

ISSN 24994553

IJCoL

Italian Journal
of Computational Linguistics

Rivista Italiana
di Linguistica Computazionale

Volume 7, Number 1-2

june-december 2021

Special Issue

Computational Dialogue Modelling:
The Role of Pragmatics and Common Ground in Interaction

aA
ccademia
university
press





IJCoL
Italian Journal of Computational Linguistics

7-1, 2 | 2021

Special Issue: Computational Dialogue Modelling: The Role of Pragmatics and Common Ground in Interaction



Electronic version

URL: <https://journals.openedition.org/ijcol/792>

DOI: 10.4000/ijcol.792

ISSN: 2499-4553

Publisher

Accademia University Press

Electronic reference

IJCoL, 7-1, 2 | 2021, "Special Issue: Computational Dialogue Modelling: The Role of Pragmatics and Common Ground in Interaction" [Online], Online since 01 December 2021, connection on 26 December 2021. URL: <https://journals.openedition.org/ijcol/792>; DOI: <https://doi.org/10.4000/ijcol.792>



IJCoL is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

editors in chief

Roberto Basili

Università degli Studi di Roma Tor Vergata

Simonetta Montemagni

Istituto di Linguistica Computazionale “Antonio Zampolli” - CNR

advisory board

Giuseppe Attardi

Università degli Studi di Pisa (Italy)

Nicoletta Calzolari

Istituto di Linguistica Computazionale “Antonio Zampolli” - CNR (Italy)

Nick Campbell

Trinity College Dublin (Ireland)

Piero Cosi

Istituto di Scienze e Tecnologie della Cognizione - CNR (Italy)

Giacomo Ferrari

Università degli Studi del Piemonte Orientale (Italy)

Eduard Hovy

Carnegie Mellon University (USA)

Paola Merlo

Université de Genève (Switzerland)

John Nerbonne

University of Groningen (The Netherlands)

Joakim Nivre

Uppsala University (Sweden)

Maria Teresa Paziienza

Università degli Studi di Roma Tor Vergata (Italy)

Hinrich Schütze

University of Munich (Germany)

Marc Steedman

University of Edinburgh (United Kingdom)

Oliviero Stock

Fondazione Bruno Kessler, Trento (Italy)

Jun-ichi Tsujii

Artificial Intelligence Research Center, Tokyo (Japan)

Cristina Bosco

Università degli Studi di Torino (Italy)

Franco Cutugno

Università degli Studi di Napoli (Italy)

Felice Dell'Orletta

Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)

Rodolfo Delmonte

Università degli Studi di Venezia (Italy)

Marcello Federico

Fondazione Bruno Kessler, Trento (Italy)

Alessandro Lenci

Università degli Studi di Pisa (Italy)

Bernardo Magnini

Fondazione Bruno Kessler, Trento (Italy)

Johanna Monti

Università degli Studi di Sassari (Italy)

Alessandro Moschitti

Università degli Studi di Trento (Italy)

Roberto Navigli

Università degli Studi di Roma "La Sapienza" (Italy)

Malvina Nissim

University of Groningen (The Netherlands)

Roberto Pieraccini

Jibo, Inc., Redwood City, CA, and Boston, MA (USA)

Vito Pirrelli

Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)

Giorgio Satta

Università degli Studi di Padova (Italy)

Gianni Semeraro

Università degli Studi di Bari (Italy)

Carlo Strapparava

Fondazione Bruno Kessler, Trento (Italy)

Fabio Tamburini

Università degli Studi di Bologna (Italy)

Paola Velardi

Università degli Studi di Roma "La Sapienza" (Italy)

Guido Vetere

Centro Studi Avanzati IBM Italia (Italy)

Fabio Massimo Zanzotto

Università degli Studi di Roma Tor Vergata (Italy)

Danilo Croce

Università degli Studi di Roma Tor Vergata

Sara Goggi

Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR

Manuela Speranza

Fondazione Bruno Kessler, Trento

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)
© 2021 Associazione Italiana di Linguistica Computazionale (AILC)



Associazione Italiana di
Linguistica Computazionale



direttore responsabile
Michele Arnese

isbn 9791280136770

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_7_1-2



Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

Computational Dialogue Modelling: The Role of Pragmatics and Common Ground in Interaction

Invited editors: *Hendrik Buschmeier and Francesco Cutugno*
co-editors: *Maria Di Maro and Antonio Origlia*

CONTENTS

Editorial Note <i>Francesco Cutugno, Hendrik Buschmeier</i>	7
Knowledge Modelling for Establishment of Common Ground in Dialogue Systems <i>Lina Varonina, Stefan Kopp</i>	9
Pragmatic approach to construct a multimodal corpus: an Italian pilot corpus <i>Luca Lo Re</i>	33
How are gestures used by politicians? A multimodal co-gesture analysis <i>Daniela Trotta, Raffaele Guarasci</i>	45
Toward Data-Driven Collaborative Dialogue Systems: The JILDA Dataset <i>Irene Sucameli, Alessandro Lenci, Bernardo Magnini, Manuela Speranza e Maria Simi</i>	67
Analysis of Empathic Dialogue in Actual Doctor-Patient Calls and Implications for Design of Embodied Conversational Agents <i>Sana Salman, Deborah Richards</i>	91
The Role of Moral Values in the Twitter Debate: a Corpus of Conversations <i>Marco Stranisci, Michele De Leonardis, Cristina Bosco, Viviana Patti</i>	113
Computational Grounding: An Overview of Common Ground Applications in Conversational Agents <i>Maria Di Maro</i>	133
Cutting melted butter? Common Ground inconsistencies management in dialogue systems using graph databases <i>Maria Di Maro, Antonio Origlia, Francesco Cutugno</i>	157
Towards a linguistically grounded dialog model for chatbot design <i>Anna Dell'Acqua, Fabio Tamburini</i>	191
Improving transfer-learning for Data-to-Text Generation via Preserving High-Frequency Phrases and Fact-Checking <i>Ethan Joseph, Mei Si, Julian Liaonag</i>	223

Introduction to the Special Issue on Computational Dialogue Modelling

Francesco Cutugno*
Università di Napoli 'Federico II'

Hendrik Buschmeier**
Bielefeld University

This special issue on 'Computational Dialogue Modelling' discusses recent approaches for modelling pragmatics and common ground in spoken human-human and human-machine interaction. Natural Language Processing (NLP), given the most recent scientific discoveries in the area of intelligent systems and distributed semantics, is now able to build interactive agents whose performance is getting more powerful from year to year. Simple 'command-based' models and dialogue state tracking methods are now widely available for very constrained tasks and domains and research in NLP is heading towards the design of more complex scenarios that need to take into account the role of pragmatics in dialogue systems as well as of grounding and common ground.

To give an example, investigating the role of pragmatics for dialogue systems enables them to adopt linguistic strategies that make internal states more observable. For example, inconsistencies between a conversational agent's internal representations and perceived contextual evidence can become visible by adopting specific superficial forms. Applying computational models of pragmatics to dialogue system design requires more complex approaches to knowledge representation and inference modelling, which can then enable conversational agents to use context-specific clarification strategies, allowing systems to be more robust as inconsistencies can be addressed efficiently. For the future, the field aims at surpassing the basic, task-oriented, applications in order to deal with more complex dialogue situations using more general structures. Managing conflicts in the internal representations among agents involved in the interaction – both human and artificial – is of critical importance for a more natural, acceptable user experience when interacting with technical systems with natural language-based interfaces.

This is a special double issue which collects ten contributions. It opens with an invited contribution by Kopp and Varonina, who provide a survey on the main theme of the issue: theories and models of Common Ground establishment in dialogue systems. The article is accompanied by a light debate starting from a number of questions raised by Cutugno and Di Maro, followed by a reply of the authors. The remaining nine articles are organised in two sections: 'Corpus Studies' and 'Technological Models'.

The five articles in the *Corpus Studies* section deal with the collection, description, or analysis of various corpora, ranging from multimodal interactions to sentiment analysis, from morality to proactivity, and grounding. Although some aspects of pragmatics have already received more attention than others, all the phenomena investigated are essential for future applications showing human-like communicative competence. For

* URBAN/ECO Research Center, University of Naples 'Federico II', Naples, Italy,
E-mail: cutugno@unina.it

** Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany,
E-mail: hbuschme@uni-bielefeld.de

such systems, large data collections are still needed – especially for languages like Italian.

The four articles in the section *Technological Models* cover topics related to the implementation of dialogue systems based on innovative models and concepts. Topics range from the general methodology to follow when designing dialogue systems, which still is – in practice – often very task-specific, to the development of specific characteristics, like conflict detection and expression, knowledge representation, and language generation, with a specific interest on the role of common ground.

The scenarios depicted in this issue show the rising interest in new aspects in the design of dialogue systems. The NLP community is creating (for what is new) and reinforcing (for what was created in the past years) the *raison d'être* of computational pragmatics, blinking an eye, simultaneously, to new forms of knowledge representation, of artificial reasoning, and introducing new components in the design of dialogue state trackers. This research area is in continuous evolution. It interacts with new discoveries in the technical fields of machine learning, knowledge representation, and NLP in general, as well as in dialogue and interaction studies, linguistics and human-agent/robot/computer interaction. It is thus to be expected that, already in the near future, new interventions will be necessary and new threads will be opened.

Finally, we – the two guest editors of this special issue – want to thank Maria Di Maro and Antonio Origlia, the co-editors and, in many cases, *deus ex machina* of the whole operation. We also want to thank the authors for their contributions, the editors in chief and copy editors of IJCoL, the committee members of the Italian Association of Computational Linguistics, as well as the reviewers – all of whom made this special issue possible.

Knowledge Modelling for Establishment of Common Ground in Dialogue Systems

Lina Varonina*
Bielefeld University

Stefan Kopp**
Bielefeld University

The establishment and maintenance of common ground, i.e. mutual knowledge, beliefs and assumptions, is important for dialogue systems in order to be seen as valid interlocutors in both task-oriented and open-domain dialogue. It is therefore important to provide these systems with knowledge models, so that their conversations could be grounded in the knowledge about the relevant domain. Additionally, in order to facilitate understanding, dialogue systems should be able to track the knowledge about the beliefs of the user and the level of their knowledgeability, e.g., the assumptions that they hold or the extent to which a piece of knowledge has been accepted by the user and can now be considered shared. This article provides a basic overview of current research on knowledge modelling for the establishment of common ground in dialogue systems. The presented body of research is structured along three types of knowledge that can be integrated into the system: (1) factual knowledge about the world, (2) personalised knowledge about the user, (3) knowledge about user's knowledge and beliefs. Additionally, this article discusses the presented body of research with regards to its relevance for the current state-of-the-art dialogue systems and several ideal application scenarios that future research on knowledge modelling for common ground establishment could aim for.

1. Introduction: Why do we need to model knowledge in dialogue systems?

When speaking about engaging in conversations with machines, most people would refer to interactions with proprietary voice-based assistants (VA), such as Amazon Alexa, Google Assistant, Siri, etc. Since their introduction to the market in the previous decade, these systems have consistently been on the rise. According to a survey conducted in the U.S. in 2020, more than a third of the adult population of the country possesses a smart speaker (Kinsella 2020). Therefore, for many people experiences with these voice-based assistants will influence their perception and expectations with regards to language-based interactions with machines. However, despite recent advances in natural language processing (NLP) the capabilities of these wide-spread VAs to lead human-like conversations are rather limited, resulting in a mismatch between expectations and reality, which is especially prevalent among users with little technical knowledge who cannot form adequate judgements about the capabilities of the system and rely on their experiences from human-human communication when interacting with VAs (Luger and Sellen 2016). Failure to engage with a voice-based assistant in a meaningful way can cause users to change their communicative behaviour, e.g., by limiting their vocabulary

* Social Cognitive Systems Group - Inspiration 1, 33619 Bielefeld, Germany.
E-mail: lvaronina@techfak.uni-bielefeld.de

** Social Cognitive Systems Group - Inspiration 1, 33619 Bielefeld, Germany.
E-mail: skopp@techfak.uni-bielefeld.de

and simplifying utterances or reducing the interactions with the system to a range of simple tasks that the users trust the system to perform correctly (Luger and Sellen 2016). The implications of such communicative failures and lessons that can be learnt from them for the design of conversational agents is one of the research topics of the project *IMPACT*¹ (The implications of conversing with intelligent machines in everyday life for people's beliefs about algorithms, their communication behaviour and their relationship-building), a cooperation of various universities and disciplines that the authors of this paper are a part of.

It is interesting to note that some researchers reject the notion of classifying interactions with VAs as *conversations* due to their fundamental differences with *actual human conversations*. Porcheron and colleagues (Porcheron et al. 2018) discussed this idea in the context of the findings of their study on everyday use of voice-based assistants in families. For instance, they argue that the predefined request-response format of the interaction with VAs cannot be equated with interactively emerging adjacency pairs, such as question-response, that serve as the basic organisational unit of many of our everyday conversations. The responses of voice-based assistants sometimes fail to coherently follow the requests of their users, which is usually treated by the users as incorrect output, rather than a reaction of an equal conversation partner. Overall, the findings of the study suggest that smart devices with voice-based assistants are not treated as interlocutors by family members, even though the interactions with them are embedded in conversational situations within a family.

These differences in treatment were also seen in open-question interviews about the nature of conversations conducted by Clark and colleagues (Clark et al. 2019). While the interviewees acknowledged the importance of similar concepts in both human-human and human-agent conversations, they operationalised them differently, as conversations with humans were characterised to have both social and transactional purposes, but descriptions of conversations with agents (which were influenced by interviewees' experiences with voice-based assistants) were mainly focused on the transactional aspect. So, for example, establishing *common ground* was identified as one of the most important parts of a good conversation with other humans. However, in a human-agent setting the interviewees rather preferred to speak about *personalisation* where certain information is used by the system to tailor user experience, which, in a long-term perspective, could create an illusion of common ground between the human and the machine. The interviewees also did not view this process as co-constructed as they would the establishment of common ground in human-human communication.

While it is not necessary to strive for the ideal of "human-like" conversation in every domain where conversational systems are used, in certain use cases, it is necessary to endow these systems with qualities that would allow them to be increasingly treated as valid conversation partners by humans (cf. (Kopp and Krämer 2021)). On the one hand, these may be the use cases in which the social aspect of the conversation is of importance, e.g., in social care or robot companionship. On the other hand, even in task-oriented, primarily transactional interactions the inclusion of certain aspects of human-human conversation is needed: the one-shot request-response format of interaction currently provided by the voice-based assistants is not sufficient. Clark and colleagues (Clark et al. 2019) offer an apt goal for task-oriented conversational systems: service desk interactions between humans. In these types of conversations, the concepts of

1 <https://www.impact-projekt.de/>

common ground and facilitation of understanding for all conversation partners become crucial for successful accomplishment of tasks.

According to the definition of *common ground* as established by Clark and Brennan, it entails "mutual knowledge, mutual beliefs and mutual assumptions" (Clark and Brennan 1991, p. 222). Thus, modelling these categories in a conversational system is a prerequisite for its capability to establish and maintain common ground. Various types of knowledge can be of relevance here, e.g., knowledge about the domain, but also knowledge about the user and their beliefs, their level of expertise in the domain, known facts and possible preconceptions, as well as the ability to track how these change as the conversation progresses, what kind of knowledge becomes *grounded* and can be used for future reference.

The goal of this article is to provide an overview of current research on knowledge modelling for the establishment of common ground in dialogue systems. Roughly, it is possible to divide this body of research into three major categories based on the type of knowledge integrated into the system. Each of these has its own research focus and use cases. These categories will be discussed in the following order:

1. factual knowledge about the world,
2. personalised knowledge about the user,
3. knowledge about user's knowledge and beliefs.

This article is by no means a comprehensive collection of work on these topics, but strives to provide a basic overview of the directions state-of-the-art research takes. Additionally, the attention currently devoted to the aforementioned topics within the research community will be discussed, along with the perspectives for and the impact of the realisation of common ground on future dialogue systems.

2. Factual knowledge about the world

The first category of knowledge that can be integrated into dialogue systems is factual knowledge about the world and the elements of the so-called commonsense knowledge, e.g., information such as "*A dog has four legs*" (Zhou et al. 2018).

The emergence of data-driven neural language models in the field of machine translation inspired the creation of end-to-end dialogue systems where similar approaches could be used, which offered an alternative to the traditional multi-component dialogue systems with separate modules for natural language understanding, generation and synthesis and dialogue management (Ritter, Cherry, and Dolan 2011; Sordani et al. 2015; Serban et al. 2016; Gao, Galley, and Li 2019). These new systems, of course, had their own challenges, such as un informativeness and the lack of diversity of utterances generated, which was addressed in different areas of research. Amongst them an idea was born to introduce *knowledge-based grounding* to neural conversational systems in order to make their responses more diverse, specific and "human-like" (Han et al. 2015; Yin et al. 2016; Zhu et al. 2017; Ghazvininejad et al. 2018; Zhou et al. 2018). This type of grounding allows the dialogue system to talk about entities not seen in the training data and also reflect changes in the domain within their responses through updates of the knowledge base (Gao, Galley, and Li 2019). *Knowledge-aware* dialogue systems are applied in both open-domain as well as task-oriented dialogue.

In such systems, external collections of knowledge are usually used. These can have varying representations, e.g., as textual data or structured knowledge bases or knowl-

edge graphs. Examples of the textual data approach can be found in (Ghazvininejad et al. 2018) where the researchers used data from social networks such as Twitter and Foursquare indexed by relevant entities, or in (Dinan et al. 2019) where Wikipedia articles organised as documents structured into paragraphs and sentences were utilised. When it comes to structured knowledge bases (Han et al. 2015; Yin et al. 2016; Zhu et al. 2017; Zhou et al. 2018; Zhang et al. 2020; Wu et al. 2020), the researchers usually use large knowledge graphs that are well-established and have been maintained by the Semantic Web community throughout the years, such as the multi-language common sense knowledge graph *ConceptNet*² (Speer, Chin, and Havasi 2017) or *DBpedia*³ (Lehmann et al. 2012) that represents the information created in Wikipedia and other Wikimedia projects. Another advantage of these graphs is that they can also be connected with each other to leverage knowledge about terms and concepts across domains as part of the *Linked Data* standard (Berners-Lee 2006). The relations in such knowledge bases are typically represented by subject-predicate-object (*s, p, o*) triples, e.g., the piece of information "a puppy can become a dog" is represented in *ConceptNet* as a triple (*/c/en/puppy, /r/CapableOf, /c/en/become_dog*) where */c/en/* and */r/* are graph-specific namespaces used for distinction of identifiers.

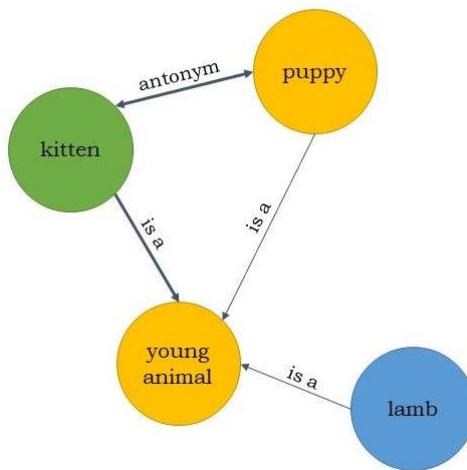


Figure 1

Part of a concept graph as defined in (Zhang et al. 2020). It is centered around the concept of *kitten* and based on relations from *ConceptNet*. Yellow-coloured nodes are one-hop concepts and the blue-coloured node is a two-hop concept.

Overall, the research on grounding of dialogues in factual or commonsense knowledge is primarily concerned with the inclusion of already available factual and commonsense knowledge bases into conversational models, i.e., selecting and extracting the knowledge relevant to the entities mentioned in user utterances and encoding this

² <http://conceptnet.io/>

³ <https://www.dbpedia.org/>

knowledge and leveraging it for response generation, not knowledge modelling in itself. These topics are outside of the scope of this paper and will not be expanded upon here. However, it is noteworthy that some approaches to knowledge integration go beyond static entity matching and acknowledge the fact that humans reference related concepts in conversations and usually shift their focus to different topics as the exchange progresses, which is modelled as *attentional state* by Grosz and Sidner (Grosz and Sidner 1986). Thus, methods are researched that would enable dialogue systems to introduce new related concepts into conversation. Consider a dialogue about *kittens* that may develop into a conversation about other *young animals*, such as *puppies* or perhaps even *lambs* if the people speaking live in the countryside. In *ConceptNet*, *young animal* and *puppy* are one-hop concepts and *lamb* is a two-hop concept with regards to *kitten* as illustrated in figure 1. It is possible to include this type of concept shift into conversational models with approaches such as the conversation generation model *ConceptFlow* that constructs a so-called *concept graph* which is a local part of the main knowledge graph centered around the currently grounded concept of the conversation (*kitten*) and extended to its one-hop (*young animal* and *puppy*) and two-hop concepts (*lamb*). This concept graph is later used for conversation modelling and response generation (Zhang et al. 2020).

3. Personalised knowledge about the user

As mentioned in the introduction, according to interviews conducted by Clark and colleagues (Clark et al. 2019), some people reject the notion of having *common ground* with machines and prefer to speak of *personalisation* in the context of human-agent conversations, i.e. the adaptation of user experience based on the information about the user collected by the system, which can create an illusion of common ground over time. However, important differences exist between the concepts of *user adaptation* or *personalisation* of dialogue systems and the establishment of *common ground* as defined by Clark and Brennan (Clark and Brennan 1991).

First, the concept of personalisation is very broad. In their survey on empathetic systems, Ma and colleagues (Ma et al. 2020) distinguish between two types of dialogue systems with regards to personalisation: *personality-aware* and *personality-infused*. While the former type only considers the personality (or certain distinct features thereof) of the *user* when composing responses, the latter type additionally infuses the *agent* with its own personality. Personality-infused systems are out of scope of this paper and will not be discussed further.

Second, personalisation is not co-constructed, as it is the system that is burdened with the collection of information about the user. Nevertheless, this collected information can be applied in the context of ensuring mutual understanding between the user and the dialogue system, e.g., by allowing the system to better understand user's intentions and react appropriately, which would be consistent with the way information exchanged for the establishment of common ground is used.

According to the aforementioned survey of Ma and colleagues (Ma et al. 2020), there are two categories of methods that can be applied to user modelling in personalised dialogue systems: *identity-based* and *knowledge-based*, while some hybrid systems also exist. Identity-based systems model the user via a set of attributes that define their basic characteristics, e.g., gender, age group, profession. The required attributes vary based on the interaction context and are oftentimes collected during the first interaction with the user. On the other hand, knowledge-based personalisation uses structured knowledge bases with facts about the user, mostly represented by subject-predicate-object triples,

and can be seen as a special case of knowledge-aware dialogue systems as described in the previous chapter. For both of these approaches, unstructured data from past interactions can also be leveraged to extract either attribute values for the identity-based models or facts to be placed in the knowledge base for knowledge-based models.

3.1 Identity-based systems

Apart from profile-building via "get-to-know" sessions during the first interaction or analysis of previously acquired interaction data, it is also possible for a dialogue system to utilise profiles of similar users in order to make assumptions about the current user. This could be beneficial for situations when the system has not yet had many interactions with the user or the profile required for personalisation is too extensive to be explicitly requested. Pei and colleagues (Pei, Ren, and de Rijke 2021), for example, propose an architecture called *Cooperative Memory Network* for this purpose, a part of which is a user profile enrichment module which maintains the profile and the dialogue memory that are represented as embeddings of the profiles of the current user and users similar to them and their dialogue history respectively. Individual user profiles are represented as numerical vectors and the utterances in dialogue history are represented as a bag-of-words. Missing values in the current profile are then inferred based on these embeddings and memory components get updated. These enriched profiles are then used to update the representation of the current user query and find the appropriate response.

Before that, Luo and colleagues (Luo et al. 2019) proposed another memory-based architecture called *Personalized MemN2N* for task-oriented dialogue systems. This architecture also leverages conversational data embeddings from similar users along with the current user profile in order to generate personalised response candidates. Of special interest here is that the researchers also use profile information to infer user preferences over entities in a knowledge base that contains facts about the task domain, e.g., whether the user would like to contact the restaurant they want to eat at via phone or social media.

3.2 Knowledge-based systems

In their position paper, Balog and Kenter (Balog and Kenter 2019) define the concept of the so-called *personal knowledge graph* (PKG), as opposed to publicly available knowledge graphs such as *DBpedia* that include knowledge about entities that are publicly significant. Despite the fact that various researchers have previously used concepts similar to a *personal knowledge base* (PKB) or graph (Kim et al. 2014; Li et al. 2014; Bang et al. 2015), Balog and Kenter (Balog and Kenter 2019) establish the key properties of PKGs and identify important research questions with regards to these. According to them, three key aspects of a PKG are:

1. inclusion of entities that are of personal interest to the user,
2. the "spiderweb" layout centered around the user,
3. possible integration with other knowledge graphs as part of the *Linked Data* idea.

The population and maintenance of these personal knowledge graphs should occur automatically, as no designated human editors exist to curate the graphs. The authors

of the article present this as one of the challenges and research questions to be explored: how to transfer the data-driven state-of-the-art neural approaches to link prediction to PKGs for which the availability of data is very limited (Balog and Kenter 2019). Previously, other approaches to personal knowledge graph population were suggested, such as the combination of support vector machines (SVM) and conditional random fields (CRF) for the classification of personal facts in dialogue data, relation extraction and subsequent slot filling to complete user-related triples that are then added to the PKG (Li et al. 2014). However, when it comes to conversational data as information source, the assertions that need to be captured in the knowledge base are rarely stated explicitly (Tigunova et al. 2019). Instead, a person who works as a teacher might often talk about school, grades and homework without explicitly saying that they are a teacher. In (Tigunova et al. 2019) a neural architecture called *Hidden Attribute Model* is presented that is trained on triples to predict scores for different objects that could complete a given subject-predicate pair by using attention both within and across user utterances, e.g., it could predict the scores for different professions X to complete the triple $(user, employedAs, X)$.

With regards to maintenance of personal knowledge graphs, it needs to be taken into consideration that PKGs are inherently more dynamic than general-purpose knowledge graphs that store information about the world and place value on established assertions that will unlikely change fast (Balog and Kenter 2019) (consider the Wikipedia-based *DBpedia* graph and the dynamics of knowledge that you can find on Wikipedia as opposed to how fast your own preferences, possessions, etc. change). To model these temporal dynamics when it comes to user-related knowledge, Kim and colleagues (Kim et al. 2014) integrate a personal knowledge base with a *forgetting model* endowing their dialogue system with a long-term memory about the user. Each entry in the PKB has two properties: retention, which models the degree of user interest in this fact, and strength, which prevents the retention value from decaying too quickly. Both of these values change over time depending on the occurrences of the respective entity in user utterances. The forgetting model used by Kim and colleagues applies Ebbinghaus's forgetting curve and spacing effect (Kim et al. 2014; Ebbinghaus 2011).

Another interesting idea that utilises a knowledge-based approach, yet concerns itself not with personalised response generation but rather with a memory-based personal question answering, is proposed in (Moon et al. 2019). In their paper, the authors represent episodic memories concerning the user, such as events they attended, as a *memory graph* consisting of the entities related to a memory connected by corresponding edges. The entities are nodes of a knowledge graph that models the related domain knowledge. An example of such a *memory graph* can be seen in figure 2. Consider that the user knows that they have once eaten at a venue in the city district Bielefeld-Mitte (the centre of the city of Bielefeld), but they do not remember in what year it was. So they can query the system with the question *When have I been to a venue in Bielefeld-Mitte?* and the system can use the proposed approach of *Memory Graph Networks* to expand memory slots with external knowledge via attention-based memory graph traversal. That way, it can eventually obtain the result that the restaurant the user has been to in 2020 for Mary's birthday is in fact located in the desired city district. The authors also mention the possibility of memory graph extraction from social media posts and tagged photo albums of a particular user.

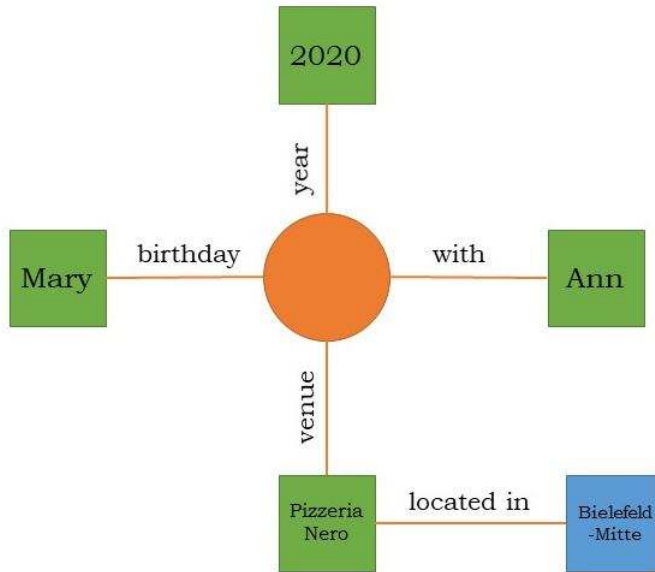


Figure 2

An episodic memory in a memory graph as defined in (Moon et al. 2019). The orange circle is the memory slot and the green squares are knowledge graph entities that are related to the memory, i.e. the birthday party of Mary in 2020 at Pizzeria Nero where the user went with their friend Ann. The blue square is the knowledge graph node that was activated after the expansion of the graph with regards to the user query *When have I been to a venue in Bielefeld-Mitte?*

4. Knowledge about user's knowledge, beliefs and mental states

The definition of *common ground* by Clark and Brennan that was mentioned throughout this paper refers to "mutual knowledge, mutual beliefs and mutual assumptions" (Clark and Brennan 1991, p. 222). However, how do the conversation partners know that they possess mutual knowledge or mutual beliefs between each other?

In an attempt to clarify the existing definitions with regards to *common ground*, Lee (Lee 2001) uses the terms of *common* and *shared knowledge* and *belief*. The author defines the concept of *common* as the information that people assume to have in common with others because of their similar background of up-bringing and the concept of *shared* as the information negotiated during a mutual interaction, while the difference between *knowledge* and *belief* lies in the certainty of truth of the information as perceived by the person. According to these definitions, in order to understand what kinds of knowledge or beliefs are common between conversation partners, they have to make assumptions about (1) the other person's background and (2) the extent to which they have understood or remembered the negotiated information. Both of these are arguably made possible by the so-called *theory of mind* (ToM).

4.1 Modelling knowledge about beliefs for ToM in human-agent interaction

One could define a theory of mind as "a basic cognitive and social characteristic that enables us to make conjectures about each others' minds through observable or latent behavioural and verbal cues" (Wang et al. 2021, p. 2). These conjectures allow humans to act accordingly in order to lead successful conversations and collaborations with others. The concept of theory of mind was also adapted for the design of human-agent interactions (Krämer, Rosenthal-von der Pütten, and Eimler 2012), primarily in the area of robotics and task-oriented collaboration (Wang et al. 2021; Scassellati 2002; Peters 2005; Devin and Alami 2016; Dissing and Bolander 2020), as perception and sensory-motor expression such as gestures are a part of the framework of ToM (Baron-Cohen 1995). Studies show that implementing ToM in robots leads to positive effects such as reduction of unnecessary communication during collaborative tasks (Devin and Alami 2016) or the perception of robots as more intelligent and natural in interaction (Hiatt, Harrison, and Trafton 2011).

As voice-based assistants fail in dialogues beyond one-shot interactions, there is a growing need and motivation to adapt aspects of the ToM concept for conversational assistants (Wang et al. 2021; Kopp and Krämer 2021). Existing neural models for question answering do not succeed at false-belief tasks, such as the classic *Sally-Anne-Experiment* (Baron-Cohen, Leslie, and Frith 1985), as was shown in an article by Nematzadeh and colleagues (Nematzadeh et al. 2018), where the researchers created a dataset of tasks that can be used for the evaluation of question answering neural models (such as memory networks, the examples of which were shown in chapter 3.1) with regards to belief reasoning. They tested several of such models and found that they make reasoning mistakes in false-belief tasks due to not having the ability to track mental states of agents that are inconsistent with the state of the real world. This might be a potential motivation to develop models that can explicitly incorporate theory of mind in conversational contexts.

Different approaches to the implementation of theory of mind in artificial agents exist and a brief overview will be given in this subsection. In general, one could divide the existing approaches into three groups:

1. models based on logic and symbolic reasoning,
2. probabilistic models, and
3. models based on machine learning.

With regards to the first group, one example could be the work of Devin and Alami (Devin and Alami 2016) that deals with the execution of shared plans in human-robot teams. Their proposed architecture features a ToM manager that maintains the mental state of the robot and other agents. The mental state is defined as (1) a set of facts about the current world state, (2) the state of current goals, (3) the state of plans, and (4) the state of actions, all of these from the respective agent's perspective. The states 2-4 denote, e.g., whether the current goal is achieved or whether a certain action has already been requested. The ToM manager then can utilise a symbolic reasoning process to make assumptions from and update the mental states of the agents.

The approach of Dissing and Bolander (Dissing and Bolander 2020) advocates for the usage of *dynamic epistemic logic* (DEL) (Bolander 2018) for theory of mind models in order to facilitate higher-order belief attribution, i.e. beliefs of agents about other agents' beliefs, as opposed to first-order belief attribution, i.e. beliefs of agents about the state

of the world. Their system maintains an epistemic state consisting of a representation of the actual world and an epistemic model over a set of possible worlds. This epistemic state is updated based on a set of rules when new actions take place in the context and can be queried with regards to a belief by using speech. The robot endowed with this approach was able to successfully pass first- and second-order false-belief tasks in an experimental setting. However, the authors state some of the limitations of their approach, e.g., the assumption that the robot is considered omniscient and cannot have false beliefs on its own or their model not accounting for agent intentions, which would be an important aspect while applying theory of mind to conversational scenarios.

In terms of probabilistic models, the most prominent one is arguably the *Bayesian theory of mind* (BToM). This approach views mentalising about the mental state of the other as Bayesian inference of the agent's hidden mental state given their behaviour in a specific context (Baker et al. 2017). The candidate mental states are defined by the agent's beliefs and desires. The beliefs are hereby represented as a probability distribution over world states in all possible worlds and their update is modelled as rational Bayesian state estimates given what was perceived by the agent and their prior beliefs. The agent's desires are represented by a utility function over situations and possible actions. The BToM adds a prior over these candidate mental states in form of a probability distribution.

