto dagli autori del testo (Carley, 1997; Popping et Roberts, 1997), tale metodo permette di modellizzare il linguaggio come rete di parole e di relazioni attraverso la creazione di una mappa cognitiva (Popping, 2000). Il concetto è il nucleo (mentale) che viene rappresentato attraverso un termine o un'espressione linguistica; i termini possono essere in relazione tra loro formando un'affermazione. Le affermazioni che condividono uno stesso concetto formano una struttura interdipendente creando così una mappa concettuale o rete testuale costituita da punti (o nodi) che rappresentano le singole parole (o concetti) e da linee, cioè i legami che li collegano.

## 2. Metodologia

L'approccio proposto prevede dapprima che i testi prodotti siano sottoposti ad un'analisi statistica dei dati testuali servendosi del software di analisi automatica T-lab, e successivamente analizzati in una prospettiva di rete mediante il software Gephi.

### 2.1 Pre-trattamento dei testi

Raggruppati all'interno di un unico corpus, la prima fase di lavorazione del testo si compone di una fase di normalizzazione del corpus e di personalizzazione del dizionario. La prima ha l'obiettivo di riconoscere le parole come forme grafiche e ciò comporta una trasformazione del corpus



(eliminazione di spazi vuoti in eccesso, marcatura degli apostrofi, riduzione delle maiuscole), e la creazione di stringhe per le locuzioni polirematiche, insiemi di parole che hanno un significato unitario non desumibile da quello delle parole che lo compongono, arrivando alla creazione delle *multiwords*. La fase di personalizzazione del dizionario è effettuata con le procedure di lemmatizzazione e disambiguazione del testo che permettono di rinominare le forme grafiche in lemmi. Lo step della disambiguazione permette di selezionare le forme omografe per disambiguarle; quello di lemmatizzazione, partendo dal riconoscimento delle forme con la stessa radice lessicale (lessema) o appartenenti alla stessa categoria lessicale, di ricondurre ogni aggettivo e sostantivo al maschile singolare, ogni verbo alla forma di infinito presente, e così via. Terminata questa fase, si procede al controllo delle caratteristiche lessicali del corpus per comprenderne la trattabilità a livello statistico, verificando i valori del *type/token ratio*, adeguato per un valore inferiore a 0.2, e gli *hapax*, adeguato per una percentuale inferiore al 50% per corpus di grandi dimensioni, e per percentuali leggermente superiori in caso di corpus di medie o piccole dimensioni. Prima di procedere all'analisi, va, inoltre, presa visione della lista delle parole chiave, creata con una procedura automatica dal software, e alla loro occorrenza all'interno del corpus, e si fissa una soglia di occorrenza minima, escludendo dall'analisi tutte le parole presenti meno di n. volte. La scelta della soglia di occorrenza dipende dalle caratteristiche lessicali e dalle dimensioni del corpus in analisi. Le parole chiave possono dunque essere prese nella loro integrità, ridotte in relazione alla soglia di occorrenza, o ancora ulteriormente ridotte in base agli scopi della ricerca.

### 2.2. Analisi dei testi mediante Analisi Tematica dei Contesti Elementari

L'Analisi Tematica dei Contesti Elementari mediante una Cluster Analysis permette di costruire ed esplorare i contenuti del corpus in analisi (Lancia, 2004). I cluster sono costituiti da un insieme di contesti elementari definiti dagli stessi pattern di parole chiave e descritti attraverso le unità lessicali che maggiormente vanno a caratterizzare i contesti elementari. La cluster analysis è eseguita mediante un metodo gerarchico-ascendente non supervisionato (algoritmo bisecting K-means), caratterizzato dalla co-occorrenza dei tratti semantici. Nello specifico, la procedura d'analisi è costituita da: analisi delle co-occorrenze mediante la creazione di una tabella dati unità di contesto*unità lessicali con valori di presenza/assenza; pre-trattamento dei dati tramite TF-IDF e trasformazione di ogni vettore riga a lunghezza 1 (norma euclidea); uso del coseno e clusterizzazione tramite algoritmo bisecting K-means; analisi comparativa con creazione della tabella di contingenza unità lessicali*cluster; test del chi-quadrato agli incroci