BToM models can get very complex depending on the scenario they are used in. To integrate the Bayesian ToM into social agents acting in contexts when quick reaction is of importance, e.g., conversational situations, it could be beneficial to simplify these models in a way that would retain sufficient accuracy, while producing reasonable costs. In their work, Pöppel and Kopp (Pöppel and Kopp 2018) investigate the potential to simplify BToM models based on various sets of assumptions about uncertainties the acting agent faces in the environment. This results in specialised models matching a specific type of uncertainty. However, they also propose a combination model capable of switching between these specialised models according to the metric of surprise which describes how well the current model explains the behaviour the agent is observing. The authors have tested their approach, comparing the simplified models, the full BToM model and the combination model by applying them to inferences over human behavioural data in situations with various degrees of uncertainty. This data was collected by letting participants complete a set of maze traversal tasks in different uncertainty conditions, e.g., uncertainty about the structure of the maze. The results show that simplified specialised models have the ability to perform both well and badly depending on the condition they were applied to, thus leading to the necessity of the flexible combination model that achieved best performance across conditions, and, importantly, in a short enough time to facilitate online behaviour evaluation, unlike the full BToM model (Pöppel and Kopp 2018).

Lastly, machine learning methods started being involved in the implementation of ToM in agents in recent time to forgo the explicit modelling of mental states and beliefs. A prominent work here is the concept of *machine theory of mind* pioneered in the article by Rabinowitz and colleagues (Rabinowitz et al. 2018) who consider the construction of a theory of mind as a meta-learning problem. Here, in a sequence of training episodes an observer gets a set of behavioural data for a novel agent in order to make predictions about their future actions. As training progresses, the observer should learn to make better predictions about new agents from the limited set of data it receives. The architecture proposed for the observer contains three neural networks: a character net, a mental state net and a prediction net. The character net parses the historical behavioural data of the agent into a character embedding, while the mental

state net creates an embedding of their mental state based on the agent's behavioural data from the current episode. Both embeddings are then given to the prediction net to form predictions over possible next steps of the agent.

In many papers in this subsection, when it comes to theory of mind, agents usually exist in the real physical world and can observe this world and the actions carried out in it in order to update the state of the world and the mental states of others. Alternatively, it can also be an artificial world that is analogous to the real world by virtue of having specific rules and laws, and the agents in the scenario at hand act within the confines of this world. However, when it comes to social interaction, it might not be enough to update mental states based on explicit actions of others in the world. People can change their mental state because of dialogues they have with others and it is important for conversational agents to be able to capture that as well (Kopp and Krämer 2021).

Qiu and colleagues (Qiu et al. 2021) have recently introduced a *hybrid mental state parser* that can transform both continuous dialogue data and discrete action data into a graphical representation of agent's beliefs about their environment and other agents in it. Their work is based on the research of Adhikari and colleagues (Adhikari et al. 2020) who developed a graph-aided transformer agent that is capable of learning to construct and update a graph representing their beliefs about the environment of a text-based game in an end-to-end fashion from textual data by using a combination of reinforcement and self-supervised learning. Inspired by this approach and aiming to design a method that can construct belief representations from dialogue data, Qiu and colleagues (Qiu et al. 2021) also situate their agent in a text-based game (however, this type of game additionally allows dialogues between players) and apply a graph-based representation of agent's beliefs in their system. In the *belief graph*, all agents and objects along with their descriptions are represented as nodes and the belief of the agent about the current state of the environment is represented in edges that define relations between the entities and can have varying strengths. The vocabulary of entities and relation types is known in this domain by virtue of it being a game. The topology of the graph is, however, unknown and needs to be learned by the agent. It is updated as new actions and dialogue history are observed. Discrete actions carried out in the game, e.g., *put* or *give*, can be mapped onto combinations of graph update operations to add or remove specific edges in the graph. Meanwhile, continuous dialogue data is used to update the graph via a recurrent neural network.

4.2 Modelling knowledge about knowledge and beliefs in dialogue systems

One important area for belief modelling in dialogue systems are argumentative dialogues, as accounting for the perspectives of those engaged in an argument is crucial here. Additionally, a lot of uncertainty exists in this type of dialogue: with regards to beliefs of your conversation partner, the completeness of information known to them and the extent of their rationality, as well as with regards to the strength of own arguments and their influence on the beliefs of the other. Hunter and colleagues (Hunter, Polberg, and Thimm 2020) aim to create a new formalism for argumentation dialogues and reasoning that could provide solutions to these challenges: the *epistemic graph*. They describe an epistemic language that can be used to define logical formulae to specify belief in arguments and relations between them given a directed argument graph, e.g., as seen in figure 3.

The beliefs of the agent are represented with probabilities: an agent believes a term (a propositional formula of an argument) to some degree if its probability is higher than 0.5, disbelieves it to some degree if its probability is lower than 0.5 and neither believes

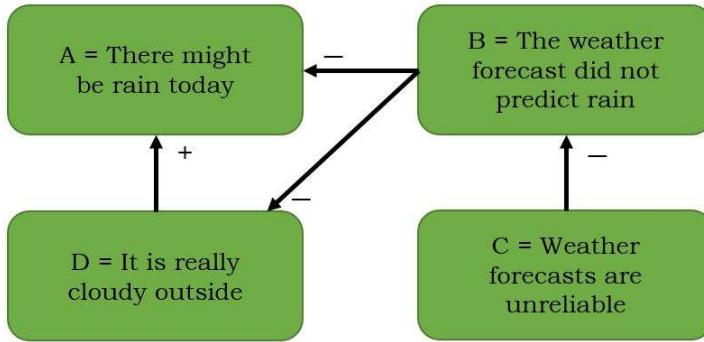


Figure 3

An example of an argument graph. Edges labelled with - and + represent attack and support respectively.

nor disbelieves it if its probability is equal to 0.5. These belief probabilities can later be used to form constraints that reflect complex beliefs, perspectives and choices. These constraints can then be reasoned with based on the logical framework developed and proved by the authors (Hunter, Polberg, and Thimm 2020). Many directions for future work are also proposed, oriented towards a practical application of epistemic graphs, for example, in computational persuasion, amongst them collection of constraints in a data-driven fashion by applying machine learning methods to crowd-sourced data on beliefs in arguments, or development of methods for belief updates during dialogues.

Another graph-based model of reasoning are *Bayesian networks*, where every node is a random variable representing a proposition and edges express statistical dependencies of one variable on another. These influences can change the belief in the target node either in a positive or a negative way, which makes Bayesian networks similar to epistemic graphs.

Bayesian networks are also used in dialogue systems when it comes to mental state representation. Buschmeier and Kopp (Buschmeier and Kopp 2011) describe the so-called *attributed listener state* (ALS) that is the assumption the speaker forms over the mental state of the listener with regards to basic communicative functions according to listener's communicative feedback the speaker receives. For example, from the listener's feedback the speaker can infer their level of understanding and form a belief about it or, more specifically, a belief about the listener's perception of their own current level of understanding. However, as this mental state attribution process is subject to uncertainty, it is necessary to understand the speaker's belief about listener's mental states in terms of their subjective *degree of belief*, i.e. the subjective confidence that this belief holds true at a given point in time, which is modelled as a probability. From this, the speaker's belief state about the listener's mental state can be defined in terms of their degree of belief in all possible worlds (Buschmeier and Kopp 2012).

Overall, the attributed listener state is modelled as a set of five discrete random variables representing the graded beliefs of the speaker about five aspects of the listener's mental state, namely, (1) them being in contact with the speaker, (2) them being able to perceive, (3) understand, (4) accept and (5) agree with the speaker. The interactions between these random variables in the ALS could be expressed with a joint

probability distribution, however, due to independence assertions for these variables it is possible to represent them in terms of five conditional distributions which is a much simpler representation that can also be expressed in terms of a graphical probabilistic model which would allow reasoning with the resulting data structures: the *Bayesian network* (Buschmeier and Kopp 2012). In fact, as seen in figure 4, *attributed listener state* is a sub-network of the larger Bayesian network of the listener where it mediates between the *conversational context* and the *information state* of the dialogue. The conversational context consists of fully observable variables, some of which are inferred from listener’s feedback, such as modality, and abstract concepts, such as difficulty of the speaker’s utterance. On the other hand, the information state (IS) of the dialogue denotes the level of grounding in the current conversation.

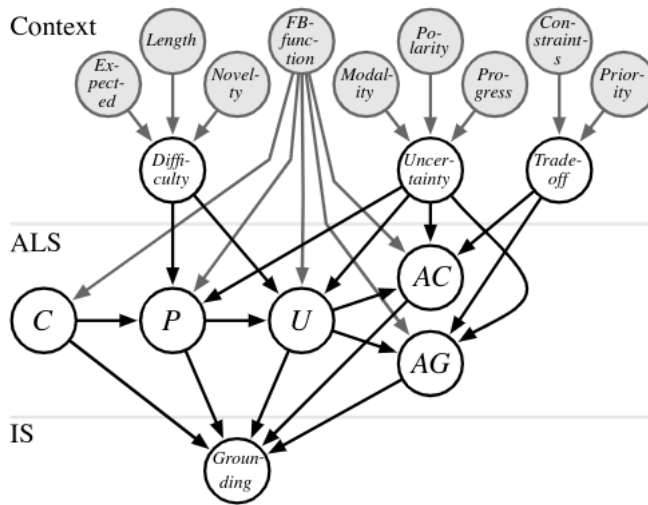


Figure 4

Structure of the Bayesian model of the listener. The variables shaded in grey are fully observable to a speaker (*FB function, modality, polarity, and progress* are derived from the listener’s feedback signal). Source of the picture: (Buschmeier and Kopp 2012).

The ALS model was later made part of the *attentive speaker agent* that is able to adapt its communicative behaviour based on user feedback. A study was conducted investigating the willingness of human listeners to provide communicative feedback in an interaction with an *attentive speaker agent* and the ability of these humans to notice the collaborative communicative behaviour of the agent (Buschmeier and Kopp 2018). In the study, the observation of the properties of feedback was done by a human who entered the corresponding context values (cf. figure 4) into the system that autonomously interpreted this feedback and in turn adapted its own communicative behaviour, including elicitation of feedback from the listener with verbal and non-verbal cues. Also, two baseline conditions were added in which the agent did not analyse the feedback, but followed a fixed strategy instead: to either always ask for feedback after presenting a unit of information or not ask at all. In general, the findings show that the participants provided feedback to the *attentive speaker agent* in a form similar to human-human interaction and stopped providing feedback to the agents that did not analyse it. Additionally, the participants recognised the attentiveness and adaptiveness of the

attentive speaker agent and also of the agent that was constantly requesting their feedback, yet only the former was ascribed a desire to be understood and helpful.

5. Discussion

In this paper we have presented an overview of methods for knowledge modelling for the representation of common ground in artificial conversational agents. Three categories of knowledge with corresponding representation formalisms were discussed: factual knowledge about the world, personalised knowledge about the user and knowledge about user's knowledge, beliefs and mental state, each of these serving its unique purpose in various types of dialogue systems. Knowledge-awareness in general allows these systems to generate more informative and helpful responses.

With the emergence of neural conversational models, many researchers in the field of dialogue systems have abandoned the classical plan-based approach to dialogue management in favor of the data-driven approach. However, as was mentioned in chapter 4.1, even advanced state-of-the-art neural networks lack the necessary representations of mental states, which results in them struggling with false-belief tasks. These representations would also allow conversational systems to establish *common ground* with the users, to model what knowledge can be considered *shared* throughout the process of interaction. Surprisingly, the availability of novel research on this topic is rather low when it comes to conversational agents and many approaches are dating back to the older plan-based systems (cf. (Kopp and Krämer 2021)).

Concepts such as theory of mind are mostly adapted for human-robot teams and are heavily grounded in observations about the world which conversational agents might not have direct access to, only learning about it indirectly through the information exchanged with the user. So if the robotic theory of mind cannot be transferred to the conversational domain one-to-one, special adaptation is necessary, as the representation of mental states and beliefs is very important for dialogue. Additionally, one needs to consider the aspect of interactivity of dialogues. If dialogue systems are supposed to make inferences about the user's mental state and beliefs by using theory of mind models, these are required to be efficient enough to be deployed online, while maintaining reasonable accuracy. The approach of Pöppel and Kopp (Pöppel and Kopp 2018) described in chapter 4.1 could be beneficial in this case, however, one needs to account for the complexity of conversational tasks. This complexity makes it challenging to identify properties of the task that can serve as the adequate basis for the creation of simplified specialised ToM models which then could be integrated into the combination model able to switch between them in order to best explain the observed behaviour.

In chapter 4.2, the domain of argumentative dialogues presented conversational scenarios where it is crucial to be able to recognise and understand the perspective of others. However, perhaps a more general and more sought-after domain where perspectives also play a major role are explanation dialogues.

Explainable AI is on the rise now and researchers argue for the social nature of explanations (Miller 2019) that should not be ignored. Explanations of the same machine learning algorithm provided to an AI expert, an elderly person with no technical experience, and a 30 y.o. technology enthusiast with a smart home would all be different. These differences can apply not only to the vocabulary used, but also potentially to dialogue structure. Consider delivering the explanation to the expert in one long turn, or allowing the technical enthusiast more room to chime in with "what-if" questions, or asking the elderly person for more feedback to ensure understanding. Ideally, in order to make these explanations different, the system needs to not only have a good factual

model of their explanandum, which was talked about in chapter 2, and not only to know the user and their personality, for which a plethora of methods were discussed in chapter 3, but also to know their mental state, to know what users believe and be able to reason about the dynamics of belief updates during the process of explanation, to know what sort of knowledge has been negotiated enough to be considered shared and can be freely referenced in the future. These are valid research areas that can be tackled, and the methods described in this paper can build a foundation for the discovery of further approaches.

To close the loop with the introduction to this paper, let us consider how voice-based assistants such as Alexa and Google Assistant could be enhanced with mental state modelling. One of the findings of the study by Porcheron and colleagues (Porcheron et al. 2018) was that the families that used smart speakers embedded them in conversational situations within the family, yet ultimately did not recognise them as interlocutors. However, if a voice-based assistant was able to maintain mental models for all family members, and to bring this to bear recognisably in dialogue, they would be able to actively participate in those conversational situations. Further, they would become able to cooperatively resolve communicative issues, for example, in case of misunderstanding or present family members having conflicting goals and desires with regards to the way they wish to use the voice-based assistant.

It would be very interesting to study how the VA would be perceived in such a case and what new group dynamics would emerge during interactions.

Acknowledgments

This research is partially supported by the Volkswagen Foundation that is funding the project IMPACT. Project IMPACT is a cooperation of several research groups belonging to University of Duisburg-Essen, Bielefeld University, University of Kassel and Lutheran University of Applied Sciences in Nuremberg.

References

- Adhikari, Ashutosh, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupard, Jian Tang, Adam Trischler, and William L. Hamilton. 2020. Learning Dynamic Belief Graphs to Generalize on Text-Based Games. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, December.
- Baker, Chris L., Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(64).
- Balog, Krisztian and Tom Kenter. 2019. Personal Knowledge Graphs: A Research Agenda. In *Proceedings of the 2019 ACM SIGIR Conference on Theory of Information Retrieval (ICTIR 2019)*, Santa Clara, CA, USA, October.
- Bang, Jeesoo, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. 2015. Example-based Chat-oriented Dialogue System with Personalized Long-term Memory. In *Proceedings of the 2015 International Conference on Big Data and Smart Computing (BIGCOMP 2015)*, pages 238–243, Jeju City, South Korea, February.
- Baron-Cohen, Simon. 1995. *Mindblindness*. MIT Press, Cambridge, MA, USA.
- Baron-Cohen, Simon, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21:37–46.
- Berners-Lee, Tim. 2006. Linked data. Online: <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed on 06.05.2021.
- Bolander, Thomas. 2018. Seeing Is Believing: Formalising False-Belief Tasks in Dynamic Epistemic Logic. In van Ditmarsch H. and Sandu G., editors, *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, volume 12 of *Outstanding Contributions to Logic*. Springer, Cham.
- Buschmeier, Hendrik and Stefan Kopp. 2011. Towards Conversational Agents That Attend to and Adapt to Communicative User Feedback. In *Proceedings of the 11th International Conference*

- on *Intelligent Virtual Agents (IVA-2011)*, pages 169–182, Reykjavik, Iceland, September.
- Buschmeier, Hendrik and Stefan Kopp. 2012. Using a Bayesian Model of the Listener to Unveil the Dialogue Information State. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*, pages 12–20, Paris, France, September.
- Buschmeier, Hendrik and Stefan Kopp. 2018. Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden, July.
- Clark, Herbert H. and Susan E. Brennan. 1991. Grounding in Communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*. American Psychology Association, Washington, D.C., pages 222–233.
- Clark, Leigh, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, Glasgow, Scotland, UK, May.
- Devin, Sandra and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot-Interaction (HRI 2016)*, Christchurch, New Zealand, March.
- Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA, May.
- Dissing, Lasse and Thomas Bolander. 2020. Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic. In *Proceedings of the 29th International Joint Conference of Artificial Intelligence (IJCAI-20)*, Yokohama, Japan, January.
- Ebbinghaus, Hermann. 2011. *Memory: A contribution to experimental psychology*. Martino Fine Books, Eastford, CT, USA, reprint of 1913 edition.
- Gao, Jianfeng, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI: Question Answering, Task-Oriented Dialogues and Social Chatbots. *Foundations and Trends © in Information Retrieval*, 13(2-3):127–298.
- Ghazvininejad, Marjan, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A Knowledge-Grounded Neural Conversation Model. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, LA, USA, February.
- Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.
- Han, Sangdo, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting Knowledge Base to Generate Responses for Natural Language Dialog Listening Agents. In *Proceedings of the 16th Annual SIGDial Meeting on Discourse and Dialogue (SIGDIAL 2015)*, Prague, Czech Republic, September.
- Hiatt, Laura M., Anthony M. Harrison, and J. Gregory Trafton. 2011. Accomodating Human Variability in Human-Robot Teams through Theory of Mind. In *Proceedings of the 22nd International Joint Conference of Artificial Intelligence (IJCAI-11)*, Barcelona, Spain, July.
- Hunter, Anthony, Sylwia Polberg, and Matthias Thimm. 2020. Epistemic Graphs for Representing and Reasoning with Positive and Negative Influences of Arguments. *Artificial Intelligence*, 281.
- Kim, Yonghee, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2014. Acquisition and Use of Long-Time Memory for Personalized Dialogue Systems. In *Proceedings of the International Workshop on Multitmodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI 2014)*, Singapore, Singapore, September.
- Kinsella, Bret. 2020. Nearly 90 million U.S. adults have smart speakers, adoption now exceeds one-third of consumers. Online: shorturl.at/idJ69. Accessed on 28.04.2021.
- Kopp, Stefan and Nicole Krämer. 2021. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12:597.
- Krämer, Nicole, Astrid M. Rosenthal-von der Pütten, and Sabrina Eimler. 2012. Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction. In M. Zacarias and de Oliviera J.V., editors, *Human-Computer Interaction: The Agency Perspective*, volume 396 of *Studies in Computational Intelligence*. Springer, Berlin, Heidelberg.

- Lee, Benny P.H. 2001. Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground. *Journal of Pragmatics*, 33:21–44.
- Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2012. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 1:1–5.
- Li, Xiang, Gokhan Tur, Dilek Hakkani-Tür, and Qi Li. 2014. Personal Knowledge Graph Population from User Utterances in Conversational Understanding. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT 2014)*, South Lake Tahoe, NV, USA, December.
- Luger, Ewa and Abigail Sellen. 2016. “Like Having a Really Bad PA”: The Gulf between User Expectation and Experience of Conversational Agents. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2016)*, pages 5286–5297, San Jose, CA, USA, May.
- Luo, Liangchen, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning Personalized End-to-End Goal-Oriented Dialog. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 6794–6801, Honolulu, HI, USA, January.
- Ma, Yukun, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A Survey on Empathetic Dialogue Systems. *Information Fusion*, 64:50–70.
- Miller, Tim. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38.
- Moon, Seungwhan, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Memory Graph Networks for Explainable Memory-grounded Question Answering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL 2019)*, pages 728–736, Hong Kong, China, November.
- Nematzadeh, Aida, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 2392–2400, Brussels, Belgium, November.
- Pei, Jiahuan, Pengjie Ren, and Maarten de Rijke. 2021. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. In *Proceedings of the Web Conference 2021 (WWW 2021)*, Online, April.
- Peters, Christopher Edward. 2005. Foundations of an agent theory of mind model for conversation initiation in virtual environments. In *Proceedings of the AISB-05 Joint Symposium on Virtual Social Agents*, Hatfield, UK, April.
- Pöppel, Jan and Stefan Kopp. 2018. Satisficing Models of Bayesian Theory of Mind for Explaining Behavior of Differently Uncertain Agents. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden, July.
- Porcheron, Martin, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2018)*, Montreal, QC, Canada, April.
- Qiu, Liang, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2021. Towards Socially Intelligent Agents with Mental State Transition and Human Utility. *Computing Research Repository (CoRR)*, arXiv:2103.07011.
- Rabinowitz, Neil, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference of Machine Learning (ICML 2018)*, Stockholm, Sweden, July.
- Ritter, Alan, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgh, Scotland, UK, July.
- Scassellati, Brian. 2002. Theory of Mind for a Humanoid Robot. *Autonomous Robots*, 12:13–24.
- Serban, Iulian Vlad, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. In *Proceedings of 30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, December.
- Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2015)*, Denver, CO, USA, June.
- Speer, Robyn, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*

- (AAAI-17), San Francisco, CA, USA, February.
- Tigunova, Anna, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the Lines: Learning Personal Attributes from Conversations. In *Proceedings of the 28th The Web Conference (WWW 2019)*, San Francisco, CA, USA, May.
- Wang, Qiaosi, Koustuv Saha, Eric Gregori, David A. Joyner, and Ashok K. Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive about a Virtual Teaching Assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2021)*, Yokohama, Japan, May.
- Wu, Sixing, Ying Li, Dawei Zhang, and Zhonghai Wu. 2020. Improving Knowledge-Aware Dialogue Response Generation by Using Human-Written Prototype Dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online, November.
- Yin, Jun, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural Generative Question Answering. In *Proceedings of the Twenty-Fifth Joint Conference on Artificial Intelligence (IJCAI-16)*, New York City, NY, USA, July.
- Zhang, Houyu, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graph. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, July.
- Zhou, Hao, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh Joint Conference on Artificial Intelligence (IJCAI-18)*, Stockholm, Sweden, July.
- Zhu, Wenya, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible End-to-End Dialogue System for Knowledge Grounded Conversation. *Computing Research Repository (CoRR)*, arXiv:1709.04264.

Knowledge Modelling for Establishment of Common Ground in Dialogue Systems

Discussion points raised by
Francesco Cutugno
and Maria Di Maro

The paper deals with the process of grounding in dialogue systems, modelled in terms of factual knowledge of the world, knowledge concerning the user, and the hypothesis of mental knowledge state of the user, i.e., theory of mind. The difficulty of describing and modelling this pragmatic process in conversational agents emerges here in the necessity to refer to and integrate other cognitive theories. Specifically, considering that there are diverse types of shared sets of knowledge, the question that can be addressed refers to their possible different modelling strategy. More in detail, how can they be differently modelled according to their functions? As these sets of knowledge can be partially represented as different aspects of the Common Ground (GC) (Clark 2020), it would be worth exploring how they also interact with one another to successfully communicate. The processes described in the paper, which reflect the state-of-the-art, point out how the success of such grounding applications requires a consistent number of interactions or dialogue turns to efficiently ground information to be used to personalise a dialogue or to infer user's mental states. In this sense, corpus-based training processes, with or without probability-based methods, could be considered as a good starting point. Citing the author's abstract, "[...] this article provides a basic overview of current research on knowledge modelling for the establishment of common ground (henceforth CG) in dialogue systems." The overview is by far more than "basic" and covers a wide range of issues related to CG deepening how to integrate three types of knowledge (i.e., factual, personalised, and beliefs about user knowledge) into any form of automatic system able to manage with task oriented (and not only with them) dialogues. Even if it is not clearly noted in the paper, the introduction of a module able to introduce and represent CG in the general architecture of an automatic dialogue system manager, needs to be strictly "synchronised" with another fundamental module in the architecture: the Dialogue State Tracker (henceforth DST), which, in the recent literature, is more and more becoming the real "pulsing hearth" of these systems. Provided that DST systems have been deeply transformed by the application of Deep Neural Networks, contextual (in a very wide sense) embeddings, inexplicable procedures whose details we all are trying to explain, it could be worth exploring how this is reflected into CG module design. In other words, provided that automatic CG representation processes are called to interact with DST at any time, it is interesting to know what the authors' vision is on the evolution of CG technologies faced to DST systems affected by a high level of complexity. More specifically, under which constraints is it imaginable that also CG technologies can go "into deep"? Natural dialogues, both task oriented and general, have a temporal dynamic. Dialogue state evolves with time and so CG does. We have found very few literature references on evolving systems, able, for example, to find inconsistencies, or re-align dialogue states along with the dialogic situations

that can appear during interaction. CG and knowledge representation can be thought “static” and encyclopaedic but some pointer, indexes, should be active and varying with time, or better, with turns advancements. What is authors’ idea on this matter? In conclusion, it appears almost clear that in the next future, online learning techniques will be introduced more and more pervasively into dialogue systems. Again, temporal evolution awareness and state tracking will take an advantage by this injection. But what about CG? And how Deep Neural Network and online learning will be integrated?

References

Clark, Herbert H. 2020. Common ground. *The International Encyclopedia of Linguistic Anthropology*, pages 1–5.

Knowledge Modelling for Establishment of Common Ground in Dialogue Systems

Response to the discussion points
by Lina Varonina
and Stefan Kopp

Here, we would like to respond to the questions and discussion points raised by Francesco Cutugno and Maria Di Maro in the wake of our paper on knowledge modelling for common ground establishment in dialogue systems. These points are concerned with the connection between the modules for common ground (CG) modelling and dialogue state tracking (DST), and how recent developments with the introduction of deep learning methods to DST can influence CG and knowledge modelling.

Recent research on DST has started to recognise the importance of connecting dialogue context with background information about a domain (Zhou and Small 2019; Ouyang et al. 2020; Chen et al. 2020; Liao et al. 2021). While this is by no means a novel insight, it has not been incorporated much into technical modelling approaches to DST. Further, even in smaller task-oriented domains it is often necessary to look beyond one turn to understand the user and successfully accomplish the task, as one turn will most likely not carry all the information the system needs. In contrast, contemporary voice assistants focus on one-shot request-response interactions with limited needs for context understanding. This is one of the problems users encounter with commercially available voice assistants, which usually expect users "by design" to provide all the necessary information and to ensure that it is understood by the system. When integrating background knowledge about the domain with the information provided by the user during dialogue, however, a system can resolve under-specified requests by making assumptions about user goals (Ouyang et al. 2020): If the user booked a hotel and a restaurant for two people and then wants to also book a taxi, it is highly likely that the taxi will be required for the same two people to transfer from the hotel to the restaurant.

The bigger point behind this argument is that communicative goals of the conversation partners are part of their mental state and the ability to infer mental state facilitates the construction of CG as per the definition of (Clark and Brennan 1991; Lee 2001) that is used in our paper, i.e. mutual knowledge, beliefs and assumptions that the parties have in common either due to similar background or because they were negotiated during the interaction. We argue that to solve the DST challenge one needs to re-recognise the importance of such mental state modelling for human-agent conversational interaction in the future. The current focus of research seems to lie on inference and prediction, while one of the main aspects of CG seen in the above-presented definition is the negotiation of knowledge as the interlocutors cannot be sure that their interpretation of the other's mental state is correct (Kopp and Krämer 2021). Thus, future research should aim to extend the capabilities of dialogue systems to include representations of knowledge that go beyond taking the information recovered from dialogue history as "objective truth". Instead, these representations should incorporate aspects of the interlocutor's mental state, such as epistemic stances or degrees of belief, in order to

account for different degrees of user's understanding or agreement with regards to a particular piece of information.

As we describe in our paper, graphs are an important representation of knowledge in the context of dialogue systems and research on DST often uses this form of representation in combination with deep neural networks (Zhou and Small 2019; Chen et al. 2020; Liao et al. 2021). However, we do believe that there is a need for bringing these methods closer together and that graph-based representations can be embedded into neural dialogue state trackers to enhance the quality of the dialogues through the introduction of cross-turn grounded context and general domain knowledge as argued above. Cutugno and Di Maro correctly note that it is crucial to account for the dynamic nature of knowledge in the context of building CG. Different types of knowledge can exhibit different temporal dynamics of change. Those distinctions can be traced back to Description Logics (Baader, Horrocks, and Sattler 2008) with its concepts of T-Box (terminological box) and A-Box (assertional box). The T-Box contains descriptions of properties and roles of general domain concepts and the relationships between these concepts (comparable with a database schema). The A-Box, on the other hand, contains properties of and relationships between individual instances of these concepts (comparable with data within a database). Looking at modern knowledge representation approaches, one can see parallels of the T-Box and A-Box with the concepts of ontology and knowledge graphs, respectively. In fact, ontology languages such as OWL (Bock et al. 2012), a widely-used web ontology language developed by W3C OWL Working Group, are often based on Description Logics.

Many examples of knowledge modelling discussed in our paper consider knowledge that is static within the use case, e.g., factual domain knowledge. However, we argue that in order to build truly conversational dialogue systems capable of co-constructing CG with their human user, the dynamic aspect of knowledge cannot be discarded. That is, characterising a dialogue state only based on static domain knowledge is insufficient because, even if the topic of a conversation is not changing, user's stance with regards to it may. These notions are currently being introduced into DST research. The work in (Zhou and Small 2019) features a dynamically changing knowledge graph for DST to represent relationships between slots and their values. In their work, they also consider the labels "not mentioned" and "user doesn't care" with regards to possible slot values. This can be considered a basic expression of dynamically changing stance about knowledge within conversation.

The other way around, even the system can have its own stance that evolves throughout the discourse. Depending on the type of the dialogue and the communicative goals of the human and the agent, it may then be necessary to align their beliefs about the domain through interaction, for instance by explanation or argumentation. Hereby it is important to separate representations of the agent's beliefs about the domain and its beliefs about the user's beliefs about the domain. Especially challenging here is that changes in user's beliefs about the domain are never directly observable and can only be inferred under uncertainty from communicative responses or feedback signals. An interesting research question is thus whether a conversational system can reduce this uncertainty with specific dialogue strategies and feedback elicitation in order to more efficiently infer the mental state of the user.

References

- Baader, Franz, Ian Horrocks, and Ulrike Sattler. 2008. Description Logics. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*. Elsevier, pages 135–179.

- Bock, Conrad, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, and Michael Smith. 2012. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). Online. Accessed on 26.10.2021.
- Chen, Lu, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, New York, NY, February 7-12.
- Clark, Herbert H. and Susan E. Brennan. 1991. Grounding in Communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*. American Psychology Association, Washington, D.C., pages 222–233.
- Kopp, Stefan and Nicole Krämer. 2021. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12:597.
- Lee, Benny P.H. 2001. Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground. *Journal of Pragmatics*, 33:21–44.
- Liao, Lizi, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. 2021. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569.
- Ouyang, Yawen, Moxin Chen, Xinyu Dai, Yinggong Zhao, Shujiang Huang, and Jiajun Chen. 2020. Dialogue state tracking with explicit slot connection modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July.
- Zhou, Li and Kevin Small. 2019. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *arXiv preprint*, (arXiv:1911.06192).

Prosody and gestures to modelling multimodal interaction: Constructing an Italian pilot corpus

Luca Lo Re*

Università di Firenze

Modeling dialogue implies detecting natural interaction. A pragmatic approach allows to consider the linguistic act composed of several and different features interacting with each other. Data collected for this project comprises three different genres of communication: monological, dialogical and conversational. The project aims to identify and analyze the pragmatic value of multimodal communication spotting the linguistic actions which carry out illocution values. We draw a pragmatic approach to study multimodal interaction combining the L-AcT annotation (Cresti 2000) with the gesture's architecture designed by Kendon (Kendon 2004). The annotation system is designed to divide the speech units (utterance, intonation units and illocution types) (Hart, Collier, and Cohen 2006) (Cresti 2005) (Moneglia and Raso 2014) from gestural units (Gesture Unit, Gesture Phrase, Gesture Phase). Keeping the Gesture Unit as a superior macro-unit at the other gestural units only for the quantitative purpose, we realize a matching between gesture and speech units. These units work together to form the communicative intention of the speaker that can be recognizable by the Illocution Type. This annotation system leads to understanding how speakers realize multimodal linguistic actions and how different modalities work.

1. Introduction

The main issue in dialogic modelling studies concerns the information management of agents participating in an interaction. Consequently, one of the basic tasks of the main theoretical models on dialogue is to understand the consistency between a dialogic move and its response (Ginzburg and Fernández 2010).