cluster*unità lessicali. Rispetto al criterio di partizione che determina il numero dei cluster, viene utilizzato un algoritmo che utilizza il rapporto tra varianza intercluster e varianza totale assumendo come partizione ottimale quella in cui questo rapporto supera la soglia del 50%. L'interpretazione della posizione occupata dai cluster nello spazio fattoriale e delle parole che li caratterizzano permettono di individuare le relazioni implicite che organizzano il pensiero dei soggetti, consentendo di cogliere il punto di vista del narratore nei confronti dell'evento narrato. Quest'ultimo comprende anche una serie di elementi valutativi, riflessioni, significati, giudizi di valore, ma anche proiezioni affettive.

### 2.3. Analisi delle reti

Il secondo step d'analisi prevede l'inserimento del corpus all'interno del software Gephi. Tale software organizza i vari lemmi in una matrice di adiacenza (lemma*lemma) consentendo la creazione di una rete *1-mode*, uno strumento utile per visualizzare la struttura di relazioni tra i vari lemmi, rappresentati da cerchi o nodi, e collegati tramite legami rappresentati da linee direzionate. Tale tecnica permette di cogliere il modo con cui i nodi sono connessi tra loro, identificando così le zone di vicinato (*neighbourhood*), e individuando quei nodi che occupano una posizione di rilevanza in differenti set o nell'intero network. A tale scopo, vengono calcolate differenti misure basate sulla centralità e, tra queste, la *degree centrality* che indica le parole usate con maggiore frequenza in connessione ad altre parole all'interno delle narrazioni e nei vari contesti di significato. Più nel dettaglio, l'incidenza di ogni nodo può essere espressa sia come *in-degree*, numero di archi entranti in un punto, individuando in questo modo i cosiddetti "predecessori" di ogni unità lessicale, sia come *out-degree*, numero di archi uscenti dal punto, mostrando invece i "successori". Tale relazione tra predecessori e successori all'interno della rete testuale aiuta a comprendere la varietà semantica generata dai nodi. Altro indice utilizzato è la *betweennes centrality*, misura di centralità globale basata sulla vicinanza, che esprime il grado con cui un nodo sta "fra" gli altri nodi del grafo. I nodi collocati in queste zone del network eserciterebbero una funzione di controllo sui flussi informativi e di "passaggio" permettendo il collegamento tra due o più set del network (Freeman, 1979). Nell'ottica dell'analisi testuale, questi lemmi, infatti, giocano un ruolo centrale nella circolazione dei significati all'interno della rete, fungendo da punto di giunzione da cui si connettono zone diverse di testo e si snodano specifici percorsi di significato, andando a definire in questo modo la varietà semantica delle narrazioni.



## 3. Caso studio

Presentiamo uno studio condotto attraverso l'uso combinato delle tecniche allo scopo di comprendere la percezione degli studenti del mondo del lavoro nel contesto italiano. Gli ultimi dati disponibili mostrano che l'Italia è tra i paesi europei con il più alto tasso di disoccupazione giovanile (Eurostat, 2017). L'instabilità, la precarietà e la discontinuità delle entrate rendono i giovani vulnerabili ai cicli economici, modificando natura e tempi della transizione al mondo del lavoro e riducendo le opportunità di sviluppare soddisfacenti piani di vita (Leccardi, 2006). La sfiducia incide sui propulsori della transizione, cioè sul mantenimento di aspirazioni elevate, sulla cristallizzazione degli obiettivi di carriera e sul comportamento intensivo della ricerca di un lavoro (Vuolo et al., 2012). Per lo studio abbiamo utilizzato una fonte di dati testuali provenienti dal blog di Repubblica "Microfono Aperto" in cui studenti delle scuole superiori, nel periodo dal 2013 al 2016, hanno risposto al promt "Quattro giovani su dieci senza lavoro. E tu che pensi? Di chi sono le colpe? Cosa vorresti che venisse fatto al più presto per garantirti un dignitoso futuro?". Raccontarsi attraverso la Rete agevola il processo di riflessione su di sé, sul proprio ruolo e sul rapporto con ciò che accade nel contesto in cui il giovane è inscritto. In una situazione di malessere per la precarietà lavorativa, il web può essere un utile contenitore per la condivisione dell'esperienza di precarietà, costituendo un ambiente di condivisione e socializzazione delle proprie esperienze (Di Fraia, 2007).