Our goal is to build a model of annotation based on spontaneous spoken language and linguistic units perceptually identified using a pragmatic approach for data segmentation. We intend to illustrate the modelling method of a multimodal corpus of Italian spontaneous speech that can help to detect information management on a computational basis that can also serve as a prototype for the creation of a larger multimodal corpus of spontaneous spoken data. Human communication is defined as multimodal as it occurs through several channels and indices (Fontana 2009). Linguistic action, realized by speakers, is composed by speech, gestures, facial expressions, and context. Each channel is also characterized by several features: speech is characterized by prosody, loudness, intonation, and voice quality, while gestures by rhythm, form, and representation mode. Multimodality is a recent and multidisciplinary field of study. The term was used by Charles Goodwin and Gunther Kress Theo van Leeuwen in the

* Dept. Lettere e Filosofia - Via della Pergola 60, 50121 Florence, Italy. E-mail: luca.lore@unifi.it

mid of 1990s in different fields of study: Goodwin referred to multimodality within the ethnomethodology and Conversational Analysis, while Kress and van Leeuwen within the socio-semiotic studies (Jewitt, Bezemer, and O'Halloran 2016). Despite the growing popularity of the concept of multimodality within different approaches, there still lacks a clear and shared notion of multimodality it is possible to argue that there is still « the need for studying how different kinds of meaning making are combined into an integrated, multimodal whole that scholars attempted to highlight when they started using the term 'multimodality'» (Jewitt, Bezemer, and O'Halloran 2016). Linguistics' interest in multimodality is recent. As a matter of fact, before Kendon and McNeill's work, gesture was regarded as non-verbal communication and only studied in psychology and sociology. Kendon and McNeill have shown that gestures have an important cognitive and linguistic function and that gestures and speech are tightly linked. For McNeill, gesture and speech, are two different sides of the thought: gestures are figurative, holistic, and concise; while speech is arbitrary, analytical, and linear. Consequently, these two aspects of language reside in two different ways of thinking: one figurative and the other propositional. McNeill considers the tension between these two ways of thinking as the urge to think and communicate. In brief, McNeill claims that the gesture is a window on "thought" (McNeill 2011), whereas Kendon sees gesture and speech as two modalities that achieve the utterance. Thus, gesture and speech work together to create the utterance's significance: «an utterance is looked upon as an 'object' constructed for others from components fashioned from spoken language and gesture» (Kendon 2004). Recent studies have shown experimentally the tight link between fluent speech and gesture production. Graziano and Gullberg examined the supposed compensatory role of gestures by detecting their distribution to speech disfluencies in Dutch and Italian speakers. They found that speakers' gestures mainly occur with a fluent gesture both in Italian and Dutch and that gestures are hold back more frequently in disfluent speech. The first finding shows a very strong connection between fluent speech and gesture production, against the Lexical Retrieval Hypothesis' (Krauss and Hadar 1999) predication according to which gestures occur more frequently during speech disfluencies. Moreover, the second finding reinforces the notion that speech and gesture form an integrated system showing that «when speech stops, so does gesture» (Graziano and Gullberg 2018). Cavicchio and Kita studied gestural communication in early bilinguals, detecting the gestural transfer through gesture' parameters (gesture rate and gesture salience), when speakers switch languages. They found that when bilinguals switch language, their gesture parameters switch accordingly with the language they talk. This result also supports the idea that human language is multimodal (Cavicchio and Kita 2013). Increasing interest in multimodal communication, especially in gesture studies, has requested more and more data to detect these fields and resulted in a considerable growth of multimodal corpora. This raises two issues that are addressed in our study. First, the increase of multimodal corpora leads to an increase in the annotation systems available: almost one per corpus. Second, the data is generally elicited, collected in the laboratory through the use of tasks, interviews, retelling, or TV videos, generating an underrepresentation of spontaneous spoken data. With this project, we propose an annotation system that is easy to use and clear, since, at the best of our knowledge, there still «lack an adequate conceptual apparatus, transcription system and terminology for dealing with the phenomena of gesture» (Kendon 2004). Furthermore, we use spontaneous spoken data which allows to capture more closely the natural occurring speech-gesture interaction and fill a gap in the language data used in this research field. The following sections describe the theoretical approach, the annotation system, and

the data collection process In this work, we define gestures each movement of the hand and head related to the interaction.

2. Theoretical approach

To spot out the model and the method to create a multimodal corpus we start from the notion of Linguistic Action. This idea, based on the Austinian theoretical framework, is developed within the Language into Act Theory designed by Emanuela Cresti (Cresti 2000). Taking into consideration the pragmatic value of gestures argued by scholars (Kendon 2004); (Müller, Ladewig, and Bressemer 2013); (Loehr 2014); (Cienki 2017),), we found necessary to extend this notion to gesture analysis. However, this approach raises the issue of speech flow segmentation. In fact, despite scholars recognize the architecture structure of gesture spotted by Kendon and reviewed by Kita (Kita, Van Gijn, and Van der Hulst 1997), it lacks a clear and shared coding of gestures' type or gestures' functions. To date, each decoding system is based on purpose study and the different theoretical approaches adopted. The following section illustrates our approach based on the Language into Act Theory and on the pragmatic value of gestures. The aim is to create a pilot corpus of spontaneous data that allows to detect speech as a multimodal unit under the assumption that the speech act is composed by different features interacting with each other. We believe that a multimodal corpus based on a pragmatic approach and on Linguistics Action notion, could allow future research to provide an empirical criterion to detect and define the notion of multimodal unit.

2.1 Language into Act Theory

L-Act is based on the Speech Act theory of Austin and elaborated on empirical observations of spontaneous speech corpora. This theory views speech as aroused by the speaker's affect toward the addressee and that is realized into a speech act with pragmatic value. In this model, the pragmatic function is considered the main function of speech that manages the linguistic feature and the syntactic structure. Prosody plays an important role within the illocutionary and locutionary act relationship, indeed it expresses the pragmatic function of the speech act making it a real audible entity. The information structure is built around the necessary and sufficient unit called Comment and that could be accompanied by other optional units with which it forms the information pattern. The additional units take on different functions: Topic, Parenthesis, Appendix, Locutive Introducer, and Discourse Markers. L-AcT has made a proposal, modelled through corpus-driven research, inside the debate on the speech flow segmentation and speech reference units. The proposal is based on two types of reference units prosodically identified: utterance and stanza. The utterance is the minimal and primary linguistic unit characterized by a terminated prosodic boundary and that accomplishes a single speech act; on the other side, a stanza is formed by a sequence of weak Comments that do not correspond to a sequence of utterances. Stanza is not strictly governed by pragmatics principles but rather follows strategies of textual construction (Moneglia and Raso 2014)(Panunzi and Scarano 2009). Thus, speech reference units are linguistic entities based on semantic, pragmatic, and prosodic features. Identification of reference units occurs prosodically through perceptual recognition of terminated or non-terminated boundaries by the annotator. L-AcT illocutionary classification is based on the speaker's affective activation toward the addressee and on corpus analysis that leads researchers to identify five mains illocutionary classes: refusal, assertion, direction, expression, and ritual. It's important to point out that, unlike other proposals for

which the illocution's accomplishment is ensured only by the change and transformation of the world, «from the L-AcT perspective, the illocutionary activation (originating from the affect) is accomplished regardless of its subsequent recognition and takes place in the world even in the absence of acceptance or understanding by some party» (Cresti 2020).

As mentioned above, the illocutionary value is expressed only by the Comment unit. Moreover, L-AcT is supported by prosodic model referring to works of IPO (Hart, Collier, and Cohen 2006). Between the Information Pattern and Prosodic Pattern, there is a correspondence (Moneglia and Raso 2014).

This framework considers speech as a pragmatic activity performed by the speaker: «it [L-AcT] stressed that prosody plays a mandatory role in the performance of the utterance and its linguistic identification. Moreover, L-AcT foresees that the internal information organization of the utterance is governed by pragmatic principles and is crucially mediated by prosody» (Moneglia and Raso 2014). Relating this with other evidence of the gesture prosody' relationship, we want to extend L-AcT to gesture analysis. We think that starting from a well-defined framework and an utterance's definition based on perception can be useful to study multimodal utterance from a pragmatic view. We believe that a pragmatic approach of this kind, which sees language as an action that arises from an affective impulse and is concretely realized in speech, represents a good method to detect how different features work to create a language action. Thus, in this framework, linguistic analysis cannot be separated from the analysis of the units physically produced through speech and perceptually recognizable by speakers. Considering that language is a multimodal linguistic act, it seems that is necessary to extend this approach to the gestural aspect as well.

2.2 Gesture and pragmatics

In the past, the gesture was a matter of pragmatics because it was not considered like a linguistics feature, this traditional view arose from the influence of generative linguistics (Cienki 2017). In recent years, several studies showed that gestures are features of verbal communication and underlined that gestures play a crucial role either in the cognitive part (McNeill 2008) and in the pragmatics (Kendon 2004) of the speech.

Kendon argued that some Italian gesture has pragmatic functions. He described gestures that mark the illocutionary force of an utterance (*illocutionary marker gestures*), and gestures that have the function to indicate the status of the unit inside a discourse (*discourse unit marker gestures*). Kendon concludes that «speakers may use gestures which can explicitly mark a given stretch of speech as being a particular type of speech act. Within a discourse, they can differentiate gesturally topic from comment, or indicate what units are 'focal' for their arguments», he named these gestures 'pragmatic' (Kendon 1995). Bressemer and Müller spotted a list of recurrent gestures in German that carry out pragmatics function and illocutionary values (Bressemer and Müller 2014). Enfield and colleagues (Enfield, Kita, and De Ruiter 2007) - studying Laos people - have distinguished two types of pointing gestures based on the role played by the gesture in constructing the information of the utterance: B-point (big in form) and S-point (small in form). The first one pointing gesture' type plays a necessary role within the multimodal utterance while added speech is merely supportive of B-point. Whereas, the S-point gestures are more dependent and more hidden in the information structure of the utterance. «While a B-point is doing the primary work of the utterance, with speech playing a supporting role, an S-point adds a backgrounded modifier to an utterance in which speech is central» (Enfield, Kita, and De Ruiter 2007). An S-point represents a

low risk communicative action, which might save the speaker against a potentially high social and interpersonal cost (Enfield 2006).

All these works, despite different approaches and theoretical views, can contribute to extending the idea of linguistic action like a multimodal action. Because gestures play an important pragmatic role (expressing several functions) it is important to recognize that the gestural part is not a correlated feature of the utterance, but gestures carry out – with speech - the linguistic action.

Indeed, Kendon defines gesture as «a name for visible action when it is used as an utterance or as a part of an utterance» (Kendon 2004) and sees the utterance as «any unit of activity that is treated by those co-present as a communicative ‘move’, ‘turn’ or contribution». Such units of activity may be constructed from speech or from visible bodily action or from combinations of these two modalities» (Kendon 2004). Bressemer and Müller based their study on the multimodal utterance in kendonian sense. Whereas Enfield speaks about a composite utterance defining it «as a communicative move that incorporates multiple signs of multiple types» (Enfield 2009). The composite utterance has a coded meaning – which consists of lexical and grammatical values (e.g. conventional linguistic sign) – and an enriched meaning that can be indexical if it explains the unclear utterance’s references – this can be realized either explicitly (by an indexical symbol like “this”) than implicitly (by the copresence in the time and the space like no-smoking notice) – or implicational according to the gricean model – so the meaning is achieved either through a codex system and by an interpretation based on a common ground (Enfield 2009). The idea of a multimodal utterance seems to be a theoretical concept, based on empirical evidence, but that cannot become a useful unit to linguistic analysis. There is not a definition based on practical features as well as the spoken utterance. If on the one hand, Kendon did not define multimodal utterance practically, on the other hand, Enfield referred to the composite utterance of the social interaction’s basic unit, that he called *move* according to Groffman’s theory which says: «a move may be defined as a recognizable unit contribution of communicative behavior constituting a single, complete pushing forward of an interactional sequence by means of making in some relevant social action recognizable (e.g., requesting the salt, passing it, saying thanks)» (Enfield 2009). Considering these theoretical frameworks, we aim to draw a pragmatic approach to study the multimodal spontaneous interaction. We start from the concept of language as action and then try to detect how and which basic units can compose the linguistic action. Specifically, how the different basic units (prosodic and gestural) interact and relate to each other in making the action. To do this we base our method on the efficient theoretical model of Language into Act (Cresti 2000).

3. The annotation system

Several studies drew annotation systems for the gesture. Each one is characterized by its method, research purpose, and tag definition. Some examples can be represented by NEUROGES, CoGesT, and LASG. NEUROGES is a coding system based on the assumption that gestures are closely linked to cognitive, emotional, and interactive processes. This system is well organized and divided into three modules (Kinesics, relation between the hands, and cognition/emotion) and several steps. This coding system is fine-grained and thus presents dozens of labels (Lausberg 2013) (Lausberg and Sloetjes 2016).

CoGesT (Conversational Gesture Transcription), was created to provide a transcription system for linguistic analysis and automatic processing of gestures. This system distinguishes gestures into Simplex gestures and Compound gestures. In the first one

there are two types, place static – a gesture that holds a specific hand configuration – and place dynamic where gestures are characterized by Source, Trajectory, and Target; these attributes are represented as a vector (Gibbon et al. 2003) (Trippel et al. 2004).

LASG (Linguistic Annotation System for Gesture), offers an annotation of gestures grounded in a cognitive linguistic approach and refers to a form-based approach for gesture analysis. It provides several levels of annotation: annotation for the gesture that includes sub-level as determining units, annotation of forms, motivation of form; annotation of speech that includes as sub-level to turn and intonation unit; and annotation of gesture about speech with other sub-levels as prosody, syntax, semantics, and pragmatics (Bressemer, Ladewig, and Müller 2013). All these examples are excellent annotation systems, but they present problems for the creation of a spontaneous speech corpus and for an annotation system that can be usable with spontaneous and large data. We aim to offer a simplified and efficient annotation system that can point out how gesture and speech create a multimodal utterance. We think that it is necessary to segment the gestural and speech flow on basic units that are perceptually detectable: the intonation units for speech, and the movement units for gestures. Furthermore, we consider intonation the crucial element of the utterance that is perceptually well-defined and linguistically meaningful. Loehr showed that gesture and prosody are tightly connected, both channels – gesture and speech – work together to construct discourse and to regulate interaction. This relationship was found either in production and perception, in all ages, and in dozens of languages (Loehr 2007) (Loehr 2014). We want to unify the L-AcT annotation – that emphasize the intonation's role in the speech – with the gesture's architecture designed by Kendon that offers an important gesture's structure composed of the single unit and phase of gestural movement (Kendon 1972).

The idea is to create a transcription and annotation system that can identify the basic units on a perceptual basis. On the one hand, as we have seen above, we have a model like Language into Action Theory that gives us the means and evidence to identify utterances and intonation units of the spoken modality and therefore on auditory perception. On the other hand, for gestural transcription and annotation, we lack a widely shared model. Kendon's and McNeill's studies provide an architecture of the gesture that manages to identify the units that make up the gesture without being able to univocally correlate linguistic values to the different units. It seems clear that there is a necessary and sufficient unity, represented by the stroke phase. And undoubtedly some studies show us some evidence on how in certain context gestures manage to express pragmatic and semantic values through means and solutions that seem conventionalized. For this reason, we found it is necessary to keep speech and gesture annotation separate. The two parallel annotations are based on a perceptual method that is auditory for speech and visual for the gesture. The multimodality of linguistic action emerges from the annotation of illocution, which represents the linguistic element that characterizes in our opinion the use of semantic, intonation, and gestural elements.

3.1 Speech transcription and annotation

Spoken language is characterized by several specific phenomena, some of which are related to the interaction – e.g. overlapping, vocalization, and retracting – other phenomena are related to linguistic features like intonation. Spoken language transcription cannot leave out these specific features that allow making a spoken text interpretable.

LABLITA corpora offer a good transcription method based on L-AcT and CHAT format (Cresti 2000). As previously mentioned, L-AcT is an extension of Austin's Speech Act theory and sees the speech as a result of the speaker's pragmatic activities. Prosody

**Figure 1**

Stretch of the speech transcription

plays a pivotal role in the performance of the utterance. Moreover, the utterance's information organization is based on pragmatic principles and is mediated by prosody. L-Act theory provides a tool of description and annotation for spontaneous speech. The format CHAT LABLITA was created in accordance with this framework, that implements format CHAT, created within the project CHILDES, including intonation and its function of demarking of utterance and information units.

As discussed above, the speech flow is segmented perceptually into tone units marked by prosodic breaks that can be terminated or non-terminated. The first one marks the utterance boundaries and is represented using two slashes //; the second one marks other prosodic units inside the utterance and is represented using only one slash /. For the transcription of other phenomena – like non-linguistic sound, fragments, words interrupted, retracting, and overlapping – the format provides a complete repertoire as is illustrated in the following table.

Table 1

Transcription symbols of CHAT-LABLITA format

<i>Symbol</i>	<i>Value</i>
//	Terminated prosodic break
?	Terminated prosodic break (interrogative intonation)
...	Terminated prosodic break (suspensive intonation)
+	Terminated prosodic break (interrupted sequence)
/	Non-terminated prosodic break
/	False start with repeat
//	False start with partial repeat
<	Overlapping start
>	Overlapping end
<	Signal to repeat relation
&	Vocalization
hhh	Paralinguistics or non-linguistics vocal phenomenon
xxx	Unintelligible word

The figure 1 shows a stretch of the speech transcription and annotation.

3.2 Gesture annotation

To transcribe gestures, we use the gesture's architecture drawn by Kendon. It is hierarchical and composed by a macro-unit called Gesture-Unit, that is «entire excursion, from the moment the articulators begin to depart from the position of relaxation until the moment when they finally return to one» (Kendon 2004). This excursion is divided into Gesture-Phrase, that is «what we call a 'gesture'» (McNeill 2008). Also, Gesture-Phase is composed by three other units called Gesture-Phases, that are preparation (the limbs



Figure 2
Stretch of the gesture transcription

that move from a rest position), stroke («it is the phase of the excursion in which the movement dynamics of effort and shape are manifested with greatest clarity») (Kendon 2004), recovery or retraction (the phase that follows the stroke, when the hand returns to a relaxed position), sometimes can be a hold phase «a phase in which the articulator is sustained in the position at which it arrived at the end of the stroke» (Kendon 2004) (Kita, Van Gijn, and Van der Hulst 1997). The figure 2 shows a stretch of the gesture transcription.

3.3 Multimodal relation between annotations

The model aims to identify and analyze how the basic units - units perceptively interpretable - interrelate each other to form the pragmatic value of the multimodal utterance performed in a spontaneous interaction. This can be achieved starting from the utterance’s idea defined by Cresti. In order to make L-AcT a multimodal model, it is necessary to correlate the gesture transcript with the speech transcript. Throughout this approach, it will be possible to spot the linguistic actions with illocution values, realized by the interaction of gestural and spoken features like Prosody units, Prosodic breaks, Illocution types, Gesture phrases, and Gesture phases.

The annotation system is designed to divide the speech units from the gestural units. The speech annotation structure has two units: a) utterance, b) prosodic units (Cresti 2000) (Moneglia and Raso 2014). The gestural annotation, instead, has these units: a) Gesture Unit, b) Gesture Phrase, c) Gesture Phase (Kendon 2004). Applying this method makes it possible to analyze two different modalities together and detect how speech acts are realized. Keeping the Gesture Unit as a superior macro-unit at the other gestural units only for the quantitative purpose, allows to match gesture and speech basic units that work together to form the communicative intention of speaker that can be recognizable by the Illocution Type. To annotate the Illocution class we use the five general class spotted out by Cresti : *Refusal, assertion, direction, expression, and ritual* (Cresti 2005) (Cresti 2020).

Table 2
Relation between units

<i>SPEECH UNITS</i>		<i>GESTURAL UNITS</i>
Utterance	↔	Gesture Phrase
Prosodic Unit	↔	Gesture Phase
	Illocution Type	

The Utterance is associated with Gesture Phrase because both are the higher units and because it is possible to identify perceptually: the Utterance by the terminate prosodic break, and Gesture Phrase by the direction’s change, the movement’s rhythm,

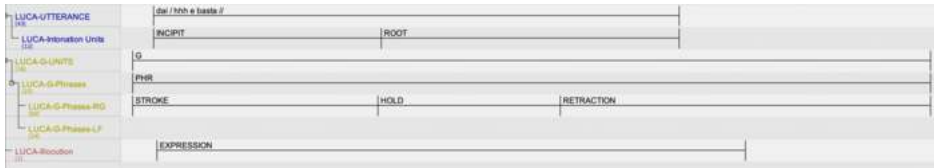


Figure 3
Stretch of transcription

or for the movement's end. The Prosodic Unit is associated with the Gesture Phase because they are the units that categorize the meaning's features: the root on one side – the intonation units of speech that is necessary and sufficient to the utterance – and the stroke on the other side – the meaningful gestural unit. We believe that applying this annotation system can lead to understanding how speakers realize multimodal linguistic actions. In particular, it makes possible to detect how the multimodal action is composed by the different features of different modalities. Data are annotated using software ELAN, which means the annotation is organized into tiers. The template is organized into two parts for each speaker: the spoken part that has the main tier called "utterance" – where speech is transcribed – and a depended tier called "Prosodic unit"; the gestural part with the main tier "G-Units" with a depended tier "G-Phrases" that is parent tier of "G-Phases". The tier with the illocution value is independent (figure 3).

This annotation system allows to detect how the different basic units interact with each other during a spontaneous interaction, performing a linguistic action. In this way, it will be possible to investigate how the phenomena of overlapping, interruption, and retracting interact in the relationship between speech units and gestural units. Most importantly, this annotation system allows us to investigate how the different linguistic actions of speakers collaborate for the construction of speech. In fact, data collected in a spontaneous context give the possibility to bring out phenomena that would not emerge with elicited data. From a computational point of view, this corpus will empower to monitor typical phenomena of spontaneous interaction to create a dialogue model. In communicative exchanges in spontaneous contexts, you can observe natural events and phenomena. In the example below (figure 4), it is possible to see how a gesture, which according to McNeill should be defined as a speech-linked gesture (a gesture that occupies a grammatical slot in a sentence) (McNeill 2008), fully realizes the semantic and illocutionary value of the utterance. Despite this, the gesture does not appear to be coded on a typological or semantic basis, according to the coding proposed by scholars. The figure 4 shows the boy responding to the girl who had asked why he had not studied. The boy responds by saying "because..." and making the gesture.

4. Collecting Data and tools

Constructing a corpus implies following several principles in the collection and organization of the data that are related to the corpus type and the research objectives. The main concept that guides these principles is quantitative and qualitative representativeness. The goal of our project is to draw a method and an approach to the creation of a multimodal corpus of spontaneous spoken Italian. For this reason, our pilot corpus is composed by different communicative situations and different data collection points. Italy has a great diatopic variation based on a large dialectal variability. We collect our



Figure 4
Example of gesture

data in two cities, Firenze and Catania, aiming capture spoken Italian that is influenced by dialectal substratum. In fact, how Interfaces Hypothesis shown (Kita and Özyürek 2003), the gestural form depends also on the information' organization of a specific linguistics system.

Following studies from Özyürek, who shows that the number of participants influences gestures and the shared space during a conversation (Özyürek 2002), we collected data from three different genres of communication: monological – which includes only one speaker (e.g., a lecture) with listeners; dialogical – that includes only two participants interacting; conversation – more than two participants. Interactions occurred in a natural context (a lecture at the university and conversations at the private homes of the participants) and were all spontaneous. With this design, we are collecting six different communicative situations: three genres for two different places. Participants were 20 – 60 years old with a secondary high school degree as the minimum education. At the start of the recording, participants are informed that they are recorded for research in linguistics. The goal of the recording is disclosed at the end of the session by handing in a piece of detailed information about the purpose of recording and its dissemination. For the recording, we use one or two cameras – GoPro Hero 6 – and one or two audio recorders – Zoom H6 – with a panoramic microphone (120°). We record participants during a real communicative event, like planning a meeting, a lecture, or a meeting with friends so the set change for each recording. In the following table, it is possible to see a resume of the interactions recorded.

5. Conclusions

The modeling of a multimodal corpus proposed shows the complexity of natural occurring interaction: speakers use several tools, like intonation and gesticulation, to communicate. Information is conveyed through different channels with different modes, hence the multimodal nature of interaction. To create a dialogic model that can be effective

Table 3
Corpus dataset

<i>INTERACTION GENRE</i>	<i>INTERACTION TYPE</i>	<i>PLACE</i>
Conversation	Three handball referees meeting	Firenze
Conversation	Three friends meeting to plan a trip	Catania
Dialogue	Scoutmasters meeting	Firenze
Dialogue	Students meeting	Catania
Monologue	Italian lecture	Firenze
Monologue	Storytelling	Catania

and close to the reality of the speakers, we believe it may be useful to base extracting a model based on a pragmatically annotated multimodal corpus. The pragmatic approach allows us to consider the linguistic act composed by several and different basic units that interact with each other: sound, prosody, gesture, metaphor, grammar, and rhythm. Our method is based on L-Act annotation scheme (Cresti 2000), adding the gestural annotation. The main contribution of our study to this field of research is the use of spontaneous data, which brings to light phenomena that cannot be elicited in a laboratory environment. To conclude, multimodal corpora represent a valuable opportunity to investigate the management of action linguistics between speakers through the two main modalities used in spontaneous interaction. This type of transcription undoubtedly allows the possibility of computational analysis of the relationships between language acts, gestures, and intonation.

References

- Bressem, Jana, Silva H. Ladewig, and Cornelia Müller. 2013. Linguistic annotation system for gestures. In *Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK) 38/1*. De Gruyter Mouton, chapter 71, pages 1098–1124.
- Bressem, Jana and Cornelia Müller. 2014. A repertoire of german recurrent gestures with pragmatic functions. In *Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK) 38/2*. De Gruyter Mouton, chapter 119, pages 1575–1591.
- Cavicchio, Federica and Sotaro Kita. 2013. Bilinguals switch gesture production parameters when they switch languages. In *Proceedings of the Tilburg Gesture Research Meeting (TIGeR) 2013*, Tilburg, The Netherlands, June.
- Cienki, Alan. 2017. From paralinguistic to variably linguistic. *The Routledge handbook of pragmatics*, 61:68.
- Cresti, Emanuela. 2000. *Corpus di italiano parlato*, volume 1. Accademia della Crusca.
- Cresti, Emanuela. 2005. Per una nuova classificazione dell'ilocuzione a partire da un corpus di parlato (LABLITA). In Elisabeth Burr, editor, *Tradizione e innovazione. Il parlato: teoria - corpora - linguistica dei corpora. Atti del VI Convegno della Società di Linguistica e Filologia Italiana (SILFI)*. Franco Cesati Editore.
- Cresti, Emanuela. 2020. The pragmatic analysis of speech and its illocutionary classification according to the language into act theory. In *In Search of Basic Units of Spoken Language: A corpus-driven approach*, volume 94. John Benjamins Publishing Company, pages 181–219.
- Enfield, Nicholas J. 2006. Social consequences of common ground. In S.C. Levinson, editor, *Roots of human sociality. Culture, cognition and human interaction*, Wenner-Gren International Symposium Series. Oxford: Berg, pages 399–430.
- Enfield, Nicholas J. 2009. *The anatomy of meaning: Speech, gesture, and composite utterances*. Language Culture and Cognition. Cambridge University Press.

- Enfield, Nick J., Sotaro Kita, and Jan Peter De Ruiter. 2007. Primary and secondary pragmatic functions of pointing gestures. *Journal of Pragmatics*, 39(10):1722–1741.
- Fontana, Sabina. 2009. *Linguaggio e multimodalità: gestualità e oralità nelle lingue vocali e nelle lingue dei segni*. ETS.
- Gibbon, Dafydd, Ulrike Gut, Benjamin Hell, Karin Looks, Alexandra Thies, and Thorsten Trippel. 2003. A computational model of arm gestures in conversation. In *Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, September.
- Ginzburg, Jonathan and Raquel Fernández. 2010. 16 computational models of dialogue. *The handbook of computational linguistics and natural language processing*, 57:1.
- Graziano, Maria and Marianne Gullberg. 2018. When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in psychology*, 9:879.
- Hart, Johan't, René Collier, and Antonie Cohen. 2006. *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press.
- Jewitt, Carey, Jeff Bezemer, and Kay O'Halloran. 2016. *Introducing multimodality*. Routledge.
- Kendon, Adam. 1972. Some relationships between body motion and speech. *Studies in dyadic communication*, 7(177):90.
- Kendon, Adam. 1995. Gestures as illocutionary and discourse structure markers in southern Italian conversation. *Journal of pragmatics*, 23(3):247–279.
- Kendon, Adam. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.
- Kita, Sotaro and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language*, 48(1):16–32.
- Kita, Sotaro, Ingeborg Van Gijn, and Harry Van der Hulst. 1997. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Proceedings of the International Gesture Workshop*, pages 23–35, Bielefeld, Germany, September. Springer.
- Krauss, Robert M. and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. *Gesture, speech, and sign*, 93.
- Lausberg, Hedda. 2013. NEUROGES – A coding system for the empirical analysis of hand movement behaviour as a reflection of cognitive, emotional, and interactive processes. In *Body - Language - Communication*, volume 1. De Gruyter Mouton, chapter 67, pages 1022–1037.
- Lausberg, Hedda and Han Sloetjes. 2016. The revised neuroges–elan system: An objective and reliable interdisciplinary analysis tool for nonverbal behavior and gesture. *Behavior research methods*, 48(3):973–993.
- Loehr, Dan. 2014. Gesture and prosody. In *Body - Language - Communication*, volume 2. De Gruyter Mouton, chapter 100, pages 1381–1391.
- Loehr, Daniel. 2007. Aspects of rhythm in gesture and speech. *Gesture*, 7(2):179–214.
- McNeill, David. 2008. *Gesture and thought*. University of Chicago press.
- McNeill, David. 2011. *Hand and mind*. De Gruyter Mouton.
- Moneglia, Massimo and Tommaso Raso. 2014. Notes on language into act theory (I-act). *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, pages 468–495.
- Müller, Cornelia, Silva H. Ladewig, and Jana Bressemer. 2013. Gestures and speech from a linguistic perspective: A new field and its history. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Tessedorf, editors, *Body - Language - Communication*, volume 1. De Gruyter Mouton, chapter 3, pages 55–81.
- Özyürek, Asli. 2002. Do speakers design their cospeech gestures for their addressees? the effects of addressee location on representational gestures. *Journal of Memory and Language*, 46(4):688–704.
- Panunzi, Alessandro and Antonietta Scarano. 2009. Parlato spontaneo e testo: analisi del racconto di vita. In L. Amenta and G. Paternostro, editors, *I parlanti e le loro storie: Competenze linguistiche, strategie comunicative, livelli di analisi*. pages 121–132.
- Trippel, Thorsten, Dafydd Gibbon, Alexandra Thies, Jan-Torsten Milde, Karin Looks, Benjamin Hell, and Ulrike Gut. 2004. CoGesT: A formal transcription system for conversational gesture. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May.

How are gestures used by politicians? A multimodal co-gesture analysis*

Daniela Trotta**
Università degli Studi di Salerno

Raffaele Guarasci†
ICAR-CNR

Gestures are an inseparable part of the language system (McNeill 2005; Kendon 2004), they are semantically co-expressive with speech serving different semantic functions to accompany oral modality (Lin 2017; McNeill 2016). To study these phenomena, we analyse the co-gesture behavior of several Italian politicians during face-to-face interviews. We add a new annotation layer to the PoliModal corpus (Trotta et al. 2020) focused on semantic function of hand movements (Lin 2017; Colletta et al. 2015; Kendon 2004). Then, we explore the patterns of co-occurrence of speech and gestures for the single politicians and from a party perspective. In particular, we address following research questions: i) Are there categories of verbs that systematically accompany hand movements in political interviews? ii) Since the corpus used presents an annotation of "speech constants" (Voghera 2001), is the Lexical Retrieval hypothesis confirmed or are gestures used in correlation with other and different constants of speech? The Lexical Retrieval hypothesis assumes that (a) gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewellyn 1969; Butterworth and Beattie 1978), and (b) that the inability to gesture can cause verbal disfluencies. Finally, we analyse semantic patterns of gesture-speech relationship.

1. Introduction

Messages can be encoded through verbal or non-verbal signals (Wagner, Malisz, and Kopp 2014). Although communication research has traditionally focused on speech – demonstrated by the fact that in recent decades a huge quantity of work, tools and approaches have been developed in the field of Spoken Corpus Linguistics (Voghera 2020; O’Keeffe and McCarthy 2010) – interest has shifted mainly towards multimodality in recent years. This is evidenced by the numerous occasions of discussions in the scientific community on this topic, focused on: technical modeling of manual gestures in human-machine interaction (i.e. the GESPIN conferences 2009 and 2011; Gesture Workshop Series), technical aspects of multimodal facial communication (i.e. The Audio-Visual Speech Processing Workshops - AVSP) and on research approaches to gesture analysis (i.e. LREC Workshops on Multimodal Corpora; the International Society for Gesture Studies). At the same time, there has been a strong increase of multimodal corpora

* Although the authors have cooperated in the research work and in writing the paper, they have individually devoted specific attention to the following sections: Daniela Trotta: 1, 2, 3 and 8; Raffaele Guarasci 4, 5, 6 and 7

** Dept. of Political Science and Communication – Via Giovanni Paolo II 132, 84084 Fisciano, Italy.
E-mail: dtrotta@unisa.it

† Institute for high performance computing and networking – Via Pietro Castellino 111, 80131 Napoli, Italy.
E-mail: raffaele.guarasci@icar.cnr.it

that stimulates sophisticated investigations into the relationship between the verbal and nonverbal components of spoken communication (Knight 2011).

This growing interest is strictly related to the fact that it is not possible to get a complete picture of human communication excluding some of the information provided during speech. As best pointed out by (Allwood 2008): “The basic reason for collecting multimodal corpora is that they provide material for more complete studies of interactive face-to-face sharing and construction of meaning and understanding which is what language and communication are all about”. In fact, every spontaneous spoken communication is accompanied by gestures (i.e. facial expressions, hand movements, postures and body movements) (Voghera 2020). Indeed – as we will explain better in the section 2 – gestures accompanying speech take on multiple functions, ranging from complete the utterance, to substitute part of the utterance and to contradict the verbal sequence (Kendon 2004; McNeill 2008; Poggi 2007).

However, developing multimodal resources is extremely time-consuming (Lin 2017), because of the difficulty of transcribing and keeping track of all the non-verbal elements. Therefore, multimodal resources currently developed for all the languages are few and of different domains. The vast majority of these resources are monolingual relying on English language only.

Concerning Italian, the recent research on multimodal corpora is limited to the experience of the IMAGACT project (Moneglia et al. 2014) which aims at setting up a cross-linguistic Ontology of Action for grounding disambiguation tasks and it makes use of the universal language of images to identify action types, avoiding the underdeterminacy of semantic definitions. There are currently no resources for the Italian language that simultaneously account for verbal and non-verbal dimensions, this lack has affected the development of lines of research focused particularly on the relationships between the co-occurrence of speech and gesture.

Given that the television interview is inherently a multimodal and multisemiotic text, in which meaning is created through the intersection of visual elements, verbal language, gestures, and other semiotic cues (Vignozzi 2019), this study focuses on the co-gesture behavior of several Italian politicians during TV face-to-face interviews.

Starting from PoliModal corpus (Trotta et al. 2019, 2020), an Italian multimodal corpus of political domain, we add a new annotation layer focused on semantic function (i.e. reinforcing, integrating, supplementary, complementary, contradictory) of hand movements (Lin 2017; Colletta et al. 2015; Kendon 2004) in order to explore the patterns of co-occurrence of speech and gestures for the single politicians and from a party perspective.

1.1 Research Objectives

This work investigates political non-verbal communication. To date, in the literature Multimodal corpora have been used to analyse how gestures are used in different contexts such as narratives (Gregersen, Olivares-Cuhat, and Storm 2009; Holler and Wilkin 2011; Parrill, Bullen, and Hoburg 2010), academic domain (Knight 2011; Ovendale 2012), child language development (Colletta et al. 2015) and in relation to Italian action verbs (Moneglia et al. 2014). This study aims to explore the patterns of co-occurrence of speech and gestures in the specific case of Italian political interviews from a multimodal corpus linguistics perspective, addressing the following research questions:

1. Are there categories of verbs that systematically accompany hand movements in political interviews? This research question is inspired by

the study presented in (Vignozzi 2019) in which the analysis of the representation of some peculiar indicators of speech (i.e. idiomatic expressions and phrasal verbs) in a corpus of English television interviews of different domain, revealed that phrasal verbs are more recurrent in political interviews, while hand movements are more often associated with business and economic interviews.

2. Since the corpus used as a case study presents an annotation of so-called “speech constant” (Voghera 2001) (i.e. pauses, interjections, false starts, repetitions, truncations), is the *Lexical Retrieval hypothesis* confirmed or are gestures used in correlation with other and different constants of speech? Note that the *Lexical Retrieval hypothesis* assumes that (a) gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewellyn 1969; Butterworth and Beattie 1978), and (b) that the inability to gesture can cause verbal disfluencies (Dobrogaev 1929).
3. In the case of political interviews, what are the semantic patterns of gesture-speech relationship?

Our examination of the co-occurrence of speech and gesture will shed light into how the two communication models interact.

2. Background

2.1 Co-Gesture Analysis: a new perspective of linguistic analysis

A gesture is a visible action of any body part, when it is used as an utterance, or as part of an utterance (Kendon 2004). If such actions are produced while speaking, we can talk about co-speech gestures. Their occurrence, simultaneous or concomitant to speech, has led to different views regarding their role in communication (Wagner, Malisz, and Kopp 2014). First of all – as pointed out by (Voghera 2020) – when we think about the relationship between verbal sequence and gestures, we should not imagine that the latter have a merely subordinate function to the word, but rather that there is a relationship of semiotic cooperation between them. The presence of gestures is useful to both the addressee and the speaker to maintain the rhythm of the speech rhythm of the speech and to mark the progression of information.

Some authors (McNeill 2005; Kendon 2004) have considered gestures as an integrative, inseparable part of the language system, since speaking itself is regarded as a variably multimodal phenomenon (Cienki and Müller 2008). Indeed gestures may provide important information or significance to the accompanying speech and add clarity to discourse (Colletta et al. 2015); they can be employed to facilitate lexical retrieval and retain a turn in conversations (Stam and McCafferty 2008) and assist in verbalizing semantic content (Hostetter, Alibali, and Kita 2007). From this point of view gestures facilitate speakers in coming up with the words they intend to say by sustaining the activation of a target word’s semantic features long enough for the process of word production to take place (Morsella and Krauss 2004). Co-gesture speech can also refer to the spoken words or phrases that are co-produced with hand gestures in face-to-face spoken conversation (Lin 2017). According to (Krauss 1998) these co-occurring words or entire lexical phrases were identified to reflect the meaning of the co-occurring gesture; they are also known as “lexical affiliates” of the gesture, especially if they play a

particular role in the lexical retrieval. Indeed if gestures play a role in a lexical retrieval, they must stand in a particular temporal relationship to the speech they are supposed to facilitate.

Over the years, studies have shown that the production of gestures is influenced by the syntax of the language itself and by the socio-cultural context of the language. As explained in a 2015 study by (Colletta et al. 2015) – focused on co-speech gesture production in children’s narratives – language syntax influences gesture production. For example – as known – some languages require an explicit subject (i.e. English, French, etc.), whereas others (i.e. Italian, Spanish, etc.) are null-subject languages. This characteristic requires distinct marking of referential continuity in the textual use of language, with less need to repeat anaphora in the latter case (Hickmann 2002). Another key factor influencing the communication is culture as a set of values and norms that helps shape the social behavior of individuals who belong to a cultural group as well as social interaction between them. Very well known is the study in (Kendon 2004), showing that Italians use a great number of gestures when communicating.

2.2 Gesturing with hands

The gestural movements of the hands and arms are probably the most studied co-speech gestures (Wagner, Malisz, and Kopp 2014). Based on the seminal works by (Kendon 1972) about the relationship between body motion and speech and by (Kendon 2011) about gesticulation and speech in the process of utterance, they are usually separated into several *gestural phases*: rest position, preparation phase, gesture stroke, holds and retraction or recovery phase (Bressemer and Ladewig 2011). More generally, gestures can be described in terms of their form, semantic and pragmatic functions, their temporal relation with other modalities, and their relationship to discourse and dialogue context.

Since hand movements serve multiple functions in communication, it is often useful to define their semantic function. One of the best known classifications in this respect is that of (McNeill 1992) which attributes five semantic functions to hand movements:

- *emblematic gestures* bear a conventionalized meaning (“thumbs up”);
- *iconic gestures* resemble a certain physical aspect of the conveyed information, e.g. they may convey the shape of a described object or the direction of a movement;
- *metaphoric gestures* are iconic gestures that resemble abstract content rather than concrete entities (McNeill 1992; Cienki and Müller 2008);
- *beat gestures* are simple and fast movements of the hands (also called batons (Ekman and Friesen 1972)).

This classification should not be understood as defining distinct categories. (McNeill 2005) argued that a simple functional classification of gestures is usually misleading. As (Wagner, Malisz, and Kopp 2014) pointed out due to the multifaceted nature of most gestures, he preferred a dimensional characterization of gestures, with dimensions including iconicity, metaphoricity, deixis, temporal highlighting (beats), and social interactivity. This acknowledges the fact that the majority of gestures can be characterized along several of these dimensions, e.g. when a pointing gesture also depicts the direction of a movement, or when a beat is superimposed onto the stroke onset of an emblematic gesture (Tuite 1993).

As we will explain more fully in Section 3.1 a further classification is proposed by (Lin 2017) adapting (Colletta et al. 2015; Kendon 2004), according to which gesture-speech relationship can assume five possible semantic functions (i.e. reinforcing, integrating, supplementary, complementary, contradictory). Since this classification can be effectively used to capture the semantic contribution of gestures the utterances, we adopt it in our study and include such classes in our classification scheme.

2.3 Using multimodal corpora for analyzing gesture and speech in interaction

The concept of a multimodal corpus has been defined by (Allwood 2008) in terms of an annotated collection of “language and communication-related material drawing on more than one modality”. Multimodal corpora (or multimedia corpora as they are often defined in the Italian literature) are used especially for pragmatic research purposes (i.e. in studies on proxemic correlates of spoken language or on the bodily manifestation of emotions), in which the starting sessions consist of videos that are transcribed and annotated (Cresti and Panunzi 2013). About what can be analyzed through the use of multimodal corpora, according to (Allwood 2001), although there are many research questions that could be answered through the use of these resources, they can be divided primarily into three major areas: *human-human face-to-face communication* (e.g. the nature of communicative gestures, multimodal communication in different national/ethnic cultures, communication and consciousness/awareness, etc.), *media of communication* (e.g. multimodality in writing, multimodality in songs and music, etc.), *applications* (e.g. better modes of multimodal human-computer communication, better modes of multimodal distance teaching/instruction, etc.).

In addition, multimodal corpora can be useful resources in the development of various computer-based applications, supporting or extending our ability to communicate, with regard to: modes of multimodal human-computer communication, better computer support for multimodal human-human communication, modes of multimodal communication for persons who are physically challenged (handicapped), modes of multimodal presentation of information from databases (for example for information extraction or for summarization), better multimodal modes of translation and interpretation, modes of multimodal distance language teaching (including gestures), better multimodal modes of buying and selling (over the internet, object presentation in shops, etc.), computerized multimodal corpora can, of course, also be useful outside of the areas of computer-based applications. In general, they can provide a basis for studying any type of communicative behavior in order to fine-tune and improve that behavior.

However these resources – probably due to the difficulty of construction – in Italy are difficult to find and consult, in fact between the 286 multimodal resources certified for all the languages by the LRE map¹ only one is in Italian, IMAGACT, a corpus-based ontology of action concepts, derived from English and Italian spontaneous speech (Moneglia et al. 2014; Bartolini et al. 2014). So this language is not well represented.

As specified in the section 1 – given that the television interview is inherently a multimodal and multisemiotic text, in which meaning is created through the intersection of visual elements, verbal language, gestures, and other semiotic cues (Vignozzi 2019) –

1 LRE map (Language Resources and Evaluation) is a freely accessible large database on resources dedicated to Natural language processing. The original feature of LRE Map is that the records are collected during the submission of different major Natural language processing conferences. The records are then cleaned and gathered into a global database called “LRE Map” (Calzolari et al. 2010). The map is freely available from the site <https://lremap.elra.info/>

this study focuses on the co-gesture behavior of several Italian politicians during TV face-to-face interviews, this therefore requires not only the presence of a multimodal resource but also of political domain.

In particular, non-verbal aspects acquire considerable importance especially in debates and interviews in the political domain, which is the area that is most suitable for this type of analysis (Seiter and Harry Jr. 2020). One of the particularly successful lines of research in recent years in the political domain is the analysis of gestures used by the speaker with the function of discrediting the opponent. These aspects have been the subject of various studies even in Italian language (D’Errico, Poggi, and Vincze 2013, 2012).

Concerning Italian language, some corpora have been made available recently, the largest one includes around 3,000 public documents by Alcide De Gasperi (Tonelli, Sprugnoli, and Moretti 2019) that has been mainly used to study the evolution of political language over time (Menini et al. 2020). All the corpora cited above are monomodal and none of them takes into account gestural traits. Indeed, corpora that include only one modality have a long tradition in the history of linguistics. According to (Lin 2017) “the construction and use of multimodal corpora is still in its relative infancy. Despite this, work using multimodal corpora has already proven invaluable for answering a variety of linguistic research questions that are otherwise difficult to consider”.

Furthermore, none of the multimodal resources currently available in Italian present a systematic annotation of gestures, since is not possible to construct a state of the art on the presence and behavior of co-gestural patterns for this language.

2.4 PoliModal corpus: description and new layer of annotation

PoliModal corpus (Trotta et al. 2019, 2020) contains transcripts of 56 TV face-to-face interviews of 14 hours, taken from the Italian political talk show “In mezz’ora in più” broadcast between 2017 and 2018, for a total of 100,870 tokens. The corpus has a double level of annotation using XML as markup language. The first one was done manually following the TEI standard for Speech Transcripts in terms of utterances and takes into account the “speech constant” (Voghera 2001). In particular:

(a) **Metadata**: these include useful information for a quick identification of transcriptions, for example the tools used for the transcription, a link to the interview, the owner account, the title of the talk show, the date of airing, the guests, etc.

(b) **Pause**: this tag is used to mark a pause either between or within utterances;

(c) **Semi-Lexical**: this tag is used to label interjections (i.e. ‘eh’, ‘ehm’ etc.), or more generally words that convey the meaning of an entire sentence, constituting a complete linguistic act demonstrated by their paraphrasability;

(d) **FalseStart**: this tag shows the speaker’s abandonment of an already produced word or sequence of words, with or without repetition of previously used linguistic material;

(e) **Repetition**: with this tag are marked cases of repetition of utterances in order to give coherence and cohesion to the speech or self-repetition as a control mechanism of the speech programming;

(f) **Truncation**: truncation indicates the deletion of a phoneme or a syllable in the final part of a word.

This annotation task addressed so far falls – from a qualitative point of view – in the first of the general types identified by (Mathet, Widlöcher, and Métivier 2015), in which the subjective interpretation is limited. Indeed, it deals with the “identification of units” (Krippendorff 2018), in which the annotator, given a written or spoken text, must iden-

tify the position and boundary of linguistic elements (e.g. identification of prosodic or gestural units, topic segmentation). In order to evaluate the reliability of our annotation scheme, we computed inter-annotator agreement by performing a double annotation of verbal and non-verbal traits of the first ten minutes of Renzi's, Di Maio's and Salvini's interview. Both annotators were expert linguists. Macro-averaged F1 computed on exact matches amounts to 0.82, which corresponds to a good agreement, given that by exact match we consider the correct choice of the trait, the position of the tag and the exact extension of the marked string, if any. This result confirms the reliability of the task and the corresponding annotation guidelines.

The second annotation level was performed automatically using ANVIL (Kipp 2001) – a tool for the annotation of audiovisual material containing multimodal dialogue – following the MUMIN (Allwood et al. 2007) annotation scheme that takes into account ten types of gestures divided into three categories:

(a) **Facial displays:** they refer to timed changes in eyebrow position, expressions of the mouth, movement of the head and of the eyes (Cassell and others 2000). The coding scheme includes features describing gestures and movements of the various parts of the face, with values that are either semantic categories such as Smile or Scowl or direction indications such as Up or Down.

(b) **Hand gesture:** we follow a simplification of the scheme from the McNeill Lab². The features, 7 in total, concern Handedness and Trajectory, so that we distinguish between single-handed and double-handed gestures, and among a number of different simple trajectories analogous to what is done for gaze movement. The value Complex is intended to capture movements where several trajectories are combined.

(c) **Body posture:** this tag comprises trajectory indications for the movement of the trunk. The categories are mutually exclusive to facilitate the annotation work.

The annotation – made at the moment by a single expert annotator – follows the criterion highlighted by (Allwood et al. 2007), claiming that annotators are expected to select gestures to be annotated only if they have a communicative function. In other words, gestures are annotated if they are either intended as communicative by the communicator (displayed or signalled) (Allwood 2001), or judged to have a noticeable effect on the recipient.

However, this last level of annotation does not take into account the semantic functions covered by these gestures and therefore would not allow to develop an in-depth analysis of the semantic contribution they could make to the discourse. So – as we will explain in depth in the Section 3.1 – we manually add a new level of annotation that takes into account the semantic functions covered by one of the gestures already tagged in the corpus: hand movements.

3. Methodology

3.1 Coding co-speech gesture in PoliModal corpus

In the paper by (Allwood 2001), the authors highlight that synchronization of information in different modalities is a crucial issue in assembling a multimodal corpus. Therefore the authors suggest to adopt the general principle of spatio-temporal contiguity. This means that a text occurs at the same point in time as the event it describes or represents. When temporal contiguity concerns the relation between transcribed speech

² Duncan, S. (2004). Coding manual. Technical Report available from <http://www.mcneilllab.uchicago.edu>.

(or gesture) and recorded speech (or gesture), it is often referred to as “synchronized alignment” of recording and transcription. What synchronization means is that for every part of the transcription (given a particular granularity), it is possible to hear and view the part of the interaction it is based on and that for every part of the interaction, it is possible to see the transcription of that part. The form of connection between the transcriptions and the material in the recordings can vary from just being a pairing of a transcription and video or audio recording, where both recording and transcription exist but they have not yet been synchronized, to being a complete temporal synchronization of recordings and transcription. In our case, audio and video signals as well as the annotations have been temporally synchronized by hand. Although the most convenient solution for synchronization is to carry it out using a computer program already while making the recording (see for example the AMI project and CHIL project), we did it manually since the recording and transcription of the corpus were done before knowing what layers would be exactly annotated.

Starting from PoliModal corpus described in 2.3, we manually add a new level of annotation that takes into account the semantic functions covered by one of the gestures already tagged in the corpus: hand movements. This is because the gestural movements of the hands and arms, i.e. spontaneous communicative movements that accompany speech (McNeill 2005), are probably the most studied co-speech gestures (Wagner, Malisz, and Kopp 2014). Based on the seminal works by (Kendon 1972) about the relationship between body motion and speech and by (Kendon 2011) about gesticulation and speech in the process of utterance, they are usually separated into several *gestural phases*: rest position, preparation phase, gesture stroke, holds and retraction or recovery phase (Bressemer and Ladewig 2011). Additionally, the point of maximal gestural excursion is often regarded as a *gestural apex*.

In PoliModal the **hand movement trajectory** tag indicates only the start and end of the movement in terms of time and the trajectory of the gesture, in particular *up*, *down*, *sideways*, *complex*. In order to keep track also of the semantic function covered by the tag, we added an additional information layer to those already present – following the classification proposed by (Lin 2017) adapting (Colletta et al. 2015) and (Kendon 1972) – which attributes five functions to hand movements:

- *Reinforcing*: the information brought by the gesture is equal to the linguistic information it is in relation with. For example, one of the interviewees emphasizes the sacrifices to which Italians have been subjected in the last fifteen years, including “il 3% del rapporto deficit/PIL” (*en.* “the 3% deficit/PIL ratio”). In saying this he makes the sign of the number three with the fingers of his right hand.
- *Integrating*: the information provided by the gesture does not add supplementary information to the verbal message, but makes the abstract concepts more precise. A frequent example in our annotation is when a politician, in order to contrast two items such as left and right parties, points one of his hands toward the right and the other toward the left.
- *Supplementary*: the information brought by gestures adds new information not coded in the linguistic content. For example, in one of the interviews, the interviewee comments on the amount of members of Parliament elected from another party saying “... non so quanti parlamentari porterà in Parlamento” (*en.* “... I don’t know how many MPs they will bring to

Parliamen”) and in the meantime he opens his arms as if to imply a large number.

- *Complementary*: the information provided by the gesture brings a necessary complement to the incomplete linguistic information provided by the verbal message. The gesture usually disambiguates the message, for example, in our annotation it is common to find cases where deictic adverbs such as *qui* (en. here) are accompanied by the corresponding pointing gesture.
- *Contradictory*: the information provided by the gesture contradicts the linguistic information provided by the verbal message. This kind of gesture was not found in our annotation.
- *Other*: within this category we include all the gestures that annotators were not able to classify with the above mentioned semantic labels.

Our annotation follows the selection criterion highlighted by (Allwood et al. 2007), claiming that annotators are expected to select gestures to be annotated only if they have a communicative function. However, as (Yoshioka 2008) points out gestures can be functionally ambiguous and thus have multiple semantic functions simultaneously. According to (Tsui 1994), the source of this multiple functions often lies in the sequential environment of the conversation in which the utterance occurs. To simplify the task, annotators are therefore asked to assign a single semantic function to the gestures under investigation, choosing the function that s/he considers prevalent in the context of use.

In order to evaluate the reliability of our annotation scheme, we compute inter-annotator agreement by performing a double annotation of the semantic functions listed above on three of the interviews considered (Matteo Renzi, Luigi Di Maio, Matteo Salvini) for a total of about 2 hours of interviews. Both annotators (one male and one female) are expert linguists. Macro-averaged F1 computed on exact matches amounts to 0.83, which corresponds to an almost perfect agreement. This result confirms that the task is well-defined and that the corresponding annotation guidelines are clear.

Figure 1 shows an example annotation with the new information layer specified the semantic function (tag ‘function’). For each observed gesture, the PoliModal corpus already contained: i) the start and end point in the video in terms of milliseconds; ii) the type of gesture observed; iii) the movement trajectory. We add to this the semantic function covered by the gesture in the context.

```
<u gender="m" length="928" role="Minister of the foreign
business and of the international cooperation" time=
"452.28" who="Angelino Alfano">C'è qualcosa di più grave
e di più profondo di cui mi sono occupato da Ministro
dell'Interno. Perché io ho gestito l'immigrazione
<movement start="470.2" end="471.2" attribute="Hand
movement trajectory" attribute_text="sideways" function=
"integrating">e ho gestito l'ordine pubblico.
</movement></u>
```

Figure 1
Annotation sample in xml

4. Systematical co-occurrence of hand-movement and specific categories of verbs

The study presented in (Vignozzi 2019) aimed to analysing the representation of some peculiar indicators of spokenness (i.e. idiomatic expressions and phrasal verbs) across TV interviews featuring different interviewees (politicians, business people and personalities from showbiz). The analysis pointed out that phrasal verbs are more recurrent in political interviews than in business and economic discussions, and that the specialized domain with which hand or arm movements are more often associated is again business and economics (60.86%). In political interviews, instead, gestures appear in 58.02% of cases, while in showbiz interviews the lowest frequency is observed, since gestures occur only in 40.42% of the cases. Besides, the study shows that beats gestures are the most frequent kind of gestures co-occurring with phrasal verbs, especially in political interviews, where they account for more than half of the total of gestures. The study was conducted on “The ESP Video Clip Corpus” in English.

In order to understand whether hand gestures (identified by the tag *hand movement trajectory*) is related in a systematic way to particular types of verbs (e.g. predicative, phrasal verbs etc.), we created a subcorpus containing only the sentences of the interviews co-occurring with the tag under investigation were extracted (for a total of 495 sentences).

The qualitative approach has been preferred in this phase for two main reasons: first of all, because the amount of data to be analyzed is controllable; moreover because existing resources for Italian such as LexIt (Lenci, Lapesa, and Bonansinga 2012), MultiWordNet (Pianta, Bentivogli, and Girardi 2002) and T-PAS (Jezek et al. 2014) do not make explicit the function that the verbs assume in the context (e.g. no tool will tell us if the verb is servile, appellative, estimative, elective, etc.).

Through a qualitative analysis, we then manually classified verbs according to their function in the text (Jezek 2003). The verbal classes identified are as follows (with the total number of occurrences in parenthesis):

- Predicative verbs: they have full lexical meaning and can independently give rise to a verbal predicate of full meaning. The class of predicative verbs encompasses the vast majority of verbs in a language, and is descriptively opposed to the class of copulative verbs that need to rely on a predicative complement to fulfill the predicate function: *sembrare* [to appear] (13), *parere* [to seem] (5), *risultare* [to result] (4), ***stare*** (131), *restare* (7), *rimanere* (2) [to stay, to remain], *diventare* (5) [to become]
- Predicative verbs which can carry a predicative complement of the subject, but only if conjugated in the passive form: *chiamare* [to call] (2), *eleggere* [to elect] (2), *giudicare* [to judge] (1) and ***fare*** [to do] (12).
- Phrasal verbs are verbs that, when combined with another non-finite mode verb with the interposition of a preposition (to, of, for, from), specify a particular time-expectant mode. They are divided into 5 groups:
 - the imminence of an action: *stare per* (3) + infinitive
 - the beginning of an action: *cominciare a* [begin to] (7) + infinitive
 - the development of an action: ***stare*** [*stay*] (38) and *venire* (15) [come] + *gerund*

- the duration and continuity of an action: *continuare a [continue to] (6) + infinitive*
- the conclusion of an action: *finire di (1) and smettere di (1) [stop to] + infinitive*
- Causative verbs: indicate that the action is caused by the subject, but that he does not perform it directly. The only causative of the Italian language that occurs in the corpus is the verb *fare [to do] (20) + infinitive*
- Performative verbs: they exist only in the first person singular of the present indicative and are so defined because pronouncing them is equivalent to performing the action they describe, i.e. to perform the action they describe one must pronounce them. The only verb belonging to this class present in the corpus is *negare [to deny](1)*.³ The other verb taking a performative function in the first person of the present indicative is *dire [to say] (26)*.

Most function verbs are predicative, that is, they have an independent meaning, forming what in syntax is called a verbal predicate. Among them we notice a more frequent use of the verb **stare** [to stay] with 131 occurrences.

- (4) Salvini: “*Ci possono essere altre sfumature, a qualcuno **sta** simpatico Macron, a qualcuno **sta** simpatica la Le Pen, è il rapporto con l’Europa che per me è determinante al di là delle simpatie.*” (en. “There may be other nuances, someone **likes** Macron, someone **likes** Le Pen, it is the relationship with Europe that for me is decisive beyond sympathies.”)

Among verbs with a predicative function of the subject (only when used in the passive form), the most commonly used are effective verbs, i.e. copulative verbs indicating a state, semblance, or transformation. In this case the most frequent is **fare** [to do] with 12 occurrences.

- (5) Padoan: “*Secondo te questa campagna elettorale sta dividendo il paese in due. Tra chi vuole continuare e rafforzare quello che è **stato fatto** e ha portato i risultati che lei ricordava, piuttosto che chi vuole eliminare.*” (en. “According to you, this election campaign is dividing the country in two. Between those who want to continue and strengthen what **has been done** and has brought the results that you recalled, rather than those who want to eliminate.”)

On the other hand, with respect to phrasal verbs, the results obtained do not confirm what emerged in (Vignozzi 2019), in which a predominance of servile verbs was noted in political domain interviews, because in our case there is a slight but not clear prevalence of verbs that indicate the performance of an action, in particular of the verb **stare** [to stay] + gerund with 38 occurrences.

³ See for example the utterance by Di Battista: *Io ho avuto credo 84 giorni di espulsione dalla Camera dei Deputati e non ho mai picchiato nessuno, mai. Anche se non le nego...* (en. I’ve had I think 84 days of expulsion from the House of Representatives and I’ve never hit anybody, ever. Although I don’t deny them...).

- (6) Veltroni: “*E quello che sta succedendo in Italia, l’affermazione non delle forze tradizionali...*” (en. “And what **is happening** in Italy, the assertion not of traditional forces...”)

Among causative verbs, the most present is the verb **fare** [to do] (20 occurrences), while the among performative ones it is **dire** [to say] (26).

- (7) Tremonti: “*E quando comincio a vedere che perfino Prodi parla di un colpo di quel tipo, avremmo dovuto andare a votare e non **ci hanno fatto** andare a votare. Perché dovevano mandarci il Governo tecnico che tecnicamente ci ha buttato giù.*” (en. “And when I start to see that even Prodi is talking about that kind of hit, we should have gone to vote and they **didn’t make us to go to vote**. Because they had to send us the technical government that technically brought us down.”)
- (8) Di Maio: “*Guardi io le **dico** noi parleremo con tutti coloro che aderiranno però...*” (en. “Look **I’ll tell you** we’re going to talk to everyone who joins though...”)

Causative verbs are verbs that express an action not performed by the subject, but made to be performed by others. In this case, we notice a prevalence of the verb **fare** [to do], mainly used with a negative valence and referred to the political opposition; in fact, this verb mainly describes actions that the subjects were forced to carry out because of the determined political circumstance of the moment.

The concept of performative act was introduced by the theory of linguistic acts elaborated in (Austin 1975). Verbs that take on this function are so defined because pronouncing them is equivalent to performing the action they describe. In other words, in order to perform the action they describe, one must pronounce them. Probably the performative verb **dire** is more present in these interviews because - being in the middle of an electoral campaign - politicians want to give an impression of being concrete and aim at emphasising their statements.

5. Is the Lexical Retrieval hypothesis confirmed?

Many studies have suggested that gestures, especially representational gestures (Krauss and Hadar 1999) play a direct role in speech production by priming the lexical retrieval of words. This view has been termed the *Lexical Retrieval hypothesis*.

The hypothesis is based on research arguing that (1) gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewellyn 1969; Butterworth and Beattie 1978), and (2) that the inability to gesture can cause verbal disfluencies (Dobrogaev 1929). In addition – as (Krauss 1998) pointed out – speakers were more disfluent overall in constrained-speech conditions than in natural conditions. Since the corpus used as the object of study presents a level of annotation that takes into account some hesitation pauses and verbal disfluencies, we decided to verify this hypothesis in the political domain, where speakers usually have to control well their communication and be persuasive.

We compute weighted mutual information between hand movements and each of the speech disfluencies reported in Table 1. This measure is calculated to show existing mutual dependencies between co-occurring tags. We consider only the interviews in the PoliModal corpus that have a minimal length of 50 turns, so to have a good amount of

annotations to consider. We report in Table 1 the tag incidence per 100 turns for each interview considered.

Table 1
Tag incidence per 100 turns for each interview

<i>Interviewee</i>	<i>Hand mov.</i>	<i>Pause</i>	<i>Semi-Lexical</i>	<i>FalseStart</i>	<i>Repetit.</i>	<i>Truncat.</i>
Matteo Renzi	35.82	0	8.50	10.16	22.45	36.89
Luigi Di Maio	22.97	0	14.86	0	18.91	18.91
Matteo Salvini1	54.38	5.20	24.56	0	24.56	19.29
Matteo Salvini2	52.87	14.94	21.83	3.44	21.83	3.44
Walter Veltroni	41.81	0	14.54	21.81	29.09	18.18
Simone Di Stefano	10.98	0	4.39	5.49	21.97	16.48
Pierluigi Bersani	32.29	1.04	26.04	0	31.25	20.83
Angelino Alfano	57.00	9.00	33.00	3.00	17.00	3.00
Giulio Tremonti	10.71	16.07	10.71	0	14.28	0
Matteo Orfini	29.85	1.49	11.94	0	14.92	0
Pier Carlo Padoan	49.27	11.94	30.43	1.44	7.24	13.5
Carlo Calenda	74.63	32.60	24.63	9.42	7.24	0.72
Alessandro Di Battista	39.02	9.26	32.19	6.82	11.70	10.58
Average	39.35	7.81	18.89	4.74	17.74	12.45

Among the politicians included in this dataset, the one that most accompanies his speech with the movements of the hands is Matteo Salvini (Lega) considering both interviews, followed by Carlo Calenda (PD) and Angelino Alfano (Il Popolo della Libertà). Their belonging to different political parties suggests that the use of hand movements is more an individual trait than a feature characterising specific political positions.

Weighted mutual information (WMI) is computed between hand movements and tags reported in Table 1. The values obtained are shown in the heatmap reported in Figure 2, with lighter colors corresponding to higher WMI values.

Overall, hand movements tend to have a higher association with semi-lexical traits and pauses, which would confirm the assumptions of *Lexical Retrieval hypothesis* according to which gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewellyn 1969; Butterworth and Beattie 1978). Indeed, semi-lexical expressions, such as ‘ah’, ‘eh’, ‘ehm’, have been associated with the fact that linguistic planning is very cognitively demanding, and it is difficult to plan an entire utterance at once. This effect is however not present for some politicians, such as Di Battista and Alfano, while it is evident for some others such as Bersani and Salvini. Therefore, our findings are not generally applicable to all interviewees in our corpus. Fig. 2 shows also evident differences in gesturing behavior among the considered politicians. For instance, although Carlo Calenda and Angelino Alfano present a high incidence of hand movements, they do not seem to be associated with specific tags. Matteo Renzi, instead, shows a gesturing behavior that is unique compared to all the other interviewees, with hand gestures that are almost always used in association with other speech phenomena.

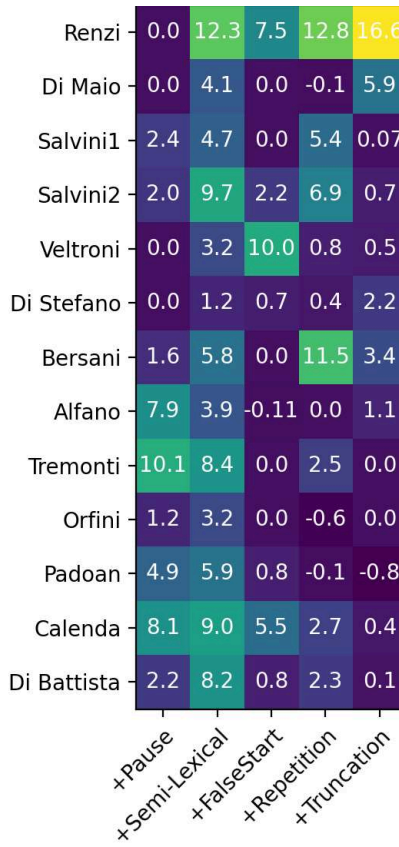


Figure 2
WMI values between hand movements and tags reported on the x-axis for each interviewee on the y-axis

In the interviews, we observe also the presence of negative values for WMI obtained in relation to false-starts (-0.11), repetitions (-0.1 and -0.6) and truncations (-0.8), suggesting that hand movements are less likely to be accompanied by such linguistic phenomena.

Notice that the results are consistent with the Tradeoff Hypothesis (De Ruiter, Bangerter, and Dings 2012). Qualitative analysis shows that when respondents are more disfluent in speech, they gesticulate more. This behavior reflects what is stated in the hypothesis “when gesturing gets harder, speakers will rely relatively more on speech, and when speaking gets harder, speakers will rely relatively more on gestures”.

6. Is the gesture-speech relationship influenced by linguistic variables?

The third analysis carried out was aimed to understand if the hand movements produced by the interviewees have significant correlations with language complexity. As in the previous analysis, a threshold was established, therefore only interviews with a minimal length of 50 turns was taken into account.

For complexity we consider the type-token ratio and the average lexical density, i.e. the number of content words divided by the total number of tokens. We do not take into account the Gulpease index (Lucisano and Piemontese 1988), despite it is considered the standard metric of readability in Italian. But its reliability is undermined by several limitations (Tonelli, Tran Manh, and Pianta 2012) like sentence length and polysyllabic words; in addition it has been specifically designed, not very suitable for transcripts.

We perform an analysis of the correlation between language complexity and hand movements, normalised by the number of tokens uttered by each politician multiplied by one thousand. Since the variables under examination are both cardinal or quantitative, the Person's correlation index had been used for each interviewee and for each political party they belong to.

Table 2
Normalized values of hand movements, TTR, and lexical density for each interviewee

<i>Interviewee</i>	<i>Hand movement</i>	<i>TTR</i>	<i>Lexical Density</i>
Matteo Renzi	35.82	0.71	0.563
Luigi Di Maio	22.97	0.8	0.562
Matteo Salvini1	54.38	0.73	0.567
Matteo Salvini2	52.87	0.82	0.569
Walter Veltroni	41.81	0.7	0.569
Simone Di Stefano	10.98	0.75	0.583
Pierluigi Bersani	32.29	0.73	0.547
Angelino Alfano	57	0.61	0.564
Giulio Tremonti	10.71	0.75	0.585
Matteo Orfini	29.85	0.72	0.566
Pier Carlo Padoan	49.27	0.75	0.570
Carlo Calenda	74.63	0.73	0.580
Alessandro Di Battista	39.02	0.8	0.568

Individual interviewee computations reveal that both the TTR and the conceptual density show a moderate negative correlation with hand movements, respectively $r = -0.3$ and $r = -0.12$. Since in all cases considered the correlation is negative it could deduce that the Information Retrieval hypothesis is confirmed. The value of the TTR could mean that the more you gesticulate the more the lexical richness decreases and therefore there are more hesitations.