### 3.1 Risultati

Il corpus conta 130 narrazioni (10110 occorrenze, 2484 forme grafiche, 1590 hapax), utilizzando come variabili descrittive la provenienza territoriale (nord, centro, sud) e il tipo di istituto frequentato (istituto tecnico-professionale e liceo) e soddisfa i criteri statistici di trattabilità. L'analisi tematica dei contesti elementari ha prodotto quattro cluster (Fig. 1; Tab. 1), rinominati CL1 "Guardare le opportunità" (14,6%); CL2 "E il governo?" (19,8%); CL3 "Dai sogni alla crisi" (38,5%); CL4 "La ricerca del lavoro, dove?" (27,1%). Le narrazioni del cluster "Guardare le opportunità" rimandano all'analisi di sacrifici e opportunità; emerge in modo marcato la necessità di una "attività", di una messa in pratica di azioni nel presente in vista di un futuro migliore. Per questo motivo, la crisi è al tempo stesso un'opportunità che i giovani devono cogliere per dimostrare le proprie capacità: *Ormai, per ciò che si sente, chiunque si chiede del proprio futuro. Per garantire che un giorno ci sia più lavoro, si deve agire ORA. [...]. Anche chi cerca lavoro, però, deve volare basso e accontentarsi, per il momento, di poco, invece di restare a casa arreso.*
*Secondo me i giovani devono avere l'opportunità di dimostrare ciò che valgono, dimostrare al mondo ciò che sanno essere e far capire a tutti che sono capaci "se si*

*impegnano" di fare qualsiasi lavoro, dal più semplice al più complesso.* I testi del secondo cluster sono maggiormente orientati alla ricerca della "colpa" e ad una richiesta di soluzioni principalmente dallo Stato: *Penso che lo Stato dovrebbe dare più spazio ai giovani assicurando loro protezione e tutela. I parlamentari devono conservare i diritti e le possibilità di ogni giovane, siamo noi il futuro di questo stato, e come tali abbiamo bisogno di opportunità.*

Il cluster "Dai sogni alla crisi" rimanda alla dimensione più interna dell'essere immersi in una società che sta attraversando un momento di crisi economica. Gli studenti rimarcano che la mancanza di lavoro annulla i sogni: *Sono davvero preoccupata, tutti noi sogniamo cosa fare da grandi e sapere che il 38,7% dei giovani non riesce a trovare lavoro mi rende indignata. I giovani sono il futuro, il progresso, si impegnano [...] Sappiamo tutti cosa dice il primo articolo della nostra splendida costituzione, eppure sembra sia ignorato. Bisogna dare più occasioni ai giovani, tenere in considerazione la nostra costituzione, per aprire le porte al futuro e rendere l'Italia migliore.* Le narrazioni dell'ultimo cluster riguardano trasversalmente tutte le difficoltà del cercare lavoro (la ricerca affannata, le aziende che non assumono a causa delle troppe tasse) e della necessità di andare all'estero: *L'Italia si ritrova in un periodo di profonda crisi e se non si riprende economicamente ridando la possibilità a noi giovani di far capire a chi di dovere che abbiamo le capacità e volontà di lavorare, l'Italia perderà tutti quei giovani ma soprattutto tutte quelle menti che andranno all'estero in cerca di condizioni di vita più favorevoli ma soprattutto di maggiori possibilità di lavoro.*