Instead in the case of conceptual density, the negative value $r = -0.12$ could mean that the more you gesticulate the more the speech tends to be simple and understandable (this could find even more justification in the format of the interview that being televised and being broadcast at a time when the audience is quite varied, it could tend to be easier to be understood by all).

Also political parties computations reveal that both the TTR and the conceptual density show a moderate negative correlation with hand movements, respectively $r = -0.7$ and $r = -0.71$ even if slightly higher than the correlation per single respondent with a deviation of 0.5 for TTR and 0.6 for conceptual density. The correlation values obtained by political party of belonging show a slight negative correlation, which could mean that the party of belonging does not significantly influence the use of the semantic communication plan and consequently the use of language.

Table 3

Values of hand movements, TTR, and lexical density for each political party

<i>Political Party</i>	<i>Avg. Hand movement</i>	<i>Avg. TTR</i>	<i>Avg. Lexical Density</i>
PD	43.94	0.73	0.566
M5S	30.99	0.80	0.565
Lega	39.32	0.76	0.574
CasaPound	10.98	0.75	0.583
Il Popolo della Libertà	57	0.61	0.564

Therefore, the first correlation values obtained allow us to state that the gesture-speech relationship is not influenced either by the political party or by the linguistic variables considered.

7. What are the semantic patterns of gesture-speech relationship?

A summary of the hand movement annotations in the corpus is reported in Table 4 and 3. In the first one, the number of annotated tags is reported for each politician, while in the second table the values are aggregated by political party. The parties include PD (left-center), Movimento 5 Stelle (center-populist), Lega (right-populist), Casa Pound (right), Popolo delle Libertà (center-right). The “Contradictory” category is not reported in the tables because it was never found in the interviews. This is probably due to the fact that in political interviews broadcast on TV, politicians try to be as clear as possible, avoiding statements and behaviour that may be misunderstood. Therefore, gestures and speech that are in contradiction are generally avoided. Probably for the same reason, supplementary movements, adding new information that is lacking in the linguistic content, are not frequent. ‘Integrating’ movements, instead, can be seen as an attempt to emphasise the speech content without adding supplementary information. This type of movement is the most frequent one, followed by “Complementary”.

A qualitative analysis of the single interviews shows interesting differences in attitude and communication style, which pertain to single politicians rather than to party positions. Matteo Renzi, for example, uses gestures very frequently to accompany his speech. We report an example of ‘Integration’ below:

Matteo Renzi: *“Quello che sta accadendo invece in queste settimane, in questi mesi, conferma che c’è una grande distanza tra la politica dei palazzi e la politica della quotidianità [integrating].”*

(Eng. *“Instead what is happening in these weeks, in these months, confirms that there is a great distance between the politics of the Palaces and the politics of everyday life.”*)

Renzi underlines that the distance between politics made by elites, detached from the real problems of the country (“politics of the Palaces”), and politics of everyday life, that is, attentive to reality and to citizens, is increasingly evident. Gesture is used to stress this difference: the speaker’s open right hand points away from his torso in correspondence with the metaphorical expression politics of the Palaces, almost as if to indicate that it is something in which he does not recognize himself. His right hand then immediately rejoins his left hand and points downwards at the moment in which the expression politics of everyday life is pronounced, as if to indicate a politics that is instead attentive to relevant and concrete things.

Concerning the *Reinforcing* type of gesture-speech relationship, it is mainly used to reiterate a concept already expressed linguistically, and it is not very used, probably because it may seem redundant. Angelino Alfano turns out to be the interviewee who makes most use of this type of gesture. In this example, Alfano, talking about the consensus obtained by one of his political opponent Matteo Salvini, claims that this consensus was obtained at his expense. So, in saying “contro di me” (against me), the open hands are close to his bust.

Angelino Alfano: “*Quindi la sfida di Salvini, avendo aggregato consenso – contro di me peraltro [reinforcing] – sull’immigrazione, è incanalarlo su un regime di legislazione democratica.*”

(Eng. “*So Salvini’s challenge, by aggregating consensus – against me by the way – on immigration, is to channel it on a regime of democratic legislation.*”)

As mentioned above, *Supplementary* gestures are used with a very low frequency. One of the few examples in the corpus is present in Simone di Stefano’s interview, where he is asked to clarify the alleged relations of the party with a convicted member of the Mafia. The interviewee tries to provide an explanation, but the interviewer continues to put him under pressure. At this point the interviewee lowers his gaze and moves his open right hand away from his torso while saying “*but I don’t want to avoid [your question]*”, as if to implicitly ask the journalist to stop her suppositions and let him explain his position.

Complementary gestures bring a necessary complement to the incomplete linguistic information provided by the verbal message. They are frequently used by the respondents in the corpus under analysis, in most cases to disambiguate the message or simply some linguistic elements. This indicates the speaker’s intention to be as clear as possible. For example, at the beginning of the interview with Carlo Calenda, he is shown a photo that portrays him wearing a worker’s helmet. The interviewee refers to the photo by pointing with his left hand away from his torso to the screen where the photo is displayed, making it easier for viewers to understand what he was referring to:

Carlo Calenda: “*Benché gli operai non si sentiranno, come posso dire, contenti dopo aver visto la mia foto con quel caschetto [complementary] in cui sembravo un totale ebete.*”

(Eng. “*Although the workers won’t feel, how can I say, happy after seeing the picture of me in that helmet where I looked like a total stupid.*”)

As noted above, a residual category has been added to the tags. The *Other* category includes all the gestures that annotators were not able to classify with the above mentioned semantic labels. This problem was found most frequently in the interviews with Pier Carlo Padoan and Carlo Calenda. These gestures are different from the others because they show a *batonic* value, that is, they are used to mark the rhythm of the enunciation, for example by tapping a finger on the table.

8. Conclusion

This paper investigate co-gesture speech of several Italian politicians during face-to-face interviews. To this purpose, we enrich PoliModal – a multimodal Italian political domain corpus – with a new layer of annotation, describing the semantic function of the different hand movements.

Concerning the type of verbs used – which in Italian can be broadly distinguished in predicative, copulative, auxiliary, phrasal, performative and causative (Jezek 2003) – it was noticed that: among the verbs with a predicative function of the subject the most commonly used are effective verbs, i.e. copulative verbs indicating a state, semblance, or transformation; with respect to phrasal verbs, the results obtained do not confirm what

Table 4
Frequency of the type of gestures annotated for each interviewee

<i>Interviewee</i>	<i>Integrat.</i>	<i>Reinforc.</i>	<i>Supplement.</i>	<i>Complement.</i>	<i>Other</i>
Matteo Renzi	32	9	2	23	1
Luigi Di Maio	6	0	1	9	1
Matteo Salvini1	16	6	3	5	1
Matteo Salvini2	17	10	0	14	5
Walter Veltroni	8	3	0	8	4
Simone Di Stefano	5	0	2	3	0
Pierluigi Bersani	13	4	0	12	2
Angelino Alfano	21	11	1	16	8
Giulio Tremonti	3	1	1	1	0
Matteo Orfini	7	0	0	10	3
Pier Carlo Padoan	16	0	0	3	15
Carlo Calenda	41	1	0	35	26
Alessandro Di Battista	29	1	0	20	0
Total	214	46	10	159	66

emerged in (Vignozzi 2019), in which a predominance of servile verbs was noted in political domain interviews, because in our case there is a slight but not clear prevalence of verbs that indicate the performance of an action, in particular of the verb **stare** + gerund with 38 occurrences. Among causative verbs, the verb **fare** (20 occurrences) is the one that occurs most frequently, while the among performative ones it is **dire** (26). Causative verbs has been detected a prevalence of the verb **fare**, mainly used with a negative valence and referred to the political opposition; in fact, this verb mainly describes actions that the subjects were forced to carry out because of the determined political circumstance of the moment. Other evidence is in favor of performative verb **dire** probably more present in these interviews because – being in the middle of an electoral campaign – politicians want to give an impression of being concrete and aim at emphasising their statements.

Furthermore, we test the *Lexical Retrieval Hypothesis* by computing the association between hand movements produced by each interviewee and speech disfluencies using *weighted mutual information*. Results show that hand movements tend to co-occur with full pauses (i.e. repetition) and empty pauses (i.e. pause) and more frequently with interjections (i.e. semi-lexical), suggesting that gesticulating may represent an attempt at lexical retrieval. In future developments we plan to extend the analysis taking into account more recent theories, e.g. the Tradeoff Hypothesis (De Ruiter, Bangertter, and Dings 2012), more general and empirically better supported.

Concerning gesture-speech relationship, the results obtained suggest that hand movements are mainly used with an integrative and complementary functions. So, the information provided by such gestures adds precision and emphasis to linguistic information.

Finally we perform an analysis of the correlation between language complexity and hand movements. Individual interviewee computations revealed negative correlation values for both TTR and conceptual density, further confirming Information Retrieval and letting us assume that probably the more you gesticulate the more your lexical

richness decreases, leading to more hesitation in speech. At the same time, the negative correlation values obtained for lexical density might suggest that the more the speaker makes use of gestures in his speech, the simpler and more comprehensible it tends to be. However, concerning the correlation by political party, again negative correlation values were obtained for both TTR and conceptual density, suggesting that party affiliation would not influence the use of gestures.

In the future we plan to make this new level of annotation freely accessible in order to make possible both comparative studies in other languages and other fields of knowledge such as political science. In addition, we will initiate a predictive study aimed at understanding which of the variables under investigation may be effective predictors of the occurrence of hand movements. A further future development could be to use a comparison between sentences with hand movements and those ones in which no movements are present – through the creation of two different subcorpora – in order to understand if the increase in complexity of language is accompanied by a parallel growth of gestures with the aim of increasing clarity of speech.

A further aspect that we propose to investigate concerns the function of gestures to discredit the opponent in political debates. This topic has been much discussed in the literature, both with regard to rhetorical and persuasive aspects, and with particular focus on multimodal communication (D'Errico, Poggi, and Vincze 2013, 2012; D'Errico and Poggi 2012; D'Errico 2019). Currently these aspects have not been considered because they are not present in the sample used as the object of analysis. However, given the nature of the interviews composing the corpus, may be a promising line of research.

References

- Allwood, Jens. 2001. Dialog coding - Function and grammar: Göteborg coding schemas. *Gothenburg Papers in Theoretical Linguistics*.
- Allwood, Jens. 2008. Multimodal corpora. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter.
- Allwood, Jens, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The mummin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287.
- Austin, John Langshaw. 1975. *How to do things with words*, volume 88. Oxford university press.
- Bartolini, Roberto, Valeria Quochi, Irene De Felice, Irene Russo, and Monica Monachini. 2014. From synsets to videos: Enriching italwordnet multimodally. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3110–3117, Reykjavik, Iceland, May. European Language Resources Association (ELRA).
- Bressemer, Jana and Silva H. Ladewig. 2011. Rethinking gesture phases: Articulatory features of gestural movement? *Semiotica*, 2011(184):53–91.
- Butterworth, Brian and Geoffrey Beattie. 1978. Gesture and silence as indicators of planning in speech. In *Recent advances in the psychology of language*. Springer, pages 347–360.
- Calzolari, Nicoletta, Claudia Soria, Riccardo Del Gratta, Sara Goggi, Valeria Quochi, Irene Russo, Khalid Choukri, Joseph Mariani, and Stelios Piperidis. 2010. The LREC map of language resources and technologies. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).
- Cassell, Justine et al. 2000. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents*, 1.
- Cienki, Alan and Cornelia Müller. 2008. *Metaphor and gesture*, volume 3. John Benjamins Publishing.
- Colletta, Jean-Marc, Michele Guidetti, Olga Capirci, Carla Cristilli, Ozlem Ece Demir, Ramona N. Kunene-Nicolas, and Susan Levine. 2015. Effects of age and language on co-speech gesture

- production: an investigation of french, american, and italian children's narratives. *Journal of child language*, 42(1):122–145.
- Cresti, Emanuela and Alessandro Panunzi. 2013. *Introduzione ai corpora dell'italiano*. Il mulino.
- De Ruiter, Jan P., Adrian Bangerter, and Paula Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4(2):232–248.
- Dittmann, Allen T. and Lynn G. Llewellyn. 1969. Body movement and speech rhythm in social conversation. *Journal of personality and social psychology*, 11(2):98.
- Dobrogaev, Sergej M. 1929. Uchenie o reflekse v problemakh iazykovedeniia [observations on reflexes and issues in language study]. *Iazykovedenie i materializm*, pages 105–173.
- D'Errico, Francesca. 2019. 'Too humble and sad': The effect of humility and emotional display when a politician talks about a moral issue. *Social Science Information*, 58(4):660–680.
- D'Errico, Francesca and Isabella Poggi. 2012. Blame the opponent! Effects of multimodal discrediting moves in public debates. *Cognitive Computation*, 4(4):460–476.
- D'Errico, Francesca, Isabella Poggi, and Laura Vincze. 2012. Discrediting signals. A model of social evaluation to study discrediting moves in political debates. *Journal on Multimodal User Interfaces*, 6(3):163–178.
- D'Errico, Francesca, Isabella Poggi, and Laura Vincze. 2013. Discrediting body. a multimodal strategy to spoil the other's image. In *Multimodal Communication in Political Speech Shaping Minds and Social Action: International Workshop, Political Speech*, volume 7688, pages 181–206. Springer, November.
- Ekman, Paul and Wallace V. Friesen. 1972. Hand movements. *Journal of communication*, 22(4):353–374.
- Gregersen, Tammy, Gabriela Olivares-Cuhat, and John Storm. 2009. An examination of l1 and l2 gesture use: What role does proficiency play? *The Modern Language Journal*, 93(2):195–208.
- Hickmann, Maya. 2002. *Children's discourse: person, space and time across languages*, volume 98. Cambridge University Press.
- Holler, Judith and Katie Wilkin. 2011. An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics*, 43(14):3522–3536.
- Hostetter, Autumn B, Martha W Alibali, and Sotaro Kita. 2007. I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes*, 22(3):313–336.
- Jezek, Elisabetta. 2003. *Classi di verbi italiani tra semantica e sintassi*. Bulzoni.
- Jezek, Elisabetta, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS: A resource of corpus-derived typed predicate argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 890–895, Reykjavik, Iceland, May. European Language Resources Association (ELRA).
- Kendon, Adam. 1972. Some relationships between body motion and speech. *Studies in dyadic communication*, 7(177):90.
- Kendon, Adam. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.
- Kendon, Adam, 2011. *Gesticulation and Speech: Two Aspects of the Process of Utterance*, pages 207–228. De Gruyter Mouton.
- Kipp, Michael. 2001. Anvil - A generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*, pages 1367–1370, Aalborg, Denmark, September.
- Knight, Dawn. 2011. *Multimodality and active listenership: A corpus approach*. A&C Black.
- Krauss, Robert M. 1998. Why do we gesture when we speak? *Current directions in psychological science*, 7(2):54–54.
- Krauss, Robert M. and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. *Gesture, speech, and sign*, 93.
- Krippendorff, Klaus. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Lenci, Alessandro, Gabriella Lapesa, and Giulia Bonansinga. 2012. Lexit: A computational resource on italian argument structure. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC2012)*, pages 3712–3718, Istanbul, Turkey, May.
- Lin, Yen-Liang. 2017. Co-occurrence of speech and gestures: A multimodal corpus linguistic approach to intercultural interaction. *Journal of Pragmatics*, 117:155–167.

- Lucisano, Pietro and Maria Emanuela Piemontese. 1988. Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31):110–124.
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.
- McNeill, David. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- McNeill, David. 2005. *Gesture and thought*. University of Chicago Press.
- McNeill, David. 2008. *Gesture and thought*. University of Chicago press.
- McNeill, David. 2016. *Why we gesture: The surprising role of hand movements in communication*. Cambridge University Press.
- Menini, Stefano, Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2020. Dadoeval@ evalita 2020: Same-genre and cross-genre dating of historical documents. In *7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. EVALITA 2020*, pages 391–397, Online, December. Accademia University Press.
- Moneglia, Massimo, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014. The IMAGACT visual ontology. An extendable multilingual infrastructure for the representation of lexical encoding of action. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3425–3432, Reykjavik, Iceland, May.
- Morsella, Ezequiel and Robert M. Krauss. 2004. The role of gestures in spatial working memory and speech. *The American journal of psychology*, pages 411–424.
- O’Keefe, Anne and Michael McCarthy. 2010. *The Routledge handbook of corpus linguistics*. Routledge.
- Ovendale, Alice. 2012. *The Role of Gesture in Cross-cultural and Cross-linguistic Learning Contexts: The Effect of Gesture on the Learning of Mathematics*. Ph.D. thesis, University of Johannesburg.
- Parrill, Fey, Jennifer Bullen, and Huston Hoburg. 2010. Effects of input modality on speech–gesture integration. *Journal of Pragmatics*, 42(11):3130–3137.
- Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302, Mysore, India, January.
- Poggi, Isabella. 2007. *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler.
- Seiter, John S. and Weger Harry Jr. 2020. *Nonverbal communication in political debates*. Lexington Books.
- Stam, Gale and Steven G. McCafferty. 2008. Gesture studies and second language acquisition: A review. *Gesture: second language acquisition and classroom research*, pages 3–24.
- Tonelli, Sara, Rachele Sprugnoli, and Giovanni Moretti. 2019. Prendo la parola in questo sesso mondiale: A multi-genre 20th century corpus in the political domain. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLIC-it 2019)*, Bari, Italy, November.
- Tonelli, Sara, Ke Tran Manh, and Emanuele Pianta. 2012. Making readability indices readable. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 40–48, Montréal, Canada, June. Association for Computational Linguistics.
- Trotta, Daniela, Alessio Palmero Aprosio, Sara Tonelli, and Annibale Elia. 2020. Adding gesture, posture and facial displays to the PoliModal corpus of political interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4320–4326, Marseille, France, May. European Language Resources Association.
- Trotta, Daniela, Sara Tonelli, Alessio Palmero Aprosio, and Elia Annibale. 2019. Annotation and analysis of the polimodal corpus of political interviews. In *Sixth Italian Conference on Computational Linguistics (CLIC-it 2019)*, Bari, November.
- Tsui, Amy B.M. 1994. *English conversation*. Oxford University Press.
- Tuite, Kevin. 1993. The production of gesture. *Semiotica*, 93(1-2):83–105.
- Vignozzi, Gianmarco. 2019. How gestures contribute to the meanings of idiomatic expressions and phrasal verbs in tv broadcast interviews: A multimodal analysis. *Lingue e Linguaggi*, 29.
- Voghera, Miriam, 2001. *Teorie linguistiche e dati di parlato*, pages 75–96. Bulzoni, Roma.
- Voghera, Miriam. 2020. What we learn about language from spoken corpus linguistics? *Caplletra. Revista Internacional de Filologia*, (69):125–154.
- Wagner, Petra, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232.

Yoshioka, K. 2008. Linguistic and gestural introduction of inanimate referents in l1 and l2 narrative. *ESL & applied linguistics professional series*, pages 211–230.

Toward Data-Driven Collaborative Dialogue Systems: The JILDA Dataset

Irene Sucameli*
Università di Pisa

Alessandro Lenci**
Università di Pisa

Bernardo Magnini†
Fondazione Bruno Kessler

Manuela Speranza‡
Fondazione Bruno Kessler

Maria Simi§
Università di Pisa

Today's goal-oriented dialogue systems are designed to operate in restricted domains and with the implicit assumption that the user goals fit the domain ontology of the system. Under these assumptions dialogues exhibit only limited collaborative phenomena. However, this is not necessarily true in more complex scenarios, where user and system need to collaborate to align their knowledge of the domain in order to improve the conversation and achieve their goals.

To foster research on data-driven collaborative dialogues, in this paper we present JILDA, a fully annotated dataset of chat-based, mixed-initiative Italian dialogues related to the job-offer domain. As far as we know, JILDA is the first dialogic corpus completely annotated in this domain. The analysis realised on top of the semantic annotations clearly shows the naturalness and greater complexity of JILDA's dialogues. In fact, the new dataset offers a large number of examples of pragmatic phenomena, such as proactivity (i.e., providing information not explicitly requested) and grounding, which are rarely investigated in AI conversational agents based on neural architectures. In conclusion, the annotated JILDA corpus, given its innovative characteristics, represents a new challenge for conversational agents and an important resource for tackling more complex scenarios, thus advancing the state of the art in this field.

1. Introduction

In recent years, mostly driven by the high performance achieved by deep learning approaches in Natural Language Processing, there has been a resurgence of interest for systems that are able to assist people in a number of tasks, interacting in a natural way. However, reproducing the peculiarity and complexity of human-human dialogues

* Department of Computer Science - Largo Bruno Pontecorvo 3, 56127 Pisa, Italy. E-mail: irene.sucameli@phd.unipi.it

** Department of Philology, Literature and Linguistics - Via Santa Maria, 56126 Pisa, Italy. E-mail: alessandro.lenci@unipi.it

† Natural Language Processing Group - Via Sommarive 18 Povo, 38123 Trento, Italy. E-mail: magnini@fbk.eu

‡ Natural Language Processing Group - Via Sommarive 18 Povo, 38123 Trento, Italy. E-mail: manspera@fbk.eu

§ Department of Computer Science - Largo Bruno Pontecorvo 3, 56127 Pisa, Italy. E-mail: simi@di.unipi.it

poses a number of scientific challenges to current conversational AI approaches, and, more generally, to computational linguistics. In this paper we present JILDA, a corpus of human-human dialogues collected with the purpose of investigating linguistic variability and collaborative phenomena in *goal-oriented dialogues*, which imply a collaborative effort to plan actions among the interlocutors in order to achieve a certain communicative goal.

9. Applicant: *Nel frattempo potrei specificarti le mie preferenze a livello geografico? Potrebbero aiutarti nel targetizzarmi meglio*

10. Navigator: *SÃ, perfetto! Grazie*

11. Applicant: *Attualmente vivo in Toscana: sono disponibile a trasferirmi in altre regioni ma anche all'Ãestero non ho problemi di mobilitÃ o limiti da questo punto di vista*

12. Navigator: *Potrei avere due offerte che mi piacerebbe proporti. Entrambe riguardano tirocini post-laurea, uno come assistente capocommessa in una azienda edile a Pistoia, e l'altra come allievo direttore a Milano presso Compass.*

(...)

15. Applicant: *Non riesco a capire bene che cosa significhi "allievo direttore"*

16. Navigator: *Certo! Le principali mansioni legate a questo impiego riguardano la pianificazione del budget e del conto economico dell'azienda. Il settore Ã quello alimentare quindi si tratta di compilare ordini e derrate alimentari, oltre che garantire la sicurezza sul lavoro e quella alimentare.*

17. Navigator: *Compiti gestionali sarebbero sicuramente al centro del lavoro.*

18. Navigator: *Ti sembra piÃ chiaro? Posso dirti altro?*

19. Applicant: *Capisco. Mi sembra interessante*

(...)

21. Navigator: *Trattandosi di un tirocinio post-laurea direi che la formazione sarÃ una componente importante.*

22. Applicant: *Capisco. C'Ã una deadline per fare domanda?*

(...)

28. Applicant: *Capisco. Potresti darmi il contatto dell'azienda? In modo tale da approfondire e mettermi in contatto diretto con loro*

9. Applicant: *In the meantime, should I specify my geographic preferences? They could help you target me better*

10. Navigator: *Yes, perfect! Thank you*

11. Applicant: *At the moment I live in Tuscany: I'm available to move to other regions and even abroad I don't have mobility problems or limitations from this point of view*

12. Navigator: *I may have two offers that I would like to propose to you. Both involve post-graduate internships, one as an assistant prime contractor in a construction company in Pistoia, and the other as a junior director in Milan at Compass.*

(...)

15. Applicant: *I can't quite understand what "junior director" means*

16. Navigator: *Sure! The main tasks related to this job concern the planning of the budget and the income statement of the company. The area is the food sector so it's a question of filling orders and foodstuffs, as well as guaranteeing work and food safety.*

17. Navigator: *Management tasks would certainly be the core of the work.*

18. Navigator: *Is it more clear now? Can I tell you more?*

19. Applicant: *I see. It seems interesting*

(...)

21. Navigator: *Since this is a post-graduate internship I would say that training will be an important component.*

22. Applicant: *I see. Is there an application deadline?*

(...)

28. Applicant: *I see. Could you give me the company's contact? This way I can take a closer look and contact them directly*

Goal-oriented dialogues contain interactions governed by shared conventions (see, for instance the work of (Grice 1975) on conversational maxims), which involve knowledge about the *pragmatics* of language (Levinson 1983), i.e., the context in which they

are produced and the speakers' communicative intentions. In this paper we focus on two pragmatic phenomena that are relevant in goal-oriented dialogues: *proactivity* and *grounding*. To give an intuition of what proactivity and grounding are, and how they are pervasive in human dialogues, let's consider the following extract, from a goal-oriented dialogue from the JILDA corpus (full version available in Appendix), where a navigator and an applicant have to find a satisfactory match between a set of job offers and the applicant's CV.

Proactivity (Balaraman and Magnini 2020b) occurs when an interlocutor offers information which was not explicitly requested, with the intention of facilitating the achievement of the conversational goal. As an example, at lines 9 and 11 of the dialogue, the applicant offers information which was not asked by the navigator (i.e., its geographical working preferences), but is assumed to facilitate the search of an appropriate job offer. The navigator, too, at line 16 provides details about a company which were actually not required by the applicant question at line 15. Even in this case the purpose is facilitating the match of a job offer with the applicant's requirements.

Grounding (Clark and Schaefer 1987; Clark and Brennan 1991; Hough and Schlangen 2017) is the process through which participants in a dialogue build and keep themselves aligned to a common knowledge ground, formed by interlocutors' shared information. Depending on the state of the dialogue, it is possible to identify several types of grounding (Traum 1999; Hough et al. 2015), such as, for instance, *feedback* and *repair*, which allow participants to demonstrate their understanding of the conversation or to correct potential misunderstandings.

Grounding is particularly relevant in goal-oriented dialogue (Mushin et al. 2003), where the participants are not supposed to share part of their knowledge. In our example dialogue from the JILDA corpus, grounding occurs in several forms. At line 15 it is the applicant who poses a clarification question *I can't quite understand what "junior director" means*. At line 18 the navigator asks for confirmation *is it more clear now? Can I tell you more?*, while at lines 19, 22 and 28 the applicant explicitly recognises to be aligned with the navigator.

Although grounding and proactivity are pervasive in human-human dialogue, both are largely under represented in current data-driven, goal-oriented, dialogue systems. This is related to the fact that both phenomena are scarcely present in training data, which, in turn, may depend on the design choices adopted by developers for the collection of dialogues. Two design choices seem to be relevant: (i) some acquisition methodologies (e.g., Wizard of Oz) constrain participants in the data collection to follow pre-defined dialogue scripts, resulting in dialogues that are quite repetitive and poor in natural pragmatic phenomena; (ii) in most cases the domain of conversation is oversimplified with respect to the real world (e.g., when booking restaurants, they are described with few characteristics), resulting in a reduced need for grounding between the system and the user.

JILDA consists of goal-oriented, chat based, Italian dialogues related to the job-offer domain. The corpus is fully annotated with semantic information, such as dialogue acts and entities, as well as proactive phenomena. It is important to underline that the annotation of proactivity has been included in the dataset to better capture the complexity of a natural, human-human dialogue. This annotation therefore represents an important characteristic of the dataset itself and is useful for conducting a linguistic analysis of the Italian language, but it is not designed to develop a system capable of producing proactive behaviour.

We describe in detail the annotation methodology adopted in JILDA and analyse and discuss the major novelties introduced in the corpus, showing high presence of

pragmatic phenomena, including grounding and proactivity. We expect that JILDA can be used to train neural dialogue models for the Italian language (JILDA is a quite new resource for this language), thereby pushing the scientific community toward more natural and effective conversational systems.

2. Background on Goal-oriented Dialogue

In this section we introduce relevant background on goal-oriented dialogues, which may help to appreciate the novelty of the JILDA corpus. First we highlight some of the characteristics of goal-oriented dialogues, then we briefly introduce some notion relevant to the realisation of automatic goal-oriented dialogue systems, and, finally, we focus on the presence of collaborative behaviours in some datasets developed to train conversational systems.

2.1 Human-human Goal-oriented Dialogue

The purpose of a typical task-oriented dialogue is to retrieve pieces of information that are supposed to correspond to user needs (e.g., booking a restaurant, finding how to open a bank account, check the weather tomorrow, etc.). It is usually assumed that the user has a rather clear goal in mind, which is then elicited by an operator during the dialogue. The operator in fact may ask questions to the user attempting to reduce the search space and to focus on those objects that fit the user goals. On the other side, the user may also intervene in the dialogue to clarify and refine the goals of the conversation. Once objects that satisfy the user needs are retrieved, an action can be executed, such as booking a restaurant, or blocking a credit card. A goal-oriented dialogue may terminate either when the goal has been achieved (e.g., a reservation has been confirmed), or when the goal can not be achieved, because it was not possible to find a match with the user needs.

As an example of human-human goal-oriented dialogue, let's consider the following excerpt from Nespole (Mana et al. 2004, 2003), a corpus consisting of spoken interactions between a professional agent and a client about vacation planning in the Trentino region.

1. Client: *Good morning; could you suggest any village in the Val di Fiemme to me; where it's possible to skate for example; that is does any skating rink exist in the Val di Fiemme;*

2. Agent: *yes; in the whole of Val di Fiemme there are some outdoor skating rinks; where you can skate usually in the afternoon; in some rinks even in the morning; and then right in Cavalese there's a skating rink an ice rink; where even some courses are organized; where they also hold hockey or skating shows; and it's indoors.*

What is interesting for our purposes is the collaborative attitude of both the Client and the Agent. Particularly, the travel agent proactively provides indications both about the opening time of skating rinks and about skating courses, which were not explicitly requested by the customer. Proactivity is a peculiar characteristics of human-human dialogues, through which the Agent anticipates the expected requests of the user, this way facilitating the achievements of the dialogue goals.

2.2 Goal-oriented Dialogue Systems

Task-oriented dialogue systems aim to assist users to accomplish a task (e.g., booking a flight, making a restaurant reservation and playing a song) through dialogue in natural language, either in a spoken or written form. As in most current approaches, we assume a system involving a pipeline of components - see Figure 1, from (Deriu et al. 2021) - where the user utterance is first processed by an Automatic Speech Recognition (ASR) module and then processed by a Natural Language Understanding (NLU) component, which interprets the user’s needs (Louvan and Magnini 2020). Then a Dialogue State Tracker (DST) (Balaraman, Sheikhalishahi, and Magnini 2021) accumulates the dialogue information as the conversation progresses and may query a domain knowledge base to obtain relevant data. A dialogue policy manager then decides the next action to be executed and, finally, a Natural Language Generation (NLG) component produces the actual response to the user.

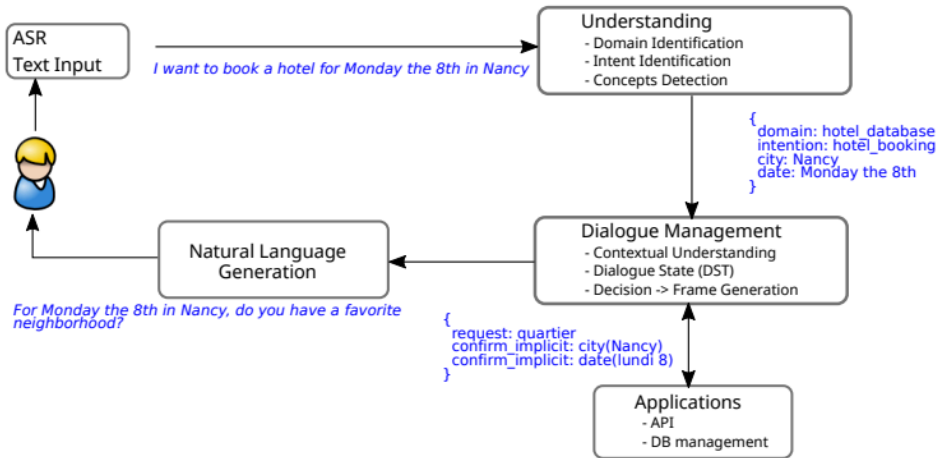


Figure 1
A standard architecture of a task-oriented dialogue system

In order to reproduce collaborative behaviours, the most relevant component is the dialogue manager, which has to decide whether a collaborative action is appropriate for the current dialogue turn, given the dialogue history and the user beliefs (i.e., the supposed user goals). For a dialogue manager the question is how to learn proactive behaviours, including knowledge about turns in which the system should be proactive, and when it should not, how to determine the information that should be proactively offered to the user, and the appropriate amount of such information (e.g., offering too much information may result in an excessive cognitive effort for the user). Similar questions apply to grounding, where the dialogue manager has to constantly monitor the level of grounding with the user, and, in case this is not satisfactory, has to take the initiative to restore it to an optimal level.

Given the inherent complexity of collaborative behaviours, it is not surprising that current dialogue systems still have limited capacities in this respect. The issue of reproducing collaborative behaviours is even more evident for a data-driven dialogue state tracker, which is assumed to learn dialogue behaviours from annotated dialogues.

In this case, the availability of dialogues displaying reach enough linguistic phenomena is crucial.

2.3 Datasets for Goal-oriented Dialogue

As dialogic annotated corpora are at the core of the capacity to learn dialogue models, this section introduces the most important available datasets, focusing on the presence of collaborative phenomena. As a case study, we have selected *WoZ* and *MultiWoZ*, two datasets developed in recent years, which are considered as benchmarks for developing deep learning methods for dialogue state tracking.

WoZ is a popular dataset for restaurant booking in Cambridge, collected using the Wizard of Oz approach, where the user and the wizard contribute a single turn to each dialogue (Wen et al. 2017). (Mrkšić et al. 2017) expanded WoZ into WoZ2.0, consisting of 1,200 dialogues. Then, MultiWOZ2.1 (Budzianowski et al. 2018) further extends WoZ including dialogues in multiple domains. To this aim, the dataset developers explicitly encouraged goal changes, in order to model more realistic conversations. Different versions of MultiWOZ2.1 have been recently published, addressing annotation errors occurring in the original dataset (Ramadan, Budzianowski, and Gasic 2018; Budzianowski et al. 2018; Eric et al. 2020; Zang et al. 2020). MultiWoZ2.1 contains 10,438 dialogues, covering several different domains (e.g., restaurants, hotels, trains and attractions).

Both datasets have been collected through the Wizard of Oz approach, (Kelley 1984), where a human (the “wizard”) plays the role of the computer within a simulated human-computer conversation, and, crucially the other speakers are not aware to talk to a human. The following is an example of a dialogue script provided to the “user” in the Wizard of Oz collection setting.

1. **User:** *You are looking for a <place to stay>. The hotel should be in the <cheap> price range.*
2. **User:** *The hotel should <include free parking> and should <include free wifi>*
3. **User:** *Once you find the <hotel> you want to book it for <6 people> and <3 nights> starting from <tuesday>*
4. **User:** *If the booking fails how about <2 nights>*
5. **User:** *Make sure you get the <reference number>*

The dialogue script is typically filled in using placeholders in a template (shown in *<italics>* in the example). It is worth to notice the amount of details that are present in the dialogue description, details that could influence the production of the user utterance for a given turn, and induce to follow a structure similar to that of the dialogue script. After being collected through Wizard of Oz, turns of each dialogue are annotated with the corresponding *dialogue state*, consisting of an intent and a set of slot-value pairs. The following is an example of the annotation provided in a portion of a MultiWoZ 2.0 dialogue:

1. **User:** *I would like a moderately priced restaurant in the west part of town.*
INFORM(PRICE=MODERATE, AREA WEST)
2. **System:** *here are three moderately priced restaurants in the west part of town. Do you prefer Indian Italian or British?*
REQUEST(FOOD)

3. User: *Can I have the address and phone number of the Italian location?*

INFORM(PRICE=MODERATE, AREA=WEST, FOOD=ITALIAN)
REQUEST(ADDRESS, PHONE-NUMBER)

Neither proactivity nor grounding are annotated in WoZ and MultiWoz. A recent study (Balaraman and Magnini 2020a), estimated that the amount of the system proactive behaviours in MultiWoz is rather low. In fact, out of 143,048 dialogue turns in the corpus, only 325 proactive turns were found with a clear proactive pattern. Although this might be an underestimation (as proactivity is not annotated in MultiWoz and it is not trivial to search for it), this is much less than we can reasonably expect in human-human goal-oriented dialogues, as the example reported in the introduction shows. Being poorly represented in the corpus, proactive behaviours can hardly be learnt by dialogue state tracking and dialogue policy models, motivating the need of richer dialogue annotations, such as those proposed in JILDA.

Other popular datasets used for dialogue state tracking include the schema-guided dataset (Shah et al. 2018), collected using a bootstrapping approach, and the TreeDST dataset (Cheng et al. 2020), with conversations covering 10 domains. These datasets mainly focus on the problem of managing a conversational domain with scarcity of training data (e.g., the problem of managing unseen slot values), proposing architectures (e.g., zero shot learning) that are robust enough for such situations. To the best of our knowledge, there is no much attention to explore collaborative phenomena in dialogue.

Finally, it is worth briefly reporting about the performance that state-of-the-art models achieve on the dialogue state tracking task. MultiWoz is probably the dataset mostly used to train a dialogue state tracker model, and several deep learning architectures have been experimented in the last years (Henderson, Thomson, and Young 2014; Balaraman and Magnini 2021), including methods proposed at various editions of the DST challenge (Henderson, Thomson, and Williams 2014). Performance are typically reported according to the *joint goal accuracy* of the model, i.e, the capacity of the model to correctly predict all dialogue states (slot-value pairs) in each turn of the dialogue. Current DST models, for instance TRADE (Wu et al. 2019), DST-QA (Zhou and Small 2019) and CHAN-DST (Shan et al. 2020), achieve a performance in the order of 50% of joint goal accuracy.

The JILDA dataset, which will be described in detail in the next sections, builds on top of the experience accumulated by MultiWoz, proposing, however, a number of methodological improvements. First of all JILDA has been collected through Map-task, a methodology that allows the participants to express themselves with more naturalness (i.e., rich language variability) than in the Wizard of Oz setting, this way overcoming some of the limitations of current datasets. Second, the selected domain, job offers, is more complex than the MultiWoz domains, which should favour grounding phenomena among interlocutors. Finally, although we basically follow the MultiWoz annotation schema, we have added categories specifically tailored to mark dialogue collaborative phenomena.

3. JILDA

JILDA is a dataset of chat-based dialogues, produced by 50 Italian native speakers and related to the job-offer domain. The dataset, which is available on GitHub,¹ includes 525 mixed-initiative dialogues collected from human-human conversations in an experiment inspired by the Map-task methodology, where one participant played the role of job consultant (or “navigator”) and the other the role of applicant, with the common goal of finding a good match between job offers and the applicant’s competences and expectations (Sucameli et al. 2020).

In a previous experiment we collected via Amazon Mechanical Turk another dataset of dialogues (Mturk), for the same domain and language as JILDA, using a template-based approach. Table 1 summarises the main characteristics of JILDA, highlighting the differences between this dataset with respect to the Mturk dataset.

Table 1

Comparison between MTurk’s and JILDA’s dialogues. Values marked with an asterisk are computed considering the average value of three JILDA’s subsets, each including the same number of tokens as MTurk

	MTurk	JILDA
# dialogues	220	525
avg turns per dialogue	8	17
# tokens	45972	217132
# sentences	5201	20644
# utterances	3380	14509
# types	1975	6519
# lemmas	1605	4913
type/token ratio	0.043	0.072*
lemma/token ratio	0.035	0.056*
avg length sentences	9.24	10.52
avg length utterances	13.58	14.94

As shown by Table 1, JILDA is characterised by a great linguistic variability and lexical complexity that we tried to capture effectively during the subsequent annotation phase.

3.1 Annotation Guidelines

The JILDA annotation scheme relies on the MultiWOZ 2.1 one (Budzianowski et al. 2018). Differently from MultiWOZ however, we annotate both applicant and navigator utterances. In fact, one of the main characteristics of JILDA is to include mixed-initiative dialogues, where both participants involved in the conversation may ask and answer questions, or volunteer information, thus conveying useful data worth extracting. In the following we will use the most standard terms “system” and “user” to refer to navigator and applicant. In fact, JILDA was created with the idea of training a dialogic system on this domain. In this scenario, the system would cover the role of navigator, while

¹ <https://github.com/IreneSucameli/JILDA>

the user would play the role of applicant. We annotate dialogue acts, which "represent the communicative intention behind a speaker's utterance in a conversation" [Chakravarty, Chava, and Fox 2019], and slots, which are specific to the JILDA job-offer domain.

3.1.1 Dialogue Acts

For our annotation we considered six *Dialogue acts*, and we annotated both user's and system's utterances. Each act describes a specific communicative intention of the speaker. More specifically, the dialogue acts used for the annotation are:

- **greet:** the speaker expresses a greeting. Example:
"Good morning, my name is Giulia and today I will be your navigator".
- **inform-basic:** the speaker provides information following a specific request. Example:
sys: "Tell me something about you: what type of studies have you done??"
usr: "I graduated from classical high school and then got a degree in nursing"
- **inform-proactive:** the speaker provides information that was not explicitly requested. For example, in the case below the system provides a piece of information (the email address) even if these data were not requested by the user:
"Could you tell me where the company is located??"
sys: "The company is in Milan. You can get in touch with them with the email address info@azienda.com"
- **request:** the speaker requests information:
sys: "Which sector would you like to work in?"
- **select:** a) the system selects the job offer suitable for the user's profile or b) the user accepts the job offer. Example:
sys: "Ok I found an offer that meets your interests: it is a post-graduate internship in the food sector."
- **deny:** the speaker is unable to satisfy a request. It includes, but is not limited to, categorizing cases in which the system does not find a suitable job offer for the user or the user does not accept the proposed offer.
Example:
usr: "I don't think this offer works for me."

Each sentence can be annotated with more than one dialogue act. For example, if the speaker, in addition to directly answering the interlocutor's question, volunteers additional information, the sentence is annotated with both *inform-basic* and *inform-proactive*. In the example proposed above to illustrate the dialogue act *inform-proactive*, sys provides the information directly requested by the user ("the company is located in Milan") as well as additional information (i.e. the company's email address).

3.1.2 Slots

A set of slots describes the relevant information we want to extract from dialogues in this specific domain. In our case each *slot* represents a specific attribute of the domain "job-offer". More specifically, we consider 14 domain-specific slots, described below:

- **age**: information referring to the age of the applicant or of the professional figure sought;
- **area**: sector of job position (e.g., “*I’d like to work in the advertising and communication area*”);
- **company-name**: name of the company or institution offering the job;
- **company-size**: company size based on the number of people who work there (e.g. “*I’d like to work in a big company*”);
- **contact**: contact information;
- **contract**: type of job contract offered or requested (e.g. “*part time*”);
- **degree**: degree or other qualification required or possessed by the applicant;
- **duties**: main tasks required by the job;
- **job-description**: title of the job position (e.g. “*web developer*”, “*receptionist*”);
- **languages**: knowledge of foreign languages required for the job or spoken by the user;
- **location**: location of the job or of the company;
- **past-experience**: user’s previous work experiences;
- **skill**: skills requested for the job or possessed by the applicant;
- **other**: all the extra information related to the job-offer domain and not fitting other slots.

```

"turn_id": 17,
"usr": "Perfect. Could you tell me the name of the company or an address I can contact to get all the information I need?",
(...)
"turn_id": 20,
"sys": "I don't have the website's link. But I can give you the email address",

"async": [
  [
    "turn_ref",
    "turn_17"
  ]
],

```

Figure 2

An example of annotation of asynchronous messages.

All the semantically informative text fragments in dialogic turns are annotated with the dialogue acts and slots names. In addition to the domain-specific slots, the annotation schema also includes two general slots. The first one, **Global slot**, is used to mark the overall results of the dialogue and it can assume only two values, *positive* or *negative*, according to the outcome of the job interview. The label *positive* is used to express success in finding a useful job position, while the label *negative* is used in case of failure. Therefore, respect to the other slots, the Global slot refers not to the single utterances but to the entire dialogue. The second one, **Async**, is used to mark the

presence of asynchronous messages, which naturally occur in chat conversations. We consider asynchronous those overlapping utterances where the answer to a question is not immediate but comes in a later turn. When this phenomenon occurs, we mark as *async* the message where the speaker replies to the question, entering as value of the slot the number of the dialogic turn where the question was asked, as in the example in Figure 2.

3.2 Annotating JILDA

The annotation task we proposed is complex since all slot fillers are open classes and the values correspond to substrings extracted from text. The selection of these values was left to annotators’ choices and therefore the boundaries of the selected text spans often differ, depending on the subjective choices made by the annotator.

JILDA and MTurk annotation process was supported by MATILDA, an open source tool specifically designed to annotate multi-turn dialogues, which was extended to support the management of collaborative annotation projects (Cucurnia et al. 2021). Each annotator is assigned subsets of the collection to annotate and can add/modify her own annotations without affecting the work of the others. The system takes care of persistence by storing in a database intermediate work of the annotators and offers management and monitoring capabilities to the project supervisor. The work of different annotators can be compared through a inter-annotator interface, which also supports the resolution of disagreements.

Annotating JILDA involved four annotators, who worked in pairs during two distinct annotation phases. Both JILDA and MTurk dialogues where annotated, thus building a dataset of over 750 fully annotated dialogues in the job search domain.



Figure 3
Dialogue annotation using MATILDA’s interface

Figure 3 shows an example of dialogue annotation via the MATILDA’s interface. Each dialogue, organised into dialogic turns, is shown in the middle of the interface screen. Each turn includes both system’s and user’s utterance. The panel on the left allows the annotator to select the relevant tags, filling the values of the slots through a text selection made directly from the input sentences. Besides the slot value, the position

in the sentence of the highlighted tokens is also stored. The annotated dialogues are then exported in json format, as shown in Figure 4.

```

"sys": "So, tell me, do you know any foreign languages?",
"sys_request": [
  [
    "languages",
    "sys[3,3][?],"
  ]
],
"turn_id": 5,
"usr": "Yes, I do know English (B2) and French (B2) and I have attended an AI German course.",
"usr_inform_basic": [
  [
    "languages",
    "usr[4,4][English],usr[5,5][B2],usr[7,7][French],usr[8,8][B2],usr[14,14][AI],usr[15,15][German],"
  ]
]

```

Figure 4

Output of the annotated dialogue, in json format

4. Analysis

4.1 First Annotation Phase

The first annotation phase involved two annotators: one worked on the entire JILDA dataset, while the other annotated the Mturk collection. When this annotation was completed, we conducted a first analysis targeting the number of tokens and types per slot, in order to understand the frequency of use of the slots, their lexical variability and for each slot the size of the linguistic dictionary that can be extracted from JILDA and Mturk.

Table 2

Tokens and types extracted per slot during the first annotation phase

	tokens	types	Type/token ratio
age	92	27	0.29
area	873	447	0.51
company-name	464	107	0.23
company-size	392	238	0.60
contact	512	49	0.09
contract	987	170	0.17
degree	863	459	0.53
duties	1206	852	0.70
job-description	660	275	0.41
languages	795	142	0.17
location	1200	257	0.21
other	106	93	0.87
past-experience	588	463	0.78
skill	1287	659	0.51
Total	10025	4238	0.42

As shown in Table 2, the *type / token ratio* of the slots' values annotated in JILDA and Mturk is 0.42 on the average. These data suggest that the two datasets have a significant semantic variability and seem to effectively capture the linguistic variety of native speakers. On the other hand, a low type/token ratio can create difficulties in training an effective linguistic model, particularly when there is the need to generalise among slot classes. To overcome this problem, without losing the linguistic richness which is typical of JILDA, we introduced specific modifications and additional indications during the second annotation phase, as described in the next section.

In addition to analysing the vocabularies of both datasets and slots, we computed the number of proactive phenomena annotated. This is an interesting analysis to conduct, since it constitutes a measure of the complexity and naturalness of the data collected.

In JILDA 17.15% of dialogue acts were proactive, while in the MTurk dataset only 1.98%. This difference between JILDA and Mturk is undoubtedly due to the different data collection methodology used to build the two datasets: a template-based approach in the case of MTurk and a less rigid approach based on the Map Task methodology in the case of JILDA.

4.2 Second Annotation Phase

At the end of the first annotation phase, we noticed some critical issues. First of all, dialogue acts and slots were not linked. This means that an utterance could be marked with one (or more) acts but could lack of slots' values and, vice versa, selected slot values could pertain to different speech acts. Consequently, it was not possible to identify a posteriori which part of the text had been marked with a specific dialogue act. Moreover, as said before, the use of open classes for the slots has led to the production of a large vocabulary for both datasets, a possibly critical issue if the data are to be used to train a dialogue model.

In order to improve the quality of the annotation and to ensure greater consistency with the Multiwoz schema, we introduced the following adjustments in the configuration model and annotation guidelines:

- One or more slots were directly associated with one of the annotated dialogue acts, in accordance with Multiwoz's annotation schema.
- We asked annotators to include in the slot's selection the smallest informative part of an utterance. In this way, sentences like *"I would like to work as web developer"* were reduced to *"web developer"*.
- To avoid losing relevant information, in case of short confirmation or denial in a speaker's utterance, the referent of this speech act was made explicit, annotating as slot's value the relevant part of the text that appeared in the previous utterance. For example, if the system says *"I find a job offer as a nurse"* and the user says *"Ok, fine"*, the latter utterance is marked as *usr-select* (as dialogue act) + *job-description* (slot) + *"nurse"* (slot value).
- To comply with the Multiwoz schema, a request is always targeted to a specific slot, and the slot value is *"?"*.

Table 3

Types extracted per slot during the second annotation phase

	tokens	types	Type/token ratio
age	130	36	0.27
area	1472	331	0.22
company-name	556	96	0.17
company-size	732	149	0.20
contact	827	44	0.05
contract	1486	131	0.08
degree	1243	315	0.25
duties	1741	956	0.54
job-description	1362	425	0.31
languages	1085	60	0.05
location	1922	168	0.08
other	559	184	0.32
past-experience	882	244	0.27
skill	1994	570	0.28
Total	15991	3709	0.23

Following these changes to the guidelines, a second annotation phase was then realised. The work involved two different annotators, who equally shared the annotation work of JILDA and Mturk. This second annotation was more accurate and led to the creation of a more detailed dataset. Furthermore, from the analysis conducted after the annotation, it seems that the changes in the revised guidelines have actually led to a reduction of the corpus vocabulary, without however losing the lexical richness of the annotated data. Indeed, Table 3 shows that the vocabularies of the two datasets are still large, although the type/token ratio, which is 0.23, is lower than before (the type/token ratio of the previous annotation was 0.42).

Moreover, the number of proactive elements is still significant, with an overall percentage of 10.4% and this is a clear indicator of the naturalness and richness of the JILDA dataset with respect to MTurk. In fact, 12.7% of the dialogue acts in JILDA are proactive, while in MTurk we observe only 2.6% of proactive acts, also due to the different features of the dialogues.

4.3 Interannotator Agreement

In order to evaluate the quality of the annotated data, we calculated the inter-annotator agreement (IAA). We decided to compute the agreements between the two annotation rounds since annotators of both rounds worked on the same datasets and they had the same task, although the guidelines changed as described in 4.2. We computed the agreement in three different steps.

Firstly, we considered if there was an overlap between the text selected as slot value by the first annotator (A1) and the second one (A2). Indeed, it was important to consider if both the annotators recognised as “informative” the same part of the utterance. We

decided to consider as an agreement also an approximated overlap. The example below shows two cases of accepted match, which is exact in the first example:

A1: ["usr-inform-proactive", "skill", "**bachelor's degree in engineering**"]
 A2: ["usr-inform-basic", "degree", "**bachelor's degree in engineering**"]

and approximated in the second one.

A1: ["usr-inform-proactive", "skill", "**bachelor's degree in engineering**"]
 A2: ["usr-inform-basic", "degree", "**degree in engineering**"]

From the 1725 strings identified by at least one of the annotators as informative, we identified **810** cases of agreement. By focusing on these overlapping values, we move on to consider whether the text fragments identified as informative were associated to the same slot by the annotators, as in the example:

A1: ["usr-inform-proactive", "**degree**", "**degree in engineering**"]
 A2: ["usr-inform-basic", "**degree**", "**degree in engineering**"]

Finally, when there is a match both on values and on slots, we evaluated if there is an agreement also in the dialogue act, as in the example:

A1: ["**usr-inform-basic**", "**degree**", "**degree in engineering**"]
 A2: ["**usr-inform-basic**", "**degree**", "**degree in engineering**"]

Using this approach, we computed three values for agreement: i.) the percentage of sub-string matches over the total number of selected values, ii.) the percentage of agreements in slot attribution over the total of matching sub-strings, and iii.) the percentage of agreements in dialogue-acts over the cases matching in both values and slots.

We computed the above agreement measures for JILDA and obtained the results shown in Table 4. We can observe that the agreement values are very low, as expected considering that changes made in the guidelines before the second round of annotation were substantial.

Table 4

IAA between first and second annotation on 10% of the dataset.

	Sub-strings	Slot	Dialogue acts
Cases	1725	810	714
Agreement	810	714	419
Accuracy	0.47%	0.88	0.58

To effectively evaluate the quality of the new annotation, we asked the two volunteers of the second phase to make a cross-annotation using a subset of JILDA, which corresponds to about 10% of the entire dataset. In this way we could evaluate if the workers had truly internalised the annotation scheme and had produced a consistent dataset. The new calculation of accuracy gives substantially higher values, as it can be seen from Table 5; this clearly proves that using the same guidelines annotators are able to create a consistent annotation of the dataset. In addition to the accuracy values, in this case we also computed Cohen's kappa both for dialogue acts and slots considering both the actual accuracy and the predicted accuracy. The results are extremely positive and are, respectively, 0.82 and 0.86. These values were computed on the basis of the the

confusion matrices between the two annotators reported in the Appendix. By looking at those matrices we can notice that, as slots are concerned, the two annotators often disagreed on the attribution to the slot *area* vs *degree* or *skill* vs *job-description*. In the attribution of slots to dialogue acts instead, most disagreements were associated, as expected, to the subtle distinction between *inform-basic* and *inform-proactive*.

Table 5

IAA between second and third annotation on 10% of the dataset.

	Sub-strings	Slots	Dialogue acts
Cases	1661	1230	1163
Agreement	1230	1163	911
Accuracy	0.73	0.87	0.84
Cohen's kappa	-	0.86	0.82

5. Grounding & Proactivity

The semantic annotations reported so far focused on slots related to the domain and to proactive dialogue acts. For what concerns the analysis of the proactivity in JILDA, we computed the number of labels used to mark information provided proactively by the speaker, as shown in Figure 5.

```

"sys": "Have you ever worked in public accounting?",
"sys_request": [
  [
    "past_experience",
    "sys[6,6][?],"
  ]
],
"turn_id": 8,
"usr": "No, but I can use Excel very well.",
"usr_deny": [
  [
    "past_experience",
    "usr[0,0][No],"
  ]
],
"usr_inform_proactive": [
  [
    "skill",
    "usr[5,5][Excel],"
  ]
]

```

Figure 5

Example of information provided proactively by the speaker.

As can be observed from Table 6, the number of proactive sentences, is quite high in JILDA, which constitutes a clear indicator of the naturalness of the data collected.

Although dialogues were not annotated with grounding phenomena, as exemplified in the introduction, we expect the JILDA dataset to include a substantial amount

Table 6
Number of proactive acts labelled in JILDA and MTurk.

	JILDA	Mturk
I annotation	2624	76
II annotation	1712	102
I ann % of proact. data	17.16%	1.98%
II ann % of proact. data	12.7%	2.6%

of instances of grounding for the fact that dialogues are natural and representative of unconstrained and cooperative human-to-human dialogues. In order to substantiate this claim with a quantitative analysis we can look at the presence of several patterns commonly associated with grounding expressions specific to this domain: expressions of confirmation, of misunderstanding and confusion, or requests for explanations.

Table 7
Grounding expressions in JILDA.

Pattern	Instances
<i>capisco, capire, capito</i>	284
<i>ok</i>	465
<i>certo</i>	402
<i>certamente</i>	188
<i>chiaro, chiarire</i>	15
<i>d'accordo</i>	115

Table 7 reports the number of instances associated to the corresponding patterns. This analysis is limited by the fact that manifestations of grounding expressed through questions are often hard to be distinguished from normal discovery questions about unknown features of the job offer or of the applicant profile.

Table 8
Grounding acts according to Traum (1999). DU stands for *Dialogue Units*.

Label	Description
initiate	Begin new DU, content separate from previous uncompleted DUs
continue	some agent adds related content to open DU
acknowledge	Demonstrate or claim understanding of previous material by other agent
repair	Correct (potential) misunderstanding of DU content
Request Repair	Signal lack of understanding
Request Ack	Signal for other to acknowledge
cancel	Stop work on DU, leaving it ungrounded and ungroundable

To give an idea of the progress of the grounding contribution within a dialogue, we have represented a portion of the JILDA dialogue presented in the Appendix as a state transition diagram, based on the model proposed in (Traum and Nakatani 2002). Using the grounding scheme proposed by Traum (see Table 8), the respective grounding acts have been identified for the first 16 turns of the dialogue, as shown in Table 9. It can be noted how *continue* and *acknowledge* constitute the core of the grounding behaviour. Particularly the applicant introduces new information (e.g., T9. *...should I specify my geographical preferences?*) only after the navigator has acknowledged, implicitly, the previous turn (T7. *let's see immediately among the offers available what could fit best for you*).

Table 9
Grounding diagram for a portion of a JILDA dialogue.

Dialog. Turns	<i>initiate</i>	<i>continue</i>	<i>acknowledge</i>	<i>repair</i>	<i>Req. Repair</i>
T1	x				
T2			x		
T3			x		
T4		x			
T5		x			
T6		x			
T7			x		
T8			x		
T9		x			
T10			x		
T11			x		
T12			x		
T13		x			
T14			x		
T15					x
T16				x	

6. Conclusion and Future Work

We have presented JILDA, a corpus of annotated human-human goal-oriented dialogues related to the job-offer domain. Differently from other datasets, JILDA has been collected through map-task, a method allowing to acquire natural dialogues. As a result, JILDA dialogues exhibit both high linguistic variability and high presence of collaborative phenomena. Annotations take as a basis the MultiWOZ scheme but, differently from the latter, we annotate both user and system utterances, highlighting the dialogue acts describing the aim of the utterance, as well as slots specific to the JILDA job-offer domain. We presented a detailed analysis of the JILDA semantic annotations, showing that the new dataset contains a large amount of pragmatic phenomena, such as *proactivity* (i.e., providing information not explicitly requested) and *grounding*, which are both rarely investigated in current AI conversational agents based on neural architectures.

Given its innovative characteristics, JILDA has the potential to foster research in conversational AI toward really collaborative goal-oriented systems. To this end, we

intend to use JILDA to experiment neural dialogue state tracking and dialogue policy models able to reproduce both grounding and proactive interactions.

References

- Balaraman, Vevake and Bernardo Magnini. 2020a. Investigating proactivity in task-oriented dialogues. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, March 1-3. CEUR-WS.org.
- Balaraman, Vevake and Bernardo Magnini. 2020b. Proactive systems and influenceable users: Simulating proactivity in task-oriented dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2020)*, Virtually at Brandeis, Waltham MA, USA, July.
- Balaraman, Vevake and Bernardo Magnini. 2021. Domain-aware dialogue state tracker for multi-domain dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:866–873.
- Balaraman, Vevake, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 2021. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralakha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li, editors, *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2021)*, pages 239–251, Singapore and Online, July 29-31. Association for Computational Linguistics.
- Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Chakravarty, Saurabh, Raja Venkata Satya Phanindra Chava, and Edward A. Fox. 2019. Dialog acts classification for question-answer corpora. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 17th International Conference on Artificial Intelligence and Law (ASAIL@ICAIL)*, Montreal, QC, Canada, June.
- Cheng, Jianpeng, Devang Agrawal, Hector Martinez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, et al. 2020. Conversational semantic parsing for dialog state tracking. *arXiv preprint arXiv:2010.12770*.
- Clark, H. Herbert and Susan E. Brennan. 1991. Grounding in communication. In L.B. Resnick, J.M. Levine, and S.D. Teasley, editors, *Perspectives on Socially Shared Cognition*. American Psychological Association, pages 127–149.
- Clark, H. Herbert and F. Edward Schaefer. 1987. Collaborating on contributions to conversations. *Language Cognition and Neuroscience*, pages 19–41.
- Cucurnia, Davide, Nikolai Rozanov, Irene Sucameli, Augusto Ciuffoletti, and Maria Simi. 2021. Multi-annotator multi-language interactive light-weight dialogue annotator. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 32 – 39, address=Online, April 19 – 23 2021.
- Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. In *Artificial Intelligence Review*, 54, 755–810.
- Eric, Mihail, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 422–428, Marseille, France, May. European Language Resources Association.
- Grice, Herbert P. 1975. Logic and conversation. In *Speech acts*. Brill, pages 41–58.
- Henderson, Matthew, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, USA, 18-20 June.
- Henderson, Matthew, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, Philadelphia, PA, USA, 18-20 June.
- Hough, Julian, Iwan Kok, David Schlangen, and Stefan Kopp. 2015. Timing and grounding in motor skill coaching interaction: Consequences for the information state. In *Proceedings of the*

- 19th Workshop on the Semantics and Pragmatics of Dialogue (*SemDial 2015*), pages 86 – 94, Gothenburg, Sweden, August.
- Hough, Julian and David Schlangen. 2017. It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017)*, pages 274–282, Vienna, Austria, March.
- Kelley, J. F. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41.
- Levinson, Stephen C. 1983. *Pragmatics*. Cambridge University Press, Cambridge, U.K.
- Louvan, Samuel and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.
- Mana, Nadia, Susanne Burger, Ronaldo Cattoni, Laurent Besacier, Victoria MacLaren, John McDonough, and Florian Metze. 2003. The NESPOLE! voIP multilingual corpora in tourism and medical domains. In *Proceedings of the 8th European Conference on Speech Communication and Technology (INTERSPEECH 2003)*, Geneva, Switzerland, September.
- Mana, Nadia, Roldano Cattoni, Emanuele Pianta, Franca Rossi, Fabio Pianesi, and Susanne Burger. 2004. The Italian NESPOLE! corpus: a multilingual database with interlingua annotation in tourism and medical domains. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).
- Mrkšić, Nikola, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada, July. Association for Computational Linguistics.
- Mushin, Ilana, Lesley Stirling, Janet Fletcher, and Roger Wales. 2003. Discourse structure, grounding, and prosody in task-oriented dialogue. *DISCOURSE PROCESSES*, 35:1–31, 01.
- Ramadan, Osman, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 432–437, Melbourne, Australia July 2018.
- Shah, Pararth, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana, June. Association for Computational Linguistics.
- Shan, Yong, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou. 2020. A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6322–6333, Online, July. Association for Computational Linguistics.
- Sucameli, Irene, Alessandro Lenci, Bernardo Magnini, Maria Simi, and Manuela Speranza. 2020. Becoming JILDA. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLIC-it 2020*, pages 409–414, Bologna, Italy (Online), March 1-3, 2021. CEUR-WS.org.
- Traum, David R. 1999. Computational models of grounding in collaborative systems.
- Traum, David R. and Christine H. Nakatani. 2002. A two-level approach to coding dialogue for discourse structure: Activities of the 1998 DRI working group on higher-level structures. In *Towards Standards and Tools for Discourse Tagging*.
- Wen, Tsung-Hsien, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.
- Wu, Chien-Sheng, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy, July. Association for Computational Linguistics.
- Zang, Xiaoxue, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for*

Conversational AI, ACL 2020, pages 109–117, Online, July.
Zhou, Li and Kevin Small. 2019. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *ArXiv*, abs/1911.06192.

Appendix

Sample of JILDA dialogues

An illustrative example of goal-oriented dialogue from JILDA.

1. **Navigator:** *Ciao! Sono Chiara e sarÃš il tuo navigator. Come posso aiutarti?.*
2. **Applicant:** *Ciao Chiara. Mi chiamo Marta e sono alla ricerca di un lavoro*
3. **Navigator:** *Ciao Marta, spero di poterti aiutare. Mentre cerco fra le offerte disponibili, puoi dirmi qualcosa sul tuo lavoro ideale?*
4. **Applicant:** *Mi sono laureata da pochi mesi in Legge*
5. **Applicant:** *Mi piacerebbe poter imparare dai professionisti del mio ambito, quindi poter essere assunta, anche per un tirocinio, in uno studio di avvocatura o notarile sarebbe per me il massimo*
6. **Applicant:** *Mi rendo conto che come ambiente sia sovraccaricato di offerta e che i posti aperti sono molto pochi perÃšÃ¶*
7. **Navigator:** *Mai perdere la speranza Marta e io sono qui proprio per aiutarti a cercare l'offerta migliore per te! Vediamo subito fra le offerte disponibile cosa potrebbe fare al caso tuo..*
8. **Applicant:** *Grazie, sei molto gentile*
9. **Applicant:** *Nel frattempo potrei specificarti le mie preferenze a livello geografico? Potrebbero aiutarti nel targetizzarmi meglio*
10. **Navigator:** *SÃ, perfetto! Grazie*
11. **Applicant:** *Attualmente vivo in Toscana: sono disponibile a trasferirmi in altre regioni ma anche all'Ãžestero non ho problemi di mobilitÃ o limiti da questo punto di vista*
12. **Navigator:** *Potrei avere due offerte che mi piacerebbe proporti. Entrambe riguardano tirocini post-laurea, uno come assistente capocommessa in una azienda edile a Pistoia, e l'altra come allievo direttore a Milano presso Compass. Ho pensato a te perchÃ entrambi gli impieghi riguardano incarichi gestionali e la tua laurea potrebbe essere molto utile in questi contesti.*
13. **Navigator:** *Uno dei due ti sembra piÃ interessante e vuoi che te lo descriva per primo?*
14. **Applicant:** *Devo essere sincera: il primo non penso che possa fare al caso mio. Potresti descrivermi il secondo lavoro?*
15. **Applicant:** *Non riesco a capire bene che cosa significhi "allievo direttore"*
1. **Navigator:** *Hello! I'm Chiara and I'll be your navigator. How can I help you?.*
2. **Applicant:** *Hi Chiara. My name's Marta and I'm looking for a job*
3. **Navigator:** *Hi Marta, I hope I can help you. While I search for available offers, can you tell me something about your dream job?*
4. **Applicant:** *I graduated in Law few months ago.*
5. **Applicant:** *I'd like to learn from experts in my area and be hired, even for an internship, in a law firm or notary's would be great for me.*
6. **Applicant:** *I realize that this sector is overloaded with requests and that there are very few places open, butÃ¶*
7. **Navigator:** *Never give up hope, Marta, I'm here to help you find the best offer available for you. Let's see immediately among the offers available what could fit best for you ..*
8. **Applicant:** *Thanks, you're very kind*
9. **Applicant:** *In the meantime, should I specify my geographic preferences? They could help you target me better*
10. **Navigator:** *Yes, perfect! Thank you*
11. **Applicant:** *At the moment I live in Tuscany: I'm available to move to other regions and even abroad I don't have mobility problems or limitations from this point of view*
12. **Navigator:** *I may have two offers that I would like to propose to you. Both involve post-graduate internships, one as an assistant prime contractor in a construction company in Pistoia, and the other as a junior director in Milan at Compass. I think they could be ok for you because both jobs involve management assignments and your degree could be very useful in these contexts.*
13. **Navigator:** *Does one of them seem more appealing? And do you want me to describe it first?*
14. **Applicant:** *I have to be honest: I don't think the first one is right for me. Could you describe the second job for me?*
15. **Applicant:** *I can't quite understand what "junior director" means*

16. Navigator: Certo! Le principali mansioni legate a questo impiego riguardano la pianificazione del budget e del conto economico dell'azienda. Il settore Ã quello alimentare quindi si tratta di compilare ordini e derrate alimentari, oltre che garantire la sicurezza sul lavoro e quella alimentare.