La posizione delle variabili descrittive mostra una differenza per la variabile provenienza territoriale e nessuna differenza per istituto frequentato. Se infatti il frequentare una scuola piuttosto che un'altra sembra non incidere sulla percezione del mondo del lavoro e sui vissuti di sfiducia, che sono invece comuni, l'appartenenza territoriale ha un suo peso. La modalità nord è, in termini di vicinanza, posta in prossimità dei cluster 1 e 4, il centro del 3 e il sud del cluster 2. Ciò indica come gli studenti del nord tendano maggiormente a problematizzare il fenomeno del precariato e la difficile ricerca del lavoro, mettendo anche l'accento sulle opportunità che i giovani hanno di dimostrare il proprio valore; le tematiche di quelli del sud vanno maggiormente nella colpevolizzazione del contesto, in linea con una maggiore risonanza del tema di discussione a causa di un'elevata incidenza della disoccupazione giovanile; le narrazioni degli studenti del centro, invece, maggiormente richiamano i propri vissuti interni.



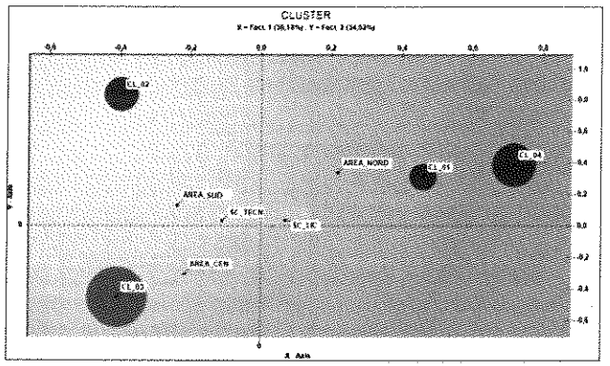
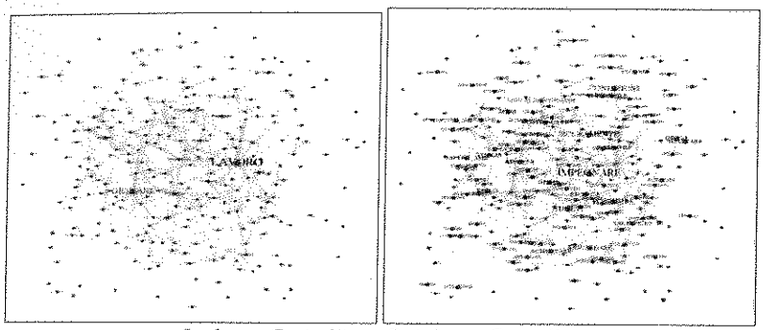
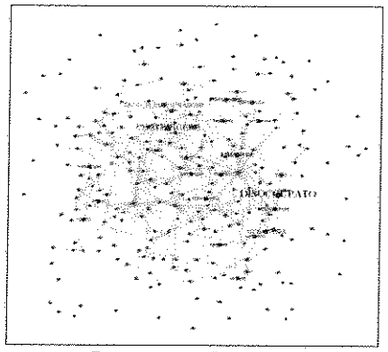



*Figura 1: Cluster Analysis*

La rete prodotta è composta da 259 nodi e 414 legami. Una prima approfondita forma di visualizzazione della struttura di relazioni tra i vari lemmi mostra i livelli più alti di degree centrality, in cui "lavoro", "giovani", "futuro", "problema" e "possibilità" rappresentano i nodi con maggiori connessioni. Inoltre, questi stessi nodi riportano anche i valori più alti di in-degree centrality, nodi "assorbenti" che presentano più legami in entrata che in uscita rispetto a tutti gli altri punti; gli studenti tendono a indirizzare i propri discorsi e, più in generale, il flusso di pensiero verso le tematiche relative al lavoro in termini sia di possibilità future sia analizzandone le problematiche ad esso legate. Dall'altro canto, "impegnare" (inteso come impegno messo in atto) e "condizioni" rappresentano il fulcro da cui muove la narrazione verso altre parole, nodi "sorgente" che hanno più legami in uscita che in entrata rispetto ai restanti nodi della rete. I lemmi che rimandano ai vissuti degli studenti, ai propri stati d'animo rispetto all'attuale condizione e ad una prospettiva lavorativa futura incerta sono quelli che giocano un ruolo centrale nella circolazione dei significati all'interno della rete, presentando difatti i valori più elevati di betweenness centrality. In particolare, "disoccupato", "costringere", "rimanere" e "scoraggiare" sono i nodi che fungono da principale punto di giunzione da cui si snodano specifici percorsi di significato: le diverse zone del network, e quindi diverse parti della narrazioni sono collegate tra loro da quei lemmi che ruotano intorno al tema della precarietà del presente, una situazione di costrizione e di forte scoraggiamento.