17. Navigator: Compiti gestionali sarebbero sicuramente al centro del lavoro.

18. Navigator: Ti sembra piÃ chiaro? Posso dirti altro?

19. Applicant: Capisco. Mi sembra interessante

20. Applicant: Sai se lâazienda offre formazione? A chi per esempio come me non ha un grande background economico ma ha fatto esami all'UniversitÃ di Economia

21. Navigator: Trattandosi di un tirocinio post-laurea direi che la formazione sarÃ una componente importante.

22. Applicant: Capisco. C'Ã una deadline per fare domanda?

23. Applicant: Sto considerando anche altre posizioni aperte in parallelo

24. Navigator: No, tranquilla, puoi inviare domanda quando preferisci. Per ora la posizione Ã aperta.

25. Applicant: Sai quanti candidati prendono? Per prepararmi al meglio alla selezione

26. Navigator: Per ora ho solo questa offerta pubblicata da questa azienda. Penso che si tratti di un solo posto aperto.

27. Navigator: E non si indica il numero di domande accettate.

28. Applicant: Capisco. Potresti darmi il contatto dell'azienda? In modo tale da approfondire e mettermi in contatto diretto con loro

29. Navigator: Certo! Eccolo info@azienda.com

30. Applicant: Ti ringrazio per lâorientamento. Ti lascio il mio contatto, nel caso in cui saltino fuori posizioni aperte nel mio campo: martamarta@gmail.com

31. Navigator: perfetto! Se ci fossero novitÃ ti contatterÃ!

32. Navigator: Spero di esserti stata comunque utile.

33. Applicant: Molto, A presto e buona giornata

34. Navigator: Buona giornata anche a te!

16. Navigator: Sure! The main tasks related to this job concern the planning of the budget and the income statement of the company. The area is the food sector so it's a question of filling orders and foodstuffs, as well as guaranteeing work and food safety.

17. Navigator: Management tasks would certainly be the core of the work.

18. Navigator: Is it more clear now? Can I tell you more?

19. Applicant: I see. It seems interesting

20. Applicant: Do you know if the company offers training? For example, for those who, like me, don't have a background in economics but took some exams at the University in Economics

21. Navigator: Since this is a post-graduate internship I would say that training will be an important component.

22. Applicant: I see. Is there an application deadline?

23. Applicant: I'm considering other open positions in parallel

24. Navigator: No, don't worry, you can apply whenever you like. The position is open for now.

25. Applicant: Do you know how many candidates they accept? To better prepare myself for the selection

26. Navigator: For now I only have this offer published by this company. I think it's just one open position.

27. Navigator: And the number of applications accepted is not indicated.

28. Applicant: I see. Could you give me the company's contact? This way I can take a closer look and contact them directly

29. Navigator: Sure! It's info@azienda.com

30. Applicant: Thank you for the assistance. I'll give you my contact info, in case open positions arise in my field: martamarta@gmail.com

31. Navigator: perfect! If there is any news I will contact you!

32. Navigator: I hope I have been helpful, anyway.

33. Applicant: Very, see you soon and have a good day

34. Navigator: Good day to you too!

Confusion matrix on slots and dialogue acts

Confusion matrix between annotator A and annotator B on 10% of the JILDA dataset in classifying overlapping text spans into slots.

A \ B	age	area	comp name	comp size	contact	contract	degree	duties	job descr	lang	location	none	other	exp	skill	Sum
age	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
area	1	150	0	0	0	0	35	1	10	0	1	0	5	3	9	215
comp name	0	0	78	0	4	0	0	0	1	0	2	0	1	0	0	86
comp size	0	1	0	83	0	1	0	1	0	0	1	0	4	0	0	91
contact	0	0	3	0	104	0	0	0	0	0	0	0	1	0	0	108
contract	0	1	0	2	0	169	0	0	6	0	0	0	6	2	2	188
degree	0	19	0	0	0	0	202	0	1	1	0	0	3	1	1	228
duties	0	4	0	0	0	0	0	217	11	2	1	0	0	0	8	243
job_descr	0	14	1	0	1	8	0	12	124	0	1	0	2	1	1	165
languages	0	0	0	0	0	0	6	0	0	134	0	0	11	0	1	152
location	0	0	2	2	1	1	0	2	3	0	225	0	1	0	0	237
none	0	1	0	0	0	0	0	0	0	0	1	294	1	3	0	300
other	0	3	0	1	0	0	3	3	1	10	0	0	2	0	1	24
experience	0	6	0	0	0	1	1	2	1	0	0	0	3	68	6	88
skills	0	4	0	0	0	0	3	15	1	2	0	0	2	11	285	323
Sum	13	203	84	88	110	180	250	253	159	149	232	294	42	89	314	2460

Figure 6
Agreement on slots

Confusion matrix between annotator A and annotator B on 10% of the JILDA dataset in classifying slots into dialogue acts.

A \ B	sys deny	sys greet	sys inform basic	sys inform proactive	sys request	sys select	usr deny	usr greet	usr inform basic	usr inform proactive	usr request	usr select	Sum
sys deny	11	0	1	1	0	0	0	0	0	0	0	0	13
sys greet	0	134	0	0	0	0	0	0	0	0	0	0	134
sys inform basic	1	0	505	57	1	1	0	0	0	0	0	0	565
sys inform proactive	0	0	56	34	1	0	0	0	0	0	0	0	91
sys request	0	0	2	0	217	0	0	0	0	0	0	0	219
sys select	0	0	1	0	0	97	0	0	0	0	0	0	98
usr deny	0	0	0	0	0	0	21	0	0	5	0	0	26
usr greet	0	0	0	0	0	0	0	160	0	0	0	0	160
usr inform	0	0	0	0	0	0	1	0	376	105	2	0	484
usr inform proactive	0	0	0	0	0	0	0	0	87	129	0	0	216
usr request	0	0	0	0	0	0	0	0	0	2	126	0	128
usr select	0	0	0	0	0	0	0	0	1	0	0	12	13
Sum	12	134	565	92	219	98	22	160	464	241	128	12	2147

Figure 7
Agreement on dialogue acts

Analysis of Empathic Dialogue in Actual Doctor-Patient Calls and Implications for Design of Embodied Conversational Agents

Sana Salman*

Faculty of Science and Engineering –
Macquarie University, Sydney

Deborah Richards**

Faculty of Science and Engineering –
Macquarie University, Sydney

Patrina Caldwell†

University of Sydney, The Children’s
Hospital at Westmead

Embodied Conversational Agents (ECAs) are virtual agents that exhibit humanlike verbal and non-verbal behaviours. When it comes to eHealth, ECAs can provide vital support to patients by being more reachable. In order to make ECAs more effective, humanlike empathy expressed during conversation through relational cues is essential. Empathy revolves around a wide range of verbal and non-verbal behaviours that include, for example, the choice of words in social dialogues. Owing to the COVID-19 situation, there was an opportunity to record online consultations in the Incontinence Clinic and Sleep Clinic at the Children’s Hospital at Westmead in Sydney, Australia. The present study analysed these human dialogues using qualitative research methods to understand the role of empathic dialogue used by the medical team. The qualitative analysis of the live calls used psychology-based relational cues derived from conversational characteristics of humans to build a coding framework around the most relevant themes. Statistical analysis was used to compare relational cue usage between healthcare roles. Finally, using the framework dialogues of the medical team and two ECAs known as Dr Evie (eVirtual Agent for Incontinence and Enuresis) and SAM (Sleep Adherence Mentor) were compared to provide recommendations for health practitioners and future ECA dialogue development.

1. Introduction

Embodied Conversational Agents (ECAs) are virtual agents that exhibit humanlike verbal and non-verbal behaviours. They are increasingly being applied in contexts where the main mode of interaction is a dialogue between two or more humans (Bickmore, Gruber, and Picard 2005). In eHealth, ECAs can provide vital support to patients by being more reachable and available in their time of need (Richards and Caldwell 2017). The agents can not only act as a source of information on health issues (Lisetti et al.

* Faculty of Science and Engineering – Macquarie University, Sydney
E-mail: sana.salman@students.mq.edu.au

** Faculty of Science and Engineering – Macquarie University, Sydney.
E-mail: deborah.richards@mq.edu.au

† University of Sydney, The Children’s Hospital at Westmead
E-mail: patrina.caldwell@health.nsw.gov.au

2012) and their prevention or cure (Yin, Ring, and Bickmore 2012), but also motivate the patients to adhere to treatments (Bickmore et al. 2010); (Richards and Caldwell 2017). For teenagers and children, a virtual agent can act as an educator, buddy and motivator (Looije, Neerinx, and Lange 2008), such as those designed for childhood obesity intervention (Kowatsch et al. 2017) or for promoting well-being and positive thoughts in young people undergoing cancer treatment (Greer et al. 2019).

In order to make ECAs more effective, humanlike empathy expressed during conversation is a vital component (Bickmore, Gruber, and Picard 2005). Empathy is an essential part of building rapport and creating a bond, or a working alliance, between the patient and therapist to improve patient adherence and satisfaction (Bennett et al. 2011). Empathy has been defined as ‘an observer reacting emotionally because he perceives that another is experiencing or about to experience an emotion’ (Paiva et al. 2005), (p.4), or ‘the process whereby one person feels her/himself into the consciousness of another person’ (Wispé 1987). Empathy is expressed through a wide range of non-verbal behaviours, such as mirroring head nods, and verbal behaviours, such as the choice of words in social dialogues, the use of greetings and farewell rituals and the art of bringing continuity in the conversation (Laver 2011). In this paper, we focus on empathy expressed through verbal dialogue.

Empathy has been researched as a vital component of emphasis devices in human computer interaction to emphasize particular qualities and context (Wright and McCarthy 2008) which leads to building dialogue sets that contain the empathic cues, such as confirmation and adherence encouragement, that act as emphasis monitors. Rather than being specialized in a single quality or domain, a good open-domain conversational agent should be able to seamlessly deliver the necessary features into one cohesive conversational flow. Facebook’s blender bot is an example of such a conversational agent that has been specially trained on empathic dialogue sets (Roller et al. 2020). As in the Blenderbot approach, there is a tendency in artificial intelligence approaches to use mimicry and to replicate the behaviour of humans based on corpora without analysis of whether that behaviour is appropriate. However, mimicking/replicating human responses to a tragic event, for example, will not always be the best response (Lundqvist 1995) and will depend on the context such as the relationship between the parties. The importance of empathy in human dialogue has led to interest in ECAs expressing empathy particularly to bring about behaviour change (McRorie et al. 2009), (Ochs, Pelachaud, and Mckeown 2017), (Ravichander and Black 2018).

In this paper we evaluate the use of relational cues in recorded patient-doctor dialogues and the ECA’s known as Dr Evie (eVirtual agent for Incontinence and Enuresis) and as SAM (Sleep Adherence Mentor) designed to increase adherent behaviour in the domains of paediatric incontinence and sleep disorders, respectively. Both conditions have long specialist waitlists. Paediatric Incontinence affects up to 20% of school aged children (Malhotra et al. 2020). It often leads to avoidance of social interactions, low self-esteem and poor quality of life. Sleep disorders also impacts quality of life for many children (Roth 2007).

Due to COVID-19, there was an opportunity to record 30 online consultations (15 minutes to 2 hours) in the Incontinence Clinic and Sleep Clinic. We used qualitative research methods that help in building structured linguistic frameworks by analysing conversations (Alam, Danieli, and Riccardi 2018) and further utilizing them in building dialogue sets for ECAs. The objective is to semantically and pragmatically analyse the actual human dialogues involving a coding framework comprised of 16 relational cues identified in the literature using NVIVO and Discursis. Relevant to the context, our analysis identifies differences between new and follow-up patients and different

clinician roles (paediatrician, psychologist, physiotherapist, nurse). The results from the coding framework are further analysed statistically to more deeply understand the relationships, context and use of relational cues between different health specialists and patients towards specification of an empathic dialogue framework that supports complex interactions. We also identify design features for future ECA dialogue development and improvement in their emphatic structures.

In the next section we describe the background domain relevant to the online consultations and evaluated ECAs (Section 2), followed by review of the literature in empathy and relational cues in dialogue (Section 3). The methodology is presented in Section 4 followed by results in Section 5. Finally, discussion, conclusions and future work are presented.

2. Background

According to the International Children Continence Society (ICCS), the medical condition of incontinence refers to intermittent or continuous bed wetting during the day or night or both (Maternik, Krzeminska, and Zurowska 2015). Paediatric incontinence is a common condition affecting up to 20 percent of school-aged children (Malhotra et al. 2020) in many of their social activities like sports and sleepovers, which often leads to avoidance of social interactions. The children feel embarrassed and anxious, which leads to frustration and low self-esteem (Theunis et al. 2002). Children report a negative self-image owing to the physical and psychological impact of having incontinence, which is often unrecognised and seen as just another milestone in their growth (Butler 1998);(Harter 1982).

Despite the fact that incontinence impacts the patient's quality of life and is a cause of stress for them and their families (Malhotra et al. 2020);(Thibodeau et al. 2013) long waiting times to receive treatment are common, up to two years. This is because of a shortage of specialists, as incontinence is categorised as non-life threatening. Hence, ECAs could provide more timely support (Richards and Caldwell 2017); (Laranjo et al. 2018).

To address the problem of long public hospital waitlists, an incontinence specialist for children aged 3-18 at the Children's Hospital at Westmead (CHW) in Sydney, Australia, created an interactive eHealth program known as eADVICE (electronic Advice and Diagnosis Via the Internet following Computerised Evaluation). eADVICE enables young patients accompanied by their parents to get an online consultation regarding incontinence treatment factoring in the patient's medical history and encoded algorithmic response scenarios that capture the domain knowledge of the health experts. Developed in 2016, the website was evaluated in several pilots that found adherence to the six possible recommended treatments was around 50 percent. To allow patients and families to 'discuss' their treatments, eADVICE was enhanced through the addition of an ECA known as Dr Evie (eVirtual agent for Incontinence and Enuresis), which gave a human embodiment to the online consultation experience. Possessing the actual voice of the incontinence specialist, this ECA significantly improved the adherence and health outcomes of patients on the hospital waiting list (Richards and Caldwell 2017). The success of Dr Evie can be attributed to its availability and its empathic and empowering dialogue (Bickmore 2004).

Owing to its success for incontinence patients, the eADVICE approach – involving a website to provide tailored recommended treatments and an ECA to discuss the treatments – has been deployed for sleep disorders (eADVICE-sleep), another condition that is not life-threatening but significantly reduces quality of life (Roth 2007). Roth

(2007) associated sleep disorders with “the presence of long sleep latency, frequent nocturnal awakenings or prolonged periods of wakefulness during sleep periods”. This condition is considered chronic if the sleep environment is comfortable but the daytime routine is full of distress, light headedness and anxiety due to lack of sleep (Kredlow et al. 2015).

As sleep disorder patients also suffer from long waiting periods to access specialists and poor treatment adherence, they can potentially benefit from an ECA (Horsch et al. 2012); (Yin, Ring, and Bickmore 2012). Known as SAM (Sleep Adherence Mentor), the ECA in eADVICE-sleep acts as a virtual sleep coach. SAM has eight dialogue sets to cover the range of treatments and to ensure the dialogues are appropriate for the child’s age.

3. Empathy, Relationship Building and Relational Cues

This review first defines empathy and then briefly reviews its role in human relationships and past ECA work involving empathy and human-ECA relationship building. To provide the basis for the coding themes in the methodology, we provide a brief review from linguistics to identify and define verbal relational cues and types of dialogue expressions that have been found to assist relationship building.

Empathy is a complex human behavioural phenomenon defined by Hoffman (2001) as ‘the cognitive awareness of another person’s internal states that is, his thoughts, feelings, perceptions and intentions’ (p.29). Hoffman refers to empathy as any emotional reaction compatible with (but not necessarily similar to) the other’s situation. (Rogers and others 1959) theory of positive psychology and his client-centred framework emphasise that ‘for a person to ‘grow’, they need an environment that provides them with genuineness (openness and self-disclosure), acceptance (being seen with unconditional positive regard) and empathy (being listened to and understood)’(Mamarimbing 2021), (p.8).

3.1 Empathy and Relationship Building with ECAs

There has been more than a decade of research on the importance of empathy in human dialogues, which has led to interest in how ECAs can express empathy to bring about behaviour change (McRorie et al. 2009);(Ochs, Pelachaud, and Mckeown 2017);(Ravichander and Black 2018). ECAs typically have a particular purpose referred to as task-oriented empathy (Bickmore, Caruso, and Clough-Gorr 2005); (Bickmore et al. 2010), which is more easily detectable in focused dialogues. However, social empathy, which is not task based, is also important, as it offers comfort and encourages long-term relationships (Bickmore 2004). According to Halpern (2007), task-oriented empathy comes more naturally in doctor-patient dialogues while a doctor gathers the patient’s background information or recommends a certain treatment. Non-task-based, or social, empathy is more generic to the conversational themes in daily life. In designing ECAs, the component of social empathy is more complex and has been less commonly analysed (Halpern 2007).

Owing to the importance of empathy in human relationships, many researchers have created and evaluated empathic ECAs such as GRETA (Hartmann, Mancini, and Pelachaud 2005) and REA, the Real Estate Agent (Cassell et al. 1999). Building and maintaining human-ECA relationships, however, is broader than congruent expression of empathy and includes other behaviours. Long-term relationships are highly influenced by the use of the right relationship-building dialogues (Stafford and Canary 1991).

According to the psychology or medical literature, a working alliance is important for successful therapy (Halpern 2007). Many scales have been developed that emphasise the use of empathic and social dialogues during health-related consultations (Looije, Neerinx, and Lange 2008); (Yin, Ring, and Bickmore 2012).

ECAs with empathic dialogues have been studied across a diverse range of health programs such as relational agents for anti-psychotic medication adherence (Bickmore et al. 2010), avatar-based health intervention to modify unhealthy lifestyles (Lisetti et al. 2012), exercise advisors that interact with older adults (Bickmore, Caruso, and Clough-Gorr 2005) and ECAs that can help cancer patients to adopt a positive lifestyle after chemotherapy (Greer et al. 2019). Research has suggested frameworks that determine the useful verbal and non-verbal behaviours for virtual agents, such as 10 cues including empathy, social dialogues and continuity (Bickmore, Gruber, and Picard 2005), the Big Five model of personality traits (Neff et al. 2010), five dialogue characteristics that exhibit relationship building (Richards and Caldwell 2017) and annotation schemes for negative emotions' handling in customer care bots (van Velsen et al. 2019).

3.2 Relational cues in Linguistics and Psychology

Bardovi-Harlig (2010) defines pragmatics as 'the study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication' (p. 221). Research on social cognition through pragmatics (Bosco et al. 2015) has resulted in empathic pragmatic models (Zhanghong and Qian 2018) that refer to empathy in the context of verbal utterances that help build strong relationships.

This paper focuses on identifying a set of verbal relational cues that have been reported in the literature, particularly in psychology, to build a working alliance or strong rapport between the patient and the health specialist. The literature has identified a number of relational cues as in the following examples.

Empathy as a cue: Empathic phrases can be divided into three categories – Queries, Clarifications and Responses – which will be the foundation of empathic cue detection in the dialogue set (Coulehan et al. 2001).

Social dialogue: Conversation can be broken into three phases: the opening, middle and closing phases with social dialogues playing their role in opening and closing phase (Laver 2011).

Reciprocal self-disclosure: Conversations use a key social strategy defined as self-disclosure to build relations and increase conversational depth and as a process of disclosing details about yourself to the listener (Ravichander and Black 2018).

Meta-relational Communication: The specific talk that results in enjoyable relationships, cooperation, building up self-esteem, giving compliments, being courteous and polite, mitigating criticism, fostering patience and forgiveness in the participants, encouraging openness and talking about the relationship's quality and needs, and helping to acknowledge the relationship is categorized as the meta-relational communication (Stafford and Canary 1991).

Continuity: During a conversation, the behavioral units that exhibit continuity are of three types: prospective, introspective and retrospective. For a doctor-patient interaction, prospective and retrospective are covered under verbal interaction (Gilbertson, Dindia, and Allen 1998).

Reference to mutual knowledge: Some examples of mutual knowledge include knowledge of participant's/patient's biography, present life or habits, recent or future events (Planalp and Benson 1992).

Affirmation: Something emotionally challenging or complex to comprehend is related to an affirmative response from the listener. The complexity can be due to multiple reasons ranging from whether or not take up a medicine or follow a recommended treatment that is not fully understood or experienced. The main purpose of affirmation is to enable to patient to express their frustration openly. Letting the patient voice out their concerns and responding in understandable utterances is the essence of affirmation. (Cameron et al. 2015).

Confirmation: Means to reiterate the facts or validating the correctness of something previously believed or suspected to be the case is defined as confirmation. In eHealth, through confirmation, the physician recognises the situation conveyed by the patient and in return receives confirmation from the patient who sees the specialist as someone to confide in and understand their issues (Abramovitch and Schwartz 1996).

4. Methodology

4.1 Data Collection

CHW recorded its live clinical calls for research purposes during the COVID-19 period from June to November 2020. In these calls, the patients are children aged 3–18 and they have specific issues such as incontinence of urine. The objective of this study is to analyse these actual human dialogues using qualitative research methods and draw findings focused on the element of empathy in doctor-patient conversations. This live-call dataset provides a unique view of how actual dialogues, recorded in a real-life environment, can suggest changes to the existing ECA dialogues and help to validate the existing component of empathy in these ECAs. The health specialists consists of two paediatricians (senior paediatrician is later referred to as senior doctor), one nurse, two physiotherapists and one registrar. All of them are experienced health specialists with experience ranging from 10-40 years. The data is composed of 23 unique patients and consultation sessions, with a total of 50,000 utterances, collected over six months from the incontinence and sleep clinics.

4.2 Data Pre-processing

The process of analysing live calls begins with transcribing the recorded calls. Transcription involves generating text files from the audio recordings followed by character identification (e.g. physio, nurse, doctor, patient or relative of patient) and validating the dialogue assignment to the respective character. To ensure privacy, elimination of the patient's personal information (e.g. name, contact number and email) is the next step. The process involved listening to the recording and removing any personal detail manually. We replaced it with random names and emails. The tools used for transcribing the data include Temi¹ and Transcribe Wreally².

1 www.temi.com

2 <https://transcribe.wreally.com/>

4.3 Qualitative Analysis Overview

Qualitative data analysis is an approach to finding patterns in conversations and interviews. The data itself can be unstructured, which means there are no predefined questions or predefined answers to choose from. The qualitative analysis method chosen depends upon not only the level of structure in the data but also how acquainted a researcher is with the subject being analysed. Figuring out the context of the data in the form of variables is known as coding. The description follows Bengtsson (2016), who used content analysis in the domain of nursing, which is close to the domain of interest. Coding has two main approaches, deductive and inductive. In the deductive approach, the researcher is familiar with the content of the discussions and has developed an understanding of the context. Hence, the codes or themes are known beforehand. In the inductive approach, the researcher reads through the conversation and figures out the common words, semantics and context before grouping them into themes to define the coding framework. The next step is to decide whether to code the exact words or phrases from the conversation as codes or themes or to go deeper and understand the underlying meaning of the dialogue content and define that as a code. The former is known as manifest analysis, where the codes are the exact content; the latter is latent analysis, where the researcher goes under the surface and defines the codes according to the research aims (Bengtsson 2016).

4.4 Coding and Annotation Approach

The present research approach is deductive latent analysis, which means that codes relevant to verbal behaviours leading to empathy will be extracted from a literature review and will have associated dialogues for further analysis. Each dialogue can be placed in multiple codes as well, depending upon the hidden context of utterances. Ensuring coding's credibility increases the measure of trustworthiness of the coding (Graneheim and Lundman 2004). Credibility mechanism is how similar or dissimilar the coding results are when another person tries to reproduce the results. While qualitative methods do not claim to be reproducible, one way to improve credibility is to seek agreement between different researchers who do the coding independently and then establish a consensus. The approach is not to validate the coding quantitatively but to open the forum for discussion in which all experts come to consensus about each other's way of coding (Woods and Catanzaro 1988).

4.4.1 Coding Process

The coding phase in qualitative analysis begins by analysing the dialogues one by one to find themes according to the literature review on verbal and non-verbal behaviours found during conversations. In this study, the focus is on verbal behaviours because audio recordings cover only the verbal aspects of conversation. While finding themes, the focus is on the element of empathy; hence, all behaviours that exhibit empathy will be considered (e.g. politeness, inclusive pronouns). It is important to consider both task-oriented and social empathy. After defining and distinguishing themes from the literature to avoid overlaps, 16 themes were identified in relation to expressing empathy through dialogue: ten themes from Bickmore et al. (2005) and six themes from Richards and Caldwell (2017) as shown in Figure 1.

For each relational cue, we have definitions, common key words and examples of sentences from the literature that can act as a set of guidelines for the application of the

annotation scheme to other data sets. These details can be made available by contacting the first author.

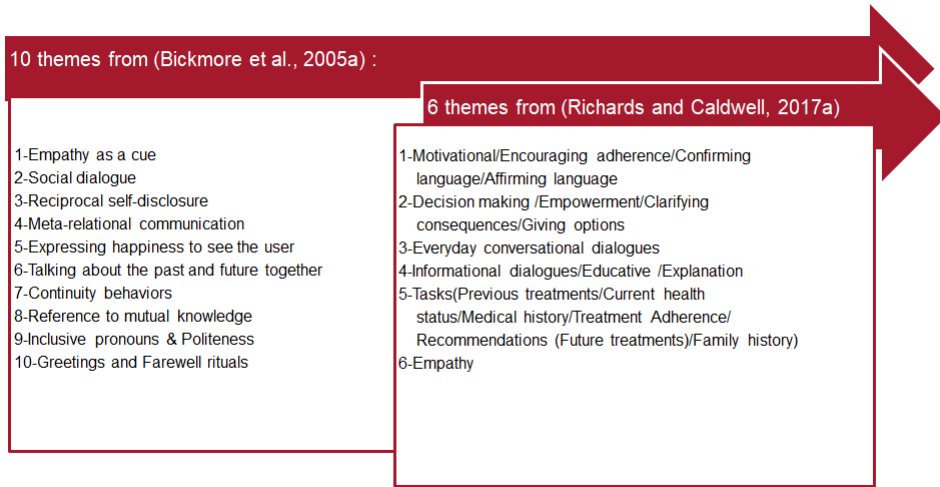


Figure 1

A breakdown of parent and sub-themes derived from the literature

4.4.2 Coding in NVIVO

Coding was done in NVIVO which began with building a framework of nodes, cases and roles based on the research questions. The detailed review of sixteen themes from literature resulted in 13 parent codes, out of which 4 parent codes were further categorized into 14 child codes as shown in Figure 2. For simplicity and for further relative comparison of codes within roles each dialogue was coded in one category only. Also the child codes were given preference over the parent codes. For example, in case of motivational dialogue, the child codes were encouraging adherence, giving option and clarifying consequences. Only in rare cases where the dialogue was motivational but did not fit any child code was it assigned to the parent code. In the end, child code dialogues were aggregated to create the total count of the parent code.

4.4.3 Independent Coders' Agreement through Cohen's Kappa

In order to seek agreement between how similar or dissimilar coding is, two annotators (SS and DR) took part in the validation process. SS was the main coder who had the context of approximately 50,000 dialogues. DR is an expert in qualitative analysis approaches but had not coded the whole dataset. After 200 dialogues were randomly selected from the first coder's assessment, the second coder was briefed on the codes' description, but the context remained missing until the first kappa was calculated so that consensus could be established in later discussions. The idea is to see how much the dialogues resonate with the theme even if the context is not given so that a more generalised coding can be obtained in the first iteration. The aim was not to validate the coding quantitatively but to open the forum for discussion in which all experts agree to each other's way of coding (Woods and Catanzaro 1988). Various coder's agreement techniques have been used in research including Krippendorff's alpha as well as Scott's pi and Cohen's kappa (Artstein and Poesio 2008). Cohen's kappa was selected as a

Parent Codes	Child Codes
1 : Continuity	
2 : Decision making	2.1 : Clarifying consequences
	2.2 : Empowerment
	2.3 : Giving options
3 : Empathy	
4 : Expressing happiness to see the Patient	
5 : Greetings and Farewells	
6 : Informational dialogues	6.1 : Educative
	6.2 : Explanation
7 : Meta-relational Communication	
8 : Motivation	8.1 : Affirmation
	8.2 : Confirmation
	8.3 : Encourage adherence
9 : Reciprocal self-disclosure	
10 : Reference to mutual knowledge	
11 : Social dialogues	
12 : Talking about the past and future together	
13 : Tasks	13.1 : Current health status
	13.2 : Family history
	13.3 : Medical history
	13.4 : Previous treatments
	13.5 : Recommendations (Future Treatments)
	13.6 : Treatment adherence

Figure 2
Relational Cues used in the Coding Process

measure of the agreement between the coder’s independent coding. An unweighted kappa score is calculated, which calculates the percentage agreement and the measure of coding by chance in each theme (Warrens 2015).

Initially, a kappa score of 0.61 was calculated on 200 dialogues with 24 themes. The coders examined the percentage agreement within each theme and agreed that three themes – everyday conversational dialogues, inclusive pronouns and politeness strategies – should be re-coded since the dialogues that were coded in them were multi-thematic and were more appropriately placed in other themes like empowerment and motivation. Both coders re-coded the original 200 dialogues in these themes. Some confusion remained concerning dialogues that were specifically questions like ‘Do you have any more questions?’ or ‘Are you with me so far?’, as they were part of multiple themes. Based on a literature review, it was agreed that the appropriate classification was confirmation. These changes and clarifications resulted in a revised kappa score of 0.75. The remaining dialogues were reviewed one by one. After discussion about the categories, the second coder agreed to change allocations from affirmation to empathy, encouraging adherence to recommendation and from child theme to parent theme, if a dialogue had more than one child theme representation. Final coding resulted in agreement on 171 dialogues and a kappa value of 0.84.

The remaining cases were resolved through discussion. In most cases, the context was mandatory for the assignment because the dialogues had more than one coding

Relational Cue	Dialogue Snippet
Continuity	Oh, good. Okay. Well, um, any questions from Sam about it. He's doing very well. So let's talk again in about three months and hopefully will discharge you from our care.
Decision making	You could be just having the wet night occasionally. I don't know. So I'm going to send you your own chart.
Clarifying consequences	I want you to check what sorts of foods you're doing.
Empowerment	So, um, how what your side effects of the medicine is it can make you thirsty. And so I think that's a Good, because you need to learn to be the boss of your own body. So questions around toilet training.
Giving options	What age was Sam toilet trained as a toddler?
Empathy	And we'll just leave it at that. You don't have to keep going, unless you absolutely want to and it's easy. need to work there.
Everyday conversational dialogues	Okay, look, it's its 5050. I find everyone neither loves it nor hates it.
Expressing happiness to see the Patient	Okay, so kind of spot on. Glad you're there with mom!
Greetings and Farewells	That's a pleasure. Talk to you soon. Take care. Have a good day. Bye bye.
Inclusive pronouns	Then we might be making some progress by doing bedwetting alarm training if bladder is not normal. So there's a whole lot of stuff we can do together.
Informational dialogues	It's harder to do it when it's wet and cold and dark. It's true, you just want to roll out and go back to sleep. But I think what's happening is your bladder is telling your brain that it's cool.
Educative	And teenagers behaviors and habits and bodies are different. So what I know about teenagers that they all need a good nine hours sleep in order to function.
Explanation	You've got the gene. And if you haven't got the pieces in place, that's when things are easier to disrupt if you think, and it's more likely to have an accident.
Meta-relational Communication	Okay, and that's what we want. We want a bladder that is well and is stable and empties well when you go to the toilet and okay we're getting close to the end of it.
Motivation	Well, so we want her to be prompted every two hours to go to the toilet and she clearly needs a little bit more. We need to try and work out, but that's what we're aiming for a bladder big enough to hold well
Affirmation	You are doing absolutely wonderfully because wearing the nappies helps you get better sleep at night.
Confirmation	Okay. So, normally you sleep at nine o'clock. Is that right, okay.
Encourage adherence	It gets better, you find that initially you would have a large wet and then the wet gets smaller and smaller, which means you're responding better
Politeness strategies	Sorry, what was that!
Reciprocal self-disclosure	But three to four hours. Probably he would have to pee 20 times a day. I would, which is what he's doing.
Reference to mutual knowledge	But as mum rightly noticed you are not quiet but you are very close to turning into a teenager.
Social dialogues	I'm just going to briefly tell her about you and you correct me if you think I've got it wrong. Okay, I just missed the fun side that she can hear you.
Talking about the past and future together	You visited back in October last year and it may be worthwhile to come back in. So we can physically examine you and scan you to make sure your bladder is normal.
Recommendations	You can just put money in your savings, slow down, just to make sure you can do it, you can practice on food first. And then if you're good, then you can take the tablet.

Figure 3
Examples of dialogues' annotation after coders' agreement

category in them such as affirmation or explanation and the context placed it correctly in empathy.

Following consensus on the 200 dialogues, the main coder reclassified any dialogues in the 50,000 dialogues that had been reassigned to three themes – everyday conversational dialogues, inclusive pronouns and politeness strategies. Finally, the main coder confirmed that all of the dialogues used everyday conversational language and that use of personal pronouns would be automatically calculated by searching for the terms 'us' and 'we' as a rough but quick method of assessment that would enable comparison between roles and individuals. The coding process resulted in parent and child codes listed in Figure 2. A few annotated dialogues after consensus are shown in Figure 3 and Figure 4 to give a glimpse of dialogues and their chosen relational cues.

4.5 Empathic Cues: Analysing the Dr Evie and SAM Dialogue Sets

Dr Evie and SAM both have structured dialogues with empathic cues already being a part of their semantics. The hypothesis of this study is that live unstructured sessions can bring out different sentence structures that are more beneficial in creating a level ground of adherence for the patients. This includes a validation process in which Dr

Relational Cue	Main Annotator	Second Annotator	Final Annotation
that's okay. That's okay. So currently, I'm thinking you don't wear any pull ups or protection to bed. Correct.	Empathy	confirmation	Empathy
So if we can break her toileting into daytime going to wee, nighttime wetting and Poo's, the plan will be more managable.	Inclusive pronouns	Educative	Educative
Make sure you're getting enough sleep, and I know that's a big ask for teenagers. But honestly, if you want to get Dry. That's really important.	Reference to mutual knowledge	Every day conversational dialogues	Empathy

Figure 4
Inter coder disagreement example dialogues

Evie and SAM would be thoroughly screened for empathic cues finalised in the prior coding process. A similarity scale would determine the threshold, and based on the similarity scores further empathic cues could be recommended. The same coding process used in Sections 4.4 was followed to evaluate these. After normalising the results, a comparison has been done to show how empathic both dialogue sets are.