*In-degree Centrality* *Out-degree Centrality*



*Betweenness Centrality*

*Figura 2*

**4. Conclusioni**

L'uso misto della TA e NTA permette di rappresentare un quadro sintetico della struttura semantica, comprendere di cosa si parla, ma anche in che modo lo si fa: la scelta delle parole e l'ordine stesso di presentazione di un'idea o opinione rispetto al tema in oggetto. L'uso congiunto delle due tecniche fornisce: a) una sintesi delle informazioni contenute nelle narrazioni; b) l'analisi dei temi affrontati; c) un focus sulla strutturazione delle frasi in termini di relazioni tra lemmi. Permette così di mettere in relazione categorie tematiche e di contenuto in quanto struttura latente, ricostruendo a ritroso il processo discorsivo.

**Bibliografia**

Bolasco S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di Statistica*, vol. 7: 1-37.



Carley K.M. (1997). Extracting team mental models through textual analysis. *Journal of organizational behavior*, 18(1): 533-558.

Di Fraia G., a cura di, (2007). *Il fenomeno blog. Blog-grafie: identità narrative in rete*. Milano: Guerini e Associati.

Eurostat (2017). Statistics on young people neither in employment nor in education or training. Report.

Freeman L.C. (1979). Centrality in Social Networks Conceptual Clarification. *Social Networks*, vol. 1: 215-239.

Hunter S. (2014). A novel method of network text analysis. *Open Journal of Modern Linguistics*, vol. 4(2): 350–366.

Lancia, F. (2004). *Strumenti per l'analisi dei testi*. Milano: Franco Angeli.

Lebart L., Salem A. and Berry, L. (1998). *Exploring textual Data*. Dordrecht: Kluwer Academic Publishers.

Leccardi C. (2006). Redefining the future: Youthful biographical constructions in the 21st century. *New directions for child and adolescent development*, vol. 113: 37-48.

Popping R. (2000). *Computer-assisted Text Analysis*. London: Sage.

Popping R. and Roberts C.W. (1997). Network approaches in text analyisis. In Klar R. and Opitz O., editors, *Classification and knowledge organization*. Berlin, New York: Springer.

Vuolo M., Staff J. and Mortimer, J. T. (2012). Weathering the great recession: Psychological and behavioral trajectories in the transition from school to work. *Developmental psychology*, vol. 48(6): 1759.

The *International Conference on the Statistical Analysis of Textual Data* (JADT, Journées d'Analyse Statistique des Données Textuelles) has been at its 14th edition. It was held for the third time in Rome, from 12 to 15 June 2018, organized by the DII - Department of Enterprise Engineering "Mario Lucertini" at Tor Vergata University of Rome and the DSS - Department of Statistical Sciences at Sapienza University of Rome. This biennial conference has continuously gained importance since its first occurrence in Barcelone (1992), and with 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). Every two years, the JADT conference presented the state of the art concerning theories, problems, methods, algorithms, software and applications in several domains, sharing a quantitative approach to the study of lexical, textual, pragmatic or discursive features of information expressed in natural language.
The proceedings of **the 2018 Conference collect 114** contributions by 243 scholars from 15 countries spread all over the world. These papers include contributions open to all scholars and researchers working in the field of textual data analysis, ranging from lexicography to the analysis of political discourse, from information retrieval to marketing research, from computational linguistics to sociolinguistics, from text mining to content analysis.

**In copertina:**
*Gioco di triangoli* - 2018
Antonietta Orsatti

ISBN 978-88-3293-137-2