5. Results

This section covers the analysis of 23 patient's consultations. The age of the patients ranges from 9-12 years with 10 patients being females while the rest are male. The consultations consist of 12 follow-ups and 11 new patient sessions. In 22 session, patients are accompanied by one or both parents and the average session length is 40 minutes. Out of 23, 16 sessions have one health specialist while 7 sessions have 2 health specialists with one always being the senior doctor. The role of the medical specialist and whether the consultation was a new or follow-up meeting are used for the comparison of results. The findings from the recorded consultations are discussed in Sections 5.1 to 5.3. Section 5.4 analyses results from the recorded consultations with the Dr Evie and SAM dialogues. In Section 5.5 the dialogue structure and topic sequencing is formulated into the coding framework.

5.1 Percentage Distribution of Relational Dialogues

The percentage of dialogue usage was compared for the following six categories:

1. Senior doctor's consultations with first-time patients
2. Senior doctor's consultations with follow-up patients
3. Average usage of relational cues by the senior doctor with follow-up versus new patients (Figure 5)
4. Physiotherapist's consultations with patients
5. Nurse's consultations with patients
6. Usage of relational cues in sessions where both the senior doctor and physiotherapist are present (Figure 6).

It is evident from the coding percentages shown in Figure 5 that patients who visit the clinic for the first time need more information about their health issue and its remedies. The senior doctor uses more informational and motivational dialogues and decision-making is also encouraged. The health specialist also needs to ask about a new patient’s medical history and current health status more than for the follow-up patients.

For follow-up patients, the percentage usage of cues is more variable as it depends on how many sessions the patient has had before and their progress as to date. For some patients, it is more about encouraging adherence to a treatment discussed in previous sessions. For others, the effects of new recommendations need clarification and dialogues confirm their understanding. Figure 5 also shows that the senior doctor’s empathic cues’ usage is the same for both new and follow-up patients but the difference lies in encouraging adherence and informational and motivational dialogues. Informational dialogues are more for new patients, whereas motivational and decision-making cues are more for follow-up patients. The reciprocal self-disclosure and social dialogues feature less in percentage usage but they are relatively more used for first-time patients.

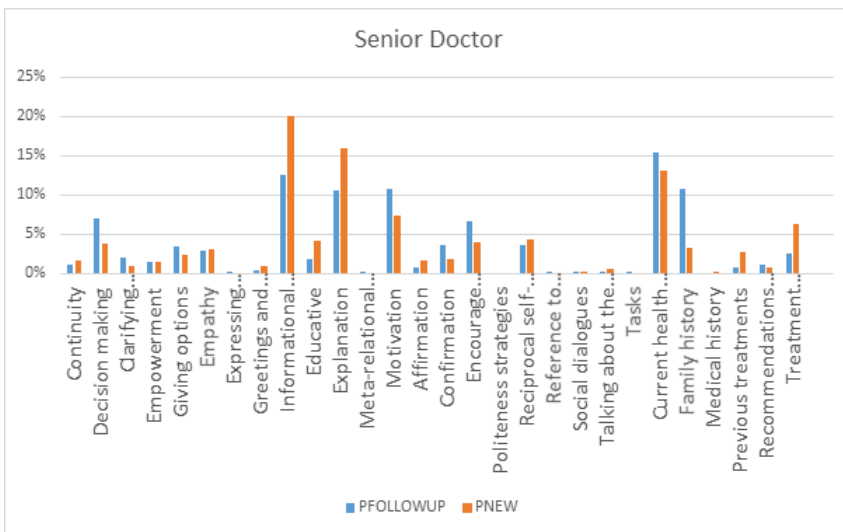


Figure 5 Senior doctor’s average usage of relational cues with first-time (PNEW) and follow-up (PFOLLOWUP) patients

The physiotherapist and nurse are the other most frequent roles found in the consultations. Most of the consultations contain one therapist role but in consultations where multiple specialists are present, as shown in Figure 6, where both senior doctor and physiotherapist appear, the physiotherapist uses more task-based dialogues and the senior doctor uses more adherence and decision-making dialogues. Social dialogues are uttered more by the physiotherapist, but empathic cues are uttered more by the senior doctor.

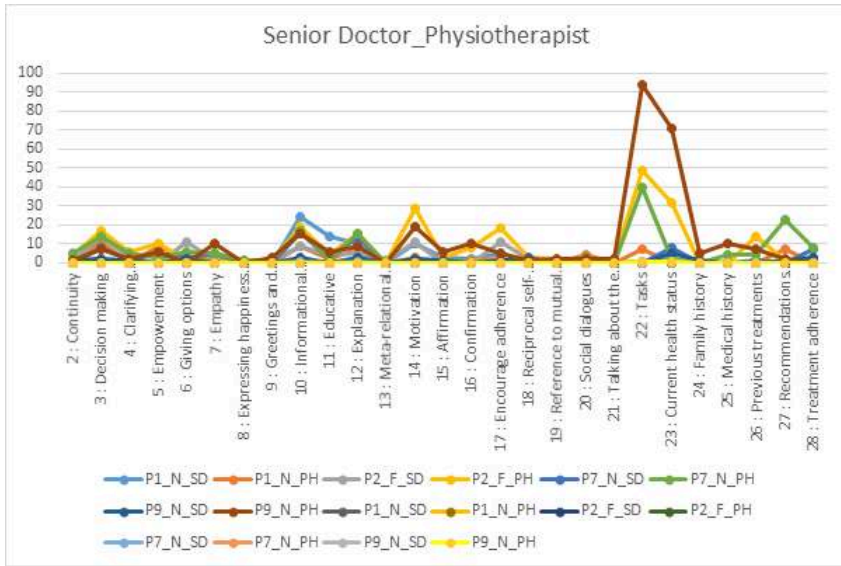


Figure 6 Senior doctor’s and Physiotherapist’s average usage of relational cues with first-time (PNEW) and follow-up (PFOLLOWUP) patients

5.2 Role Differences and Similarities in the Use of Relational Dialogues

The data helps in analysing the use of relational cues in multiple roles. In total, there were six roles to analyse, with more consultation sessions for the senior doctor, physicians and nurse. The remainder of the medical team – a paediatrician, registrar and male physiotherapist – had only one consultation each. More data is needed to analyse their use of relational cues, but for the purpose of this research they are grouped into one category under ‘others’. Owing to few data samples and lack of confirmation of normal distribution in the usage of a particular theme within a category under observation, non-parametric (i.e. Mann Whitney U) tests were chosen to understand the differences and similarities in the use of relational dialogues. The Mann Whitney U test compares outcomes between two independent groups based on the median of two distributions. The test was performed on all relational cues. The results that are significant at 90 percent confidence level or have high U-values include empowerment, explanation, social dialogues and reciprocal self-disclosure when comparing senior doctor’s usage of relational cues with junior doctors. These categories were further analysed to determine whether the senior or junior doctors used more of these cues.

Dominance is defined as the higher usage percentage of the relational cue in most of the sample points for that role. It is evident that Senior doctor is dominant in using relational cues like explanation, reciprocal self-disclosure and clarifying consequences while junior doctors are dominant in using relational cues like empowerment, social dialogues and task-based dialogues. Examples of visualisations that support the statistics are presented in Figure 7.

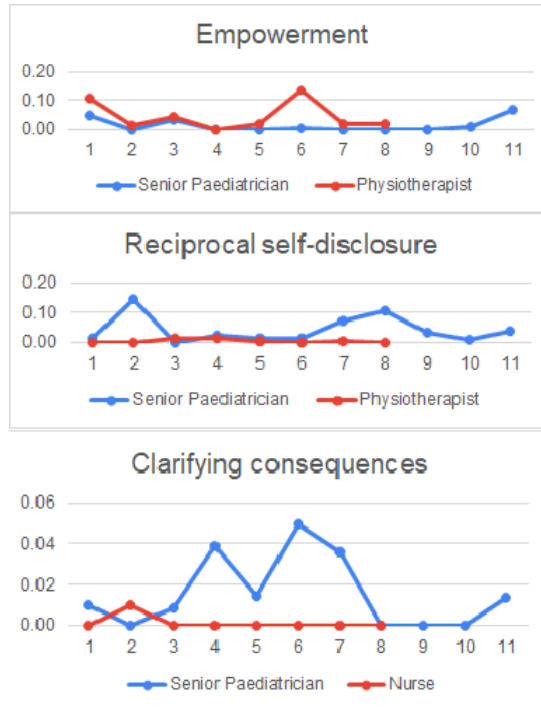


Figure 7
Comparison of Senior Paediatrician and Physiotherapist. X-axis represents specific patients, Y-axis is percentage of relational cue usage

5.3 Demographic Influence on Use of Relational Dialogues

The influence of patient demographics including gender, age and first time visitation on the use of relational dialogues has been analysed using Mann-Whitney U scores in Figure 8. The higher the U-value, more significant is the difference of relational cues usage in the respective groups. The p-value and z-score further determine the significance of the differences in the given empathic cue’s group. Figure 8 also shows U-values of new and follow-up consultations. Although first time visitation is not the demographics of the patient but it has a major contextual influence on the usage of relational cues. The new patients needed to be walked through the whole treatment details and were encouraged to speak up about their current health issues. The follow-up patients were either asked for time and volume charts or about their adherence to a certain treatment that had been discussed in previous consultations.

5.4 Comparison of Dr Evie’s and SAM’s Dialogues with Live-call Dialogues

Dr Evie’s dialogue set consists of multiple treatment-based dialogue streams, including alarm training, bowel program, caffeine intake, fluid increase, medication and time voiding. The same coding framework used for the recorded consultations was applied to the dialogues used in Dr Evie to facilitate a comparison between the usage of rela-

Group	U-value	p-value	z-score
Males versus Females			
Motivational dialogues	49	0.044	-2.014
Explanation	54	0.077	1.771
Age <10 versus Age >=10			
Previous treatments	39	0.014	-2.464
Clarifying consequences	56	0.095	-1.674
New versus follow-up			
Motivational dialogues	35	0.007	-2.693
Encourage adherence	50	0.049	-1.965
Medical history	51	0.055	1.917

Figure 8
Mann Whitney (U-scores) for significant relational cues in patient’s six demographically variant groups.

tional cues on live clinical calls and Dr Evie. Since the dialogue sets are those used for creating the Dr Evie application, they cannot be compared in terms of actual sessions that patients have with Dr Evie. The whole dataset provides a statistical presence of relational cues, which are shown in Figure 9.

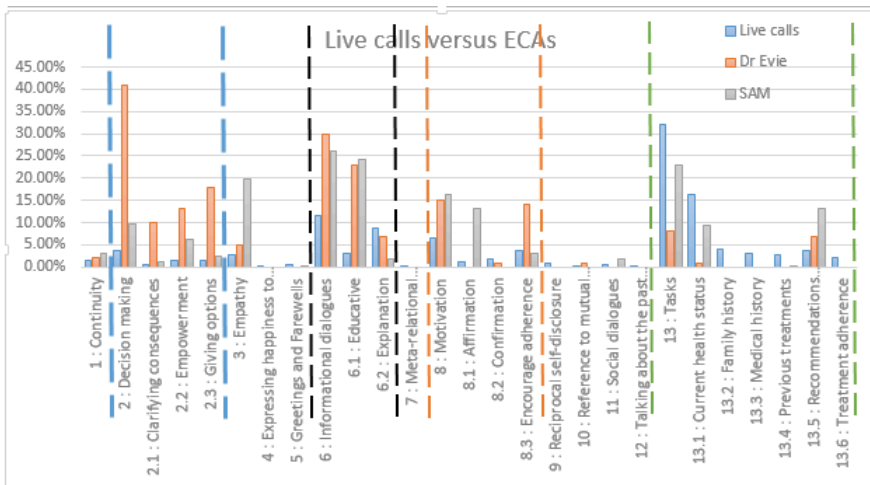


Figure 9
Coding comparison of live calls, Dr Evie and SAM. Dotted segregation is for each parent theme and corresponding sub themes. Sub themes counts add up to make parent theme counts.

The dialogue set of Dr Evie was designed by experienced health professionals who are experts in incontinence and other related disorders; hence, percentage usage of dialogue cues in live clinical calls is similar to that of relational cues in Dr Evie, especially for education, empowerment, encouraging adherence and giving options. The live clinical calls had a lower percentage usage of cues such as reference to mutual knowledge, greetings and farewells and the same can be seen with Dr Evie. The other cues such as empathy, explanation, recommendation, confirmation and clarifying consequences are also present in Dr Evie in good proportion and comparable with their usage in live clinical calls. Relational cues such as talking about the past and future together and social dialogues are missing in Dr Evie's context, Dr Evie's use of relational cues could be enriched by recommendations from live calls.

SAM's dialogue sets are mostly around sleep routine management and diet options: caffeine intake, regular sleep, night terrors, sleep hygiene and snoring issues. The recommendations are also about the sleep routine and diet habits. The dialogue flows have informational dialogues, current health assessment and recommendation dialogues incorporated in empathic language (e.g. 'I understand' and 'I know this is hard but'). SAM's dialogues are rich in empathic cues and affirmation dialogues, but lack cues for encouraging adherence, clarifying consequences and explanation. Live calls dialogue cues for these themes can be used to enrich SAM's dialogue set.

5.5 Structure, Topic and Inter-speaker Relationship Analysis

Health consultations have a defined and specific structure that exhibits the ontology of the subject area (Bickmore, Gruber, and Picard 2005). In order to embed relational cues in a logical manner, it is important to understand the conversational structure, topic variance and inter-speaker contribution in the overall consultation. The recurrence of topics and the time taken by speakers determine the level of engagement and understanding among the participants.

The dialogue structure in health consultations presented in Baker, Richards, and Caldwell (2014) places the dialogue cues from Bickmore et al. (2010) into a structure found in real consultations. Starting with greetings and farewells, social dialogues and previous treatment-related dialogues, it continues into more empathic dialogues and reciprocal self-disclosure cues. The last part of the conversation is more about future recommendations, adherence and continuity of the consultations. Our dataset is also mapped onto this logical structure to validate its existence in live calls, which suggests that even if the ECAs are built on natural language instead of structured questions, they would follow a similar structure as shown in Figure 10.

6. Discussion

The role of relational cues was analysed by the percentage of their occurrence in conversations. The analysis showed that two relational cues were used in all conversations but their percentage of usage in terms of utterance count by each role is very low – greetings, farewells and continuity. This confirms the (Laver 2011) finding that social dialogues are uttered mainly during the greeting and farewell phases of a conversation. Four relational cues were used very rarely in all conversations – expressing happiness to see the users, reference to mutual knowledge, talking about the past and future together and reciprocal self-disclosure. These cues are highly connected to how long term the relationship is with the patient and the level of comfort. The highest usage of relational cues is for empathy and motivational, informational and decision-making

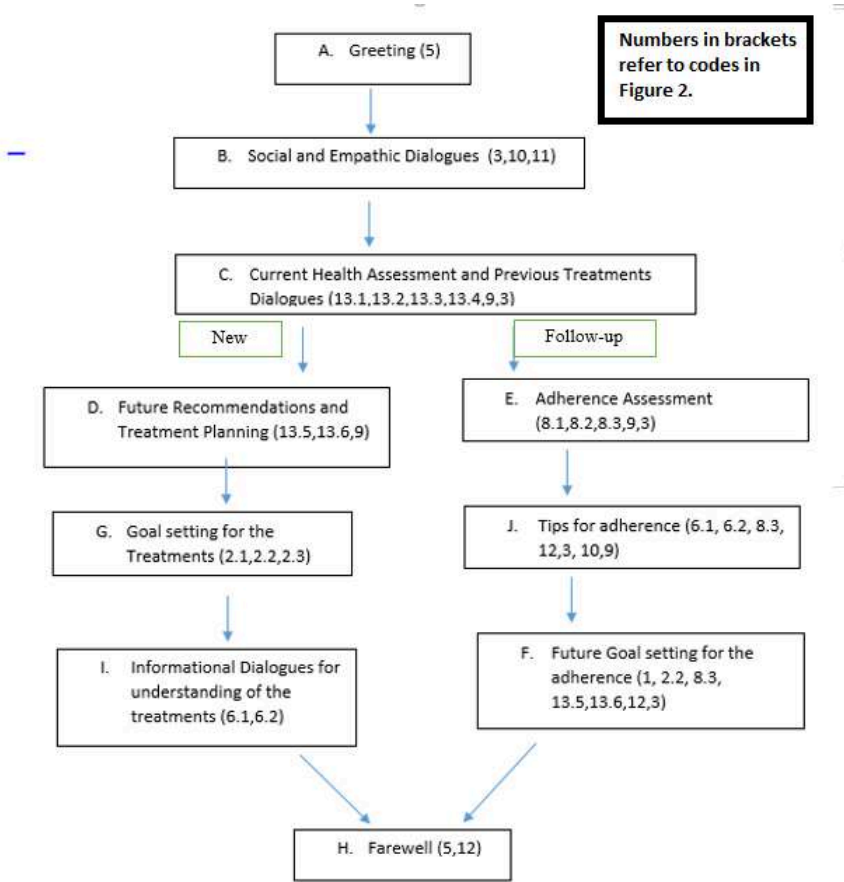


Figure 10
The Dialog Structure of the Live Calls

codes. Relational cues differ according to the session type and whether it is with a new or follow-up patient (Norfolk, Birdi, and Walsh 2007). A new patient needs to be onboarded with more communication related to treatment mechanics and education of the health issues. A follow-up patient needs more encouragement and dialogues related to clarifying consequences.

This analysis provided detailed findings on roles and demographics in the live calls. These insights can be used to create customised responses for different gender and age groups. Two factors that make the ECAs humanlike are their appearance and intelligence (Raval 2020), which depend highly on dialogue management producing dialogues similar to actual human dialogues. The live calls were more contextualised and customised. They included cues such as self-disclosure, confirmation and explanation. If the ECA dialogues were modified to capture more of a patient’s context, these cues could be used to enrich Dr Evie and SAM dialogues.

Multiple roles can be introduced in Dr Evie and SAM, as the senior doctor uses more explanation and empowering cues and the junior doctor takes care of certain routine tasks. This can make dialogue sets more empathic. To implement an approach where multiple ECAs can support a patient in different ways to provide holistic care, the

Council of Coaches platform could be used (op den Akker et al. 2018) where multiple ECAs review the patient's situation together and have separate conversations with the user according to their specific area of expertise or role (e.g. dietician, physiotherapist, friend).

The above suggested extensions to Dr Evie and SAM, such as more data gathering leading to more personalised use of relational cues, can be applied more widely to the design of other ECAs. This would involve the inclusion of user/patient models that persist between consultations, which allow the knowledge of the ECA to grow and be updated each time it meets the patient. This would also allow the ECA to tailor its social dialogue to the interests of each human. The relational cues and approach used in this article can be used to evaluate other ECAs and more relational cues could be incorporated to improve the relationship built with the user, with the aim of improving health outcomes. Another contribution of this paper is the capture of valuable datasets. The dataset and findings can be used by others in different ways. It is evident that live calls contain relational cues (e.g. disclosure and social dialogue) and dialogue patterns that could be used to design relational dialogue for ECAs. This dataset can also be used in machine learning and AI-based agents (Van Welbergen, Yaghoubzadeh, and Kopp 2014), which can learn to respond and formulate conversation using natural language generation. The live-call transcripts could also be used for the medical training of patient-doctor communication. This would be an alternative or supplement to approaches that offer guidance for health practitioners such as those provided by Rogers' (1959) client-centred therapy, which includes empathy, genuineness and unconditional acceptance. There is clear overlap in some of the relational cues and in the approach suggested by Rogers.

Finally, live clinical calls provide utterances from various roles. These roles include not only health specialists but also patients. Hence, the data can be used to build multiple ECAs that have different roles to facilitate the training of practitioners. While the quality of the recordings of patients and family members was too poor to allow transcription and qualitative analysis and outside the scope of this study due to its focus on practitioner use of relational cues, some specific calls in the live-call dataset can be used to build virtual patient ECAs, so that doctors can practise and refine their patient-doctor conversational expertise.

7. Limitations and Future Work

The dataset of live clinical calls was collected during the COVID-19 timeframe when normal clinical practice was disrupted. Live online clinical consultations became the new norm, which made their recording possible. Nevertheless, not all patients or practitioners were comfortable with this form of consultation, and it is possible that the dialogues were different to what might have been recorded in live face-to-face sessions in consultation rooms. The delay at the start to obtain consent prior to recording may have also inhibited the naturalness of the conversation and relationship. The ECAs for incontinence and sleep have been studied in conjunction with live calls, but more ECAs in domains specific to children should be explored to establish the use of relational cues especially in terms of the health specialist. As recommended by an anonymous reviewer, in the future it would be interesting to analyse empathy in a diachronic perspective, analysing the evolution of elements of empathy in doctor-patient conversations over time.

Conversational unit interfaces (CUIs) in health care are able to analyse natural languages (Laranjo et al. 2018) and to build responses according to the patient's situation

and history. As we move into the digital era, reliance on virtual agents that talk and understand like humans is a big research area (Sas, Whittaker, and Zimmerman 2016). Dr Evie uses scripted dialogues, whereas SAM uses more sophisticated technology that takes into account the user's goals and beliefs. The architecture SAM uses also allows preferences, medical history and other contextual features to be included in the ECA's reasoning and to provide explanations (Abdulrahman and Richards 2019). As Sam was developed later, we tried to incorporate more empathy and affirmation cues into its dialogues. Now that we have the data from live calls, Sam can further be enriched with encouraging adherence and clarification cues.

SAM and Dr Evie use fixed choice responses, primarily to ensure patient safety and accuracy, which is a current risk in health domains due to limitations in natural language processing (Xu et al. 2020). In the future when these limitations are addressed, safe and reliable solutions that use natural language input can be evaluated with a mix of controlled responses.

In the future, the recommendations for ECA dialogue design can be utilised to produce more dialogue sets that can be generalised over certain situations and cultures. Since negative thoughts can aggravate health issues, an empathic ECA that is personalised to the individual could help both mental and physical well-being. Hence, future agents for all health issues can potentially benefit from the relational cues and their usage presented in this paper.

8. Acknowledgments

Thanks to the health professionals, patients and families who agreed to have their consultations recorded

References

- Abdulrahman, Amal and Deborah Richards. 2019. Modelling working alliance using user-aware explainable embodied conversational agent for behaviour change: Framework and empirical evaluation. In *40th International Conference on Information Systems, ICIS 2019*, pages 1–17, Munich, Germany, December. Publisher Association for Information Systems.
- Abramovitch, Henry and Eliezer Schwartz. 1996. Three stages of medical dialogue. *Theoretical Medicine*, 17(2):175–187.
- Alam, Firoj, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61.
- Artstein, Ron and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 12.
- Baker, Scott, Deborah Richards, and Patrina Caldwell. 2014. Putting a new intelligent virtual face on a medical treatment advice system to improve adherence. In *Proceedings of the 2014 Conference on Interactive Entertainment*, pages 1–9, Newcastle NSW Australia, December.
- Bardovi-Harlig, Kathleen. 2010. Exploring the pragmatics of interlanguage pragmatics: Definition by design. *Pragmatics across languages and cultures*, 7:219–259.
- Bengtsson, Mariette. 2016. How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2:8–14.
- Bennett, Jennifer K., Jairo N. Fuertes, Merle Keitel, and Robert Phillips. 2011. The role of patient attachment and working alliance on patient adherence, satisfaction, and health-related quality of life in lupus treatment. *Patient education and counseling*, 85(1):53–59.
- Bickmore, Timothy, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling*, 59(1):21–30.
- Bickmore, Timothy W. 2004. Unspoken rules of spoken interaction. *Communications of the ACM*, 47(4):38–44.
- Bickmore, Timothy W., Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *CHI'05 extended abstracts on Human factors in*

- computing systems, pages 1212–1215, Portland OR USA, April.
- Bickmore, Timothy W., Kathryn Puskar, Elizabeth A. Schlenk, Laura M. Pfeifer, and Susan M. Sereika. 2010. Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers*, 22(4):276–288.
- Bosco, Francesca M., Ilaria Gabbatore, Claus Lamm, Rosalba Morese, Giorgia Silani, and Soile Loukusa. 2015. Social cognition: from empathy to pragmatic ability. In *11th International Conference on Cognitive Science*, Torino, Italy, September.
- Butler, Richard J. 1998. Night wetting in children: Psychological aspects. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 39(4):453–463.
- Cameron, Rachel A., Benjamin L. Mazer, Jane M. DeLuca, Supriya G. Mohile, and Ronald M. Epstein. 2015. In search of compassion: a new taxonomy of compassionate physician behaviours. *Health Expectations*, 18(5):1672–1685.
- Cassell, Justine, Timothy Bickmore, Mark Billingham, Lee Campbell, Kenny Chang, Hannes Vilhjálmsson, and Hao Yan. 1999. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 520–527, Pittsburgh Pennsylvania USA, May.
- Coulehan, John L., Frederic W. Platt, Barry Egner, Richard Frankel, Chen-Tan Lin, Beth Lown, and William H. Salazar. 2001. “Let me see if I have this right...”: words that help build empathy. *Annals of Internal Medicine*, 135(3):221–227.
- Gilbertson, Jill, Kathryn Dindia, and Mike Allen. 1998. Relational continuity constructional units and the maintenance of relationships. *Journal of Social and Personal Relationships*, 15(6):774–790.
- Graneheim, Ulla H. and Berit Lundman. 2004. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse education today*, 24(2):105–112.
- Greer, Stephanie, Danielle Ramo, Yin-Juei Chang, Michael Fu, Judith Moskowitz, and Jana Haritatos. 2019. Use of the chatbot “Vivibot” to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. *JMIR mHealth and uHealth*, 7(10):e15018.
- Halpern, Jodi. 2007. Empathy and patient–physician conflicts. *Journal of general internal medicine*, 22(5):696–700.
- Harter, S. 1982. A developmental perspective on some parameters of self-regulation in children. *Self-management and behavior change: From theory to practice*, pages 165–204.
- Hartmann, Björn, Maurizio Mancini, and Catherine Pelachaud. 2005. Implementing expressive gesture synthesis for embodied conversational agents. In *International Gesture Workshop*, pages 188–199, Berder Island, France, May. Springer.
- Hoffman, Martin L. 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.
- Horsch, Corine, Willem-Paul Brinkman, Rogier van Eijk, and Mark Neerinx. 2012. Towards the usage of persuasive strategies in a virtual sleep coach. In *The 26th BCS Conference on Human Computer Interaction 26*, pages 1–4, Birmingham United Kingdom, September.
- Kowatsch, Tobias, Marcia Nißen, Chen-Hsuan Iris Shih, Dominik Rügger, Dirk Volland, Andreas Filler, Florian Künzler, Filipe Barata, Dirk Büchter, Björn Brogle, et al. 2017. Text-based healthcare chatbots supporting patient and health professional teams: Preliminary results of a randomized controlled trial on childhood obesity. In *Conference: Persuasive Embodied Agents for Behavior Change (PEACH 2017) Workshop, co-located with the 17th International Conference on Intelligent Virtual Agents (IVA 2017)*, Stockholm, Sweden, August.
- Kredlow, M. Alexandra, Michelle C. Capozzoli, Bridget A. Hearon, Amanda W. Calkins, and Michael W. Otto. 2015. The effects of physical activity on sleep: a meta-analytic review. *Journal of behavioral medicine*, 38(3):427–449.
- Laranjo, Liliana, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.
- Laver, John. 2011. Communicative functions of phatic communion. In *Organization of behavior in face-to-face interaction*. De Gruyter Mouton, pages 215–238.
- Lisetti, Christine, Ugan Yasavur, Claudia De Leon, Reza Amini, Ubbo Visser, and Naphtali Rishe. 2012. Building an on-demand avatar-based health intervention for behavior change. In *Twenty-Fifth International FLAIRS Conference*, Marco Island, Florida, May.

- Looije, Rosemarijn, Mark A. Neerincx, and Vincent de Lange. 2008. Children's responses and opinion on three bots that motivate, educate and play. *Journal of Physical Agents*, 2(2):8.
- Lundqvist, Lars-Olov. 1995. Facial emg reactions to facial expressions: A case of facial emotional contagion? *Scandinavian journal of psychology*, 36(2):130–141.
- Malhotra, Neha R., Karen A. Kuhlthau, Iliana Rosoklija, Matthew Migliozi, Caleb P. Nelson, and Anthony J. Schaeffer. 2020. Children's experience with daytime and nighttime urinary incontinence—a qualitative exploration. *Journal of Pediatric Urology*, 16(5):535–e1.
- Mamarimbing, Stevanus Natanael. 2021. *The humanistic approach by Erin Gruwell in the freedom writers movie*. Ph.D. thesis, Widya Mandala Surabaya Catholic University.
- Maternik, Michal, Katarzyna Krzeminska, and Aleksandra Zurowska. 2015. The management of childhood urinary incontinence. *Pediatric Nephrology*, 30(1):41–50.
- McRorie, Margaret, Ian Sneddon, Etienne de Sevin, Elisabetta Bevacqua, and Catherine Pelachaud. 2009. A model of personality and emotional traits. In *International Workshop on Intelligent Virtual Agents*, pages 27–33, Amsterdam, The Netherlands, September. Springer.
- Neff, Michael, Yingying Wang, Rob Abbott, and Marilyn Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*, pages 222–235, Philadelphia, PA, USA, September. Springer.
- Norfolk, Tim, Kamal Birdi, and Deirdre Walsh. 2007. The role of empathy in establishing rapport in the consultation: a new model. *Medical education*, 41(7):690–697.
- Ochs, Magalie, Catherine Pelachaud, and Gary Mckeown. 2017. A user perception-based approach to create smiling embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–33.
- op den Akker, Harm, Riëks op den Akker, Tessa Beinema, Oresti Banos, Dirk Heylen, Björn Bedsted, Alison Pease, Catherine Pelachaud, Vicente Traver Salcedo, Sofoklis Kyriazakos, et al. 2018. Council of coaches a novel holistic behavior change coaching approach. In *4th International Conference on Information and Communication Technologies for Ageing Well and e-Health, ICT4AWE 2018*, pages 219–226, Porto, Portugal, March. SciTePress.
- Paiva, Ana, Joao Dias, Daniel Sobral, Ruth Aylett, Sarah Woods, Lynne Hall, and Carsten Zoll. 2005. Learning by feeling: Evoking empathy with synthetic characters. *Applied Artificial Intelligence*, 19(3-4):235–266.
- Planalp, Sally and Anne Benson. 1992. Friends' and acquaintances' conversations i: Perceived differences. *Journal of Social and Personal Relationships*, 9(4):483–506.
- Raval, Raturaj. 2020. An improved approach of intention discovery with machine learning for pomdp-based dialogue management. *arXiv preprint arXiv:2009.09354*.
- Ravichander, Abhilasha and Alan W. Black. 2018. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–263, Melbourne, Australia, July.
- Richards, Deborah and Patricia Caldwell. 2017. Improving health outcomes sooner rather than later via an interactive website and virtual specialist. *IEEE journal of biomedical and health informatics*, 22(5):1699–1706.
- Rogers, Carl Ransom et al. 1959. *A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework*, volume 3. McGraw-Hill New York.
- Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Éric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Roth, Thomas. 2007. Insomnia: definition, prevalence, etiology, and consequences. *Journal of clinical sleep medicine*, 3(5 suppl):S7–S10.
- Sas, Corina, Steve Whittaker, and John Zimmerman. 2016. Design for rituals of letting go: An embodiment perspective on disposal practices informed by grief therapy. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4):1–37.
- Stafford, Laura and Daniel J. Canary. 1991. Maintenance strategies and romantic relationship type, gender and relational characteristics. *Journal of Social and Personal relationships*, 8(2):217–242.
- Theunis, Marleen, Eline Van Hoecke, S Paesbrugge, Piet Hoebeke, and J Vande Walle. 2002. Self-image and performance in children with nocturnal enuresis. *European urology*, 41(6):660–667.
- Thibodeau, Betty Ann, Peter Metcalfe, Priscilla Koop, and Katherine Moore. 2013. Urinary incontinence and quality of life in children. *Journal of pediatric urology*, 9(1):78–83.

- van Velsen, Lex, Marijke Broekhuis, Stephanie Jansen-Kosterink, and Harm op den Akker. 2019. Tailoring persuasive electronic health strategies for older adults on the basis of personal motivation: Web-based survey study. *Journal of medical Internet research*, 21(9):e11759.
- Van Welbergen, Herwin, Ramin Yaghoubzadeh, and Stefan Kopp. 2014. Asaprealizer 2.0: The next steps in fluent behavior realization for ecas. In *International Conference on Intelligent Virtual Agents*, pages 449–462, Boston, MA, USA, August. Springer.
- Warrens, Matthijs J. 2015. Five ways to look at cohen’s kappa. *Journal of Psychology & Psychotherapy*, 5(4):1.
- Wispé, Lauren. 1987. History of the concept of empathy. *Empathy and its development*, 2:17–37.
- Woods, Nancy Fugate and Marci Catanzaro. 1988. *Nursing research: Theory and practice*. Mosby Incorporated.
- Wright, Peter and John McCarthy. 2008. Empathy and experience in hci. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 637–646, Florence, Italy, April.
- Xu, Jing, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.
- Yin, Langxuan, Lazlo Ring, and Timothy Bickmore. 2012. Using an interactive visual novel to promote patient empowerment through engagement. In *Proceedings of the International Conference on the foundations of digital Games*, pages 41–48, Raleigh North Carolina, May.
- Zhanghong, Xu and Wang Qian. 2018. Pragmatic empathy as a grand strategy in business letter writing. *English Language Teaching*, 11(8):14–27.

The Expression of Moral Values in the Twitter Debate: a Corpus of Conversations

Marco Stranisci*
Università degli Studi di Torino

Michele De Leonardis**
Università degli Studi di Torino

Cristina Bosco†
Università degli Studi di Torino

Viviana Patti‡
Università degli Studi di Torino

The present work introduces MoralConvITA, the first Italian corpus of conversations on Twitter about immigration whose annotation is focused on how moral beliefs shape users interactions. The corpus currently consists of a set of 1,724 tweets organized in adjacency pairs and annotated by referring to a pluralistic social psychology theory about moral values, i.e. the Moral Foundations Theory (MFT). According to the MFT, different configurations of moral values determines the stance of individuals on sensitive topics, such as immigration, civil rights, and gender equality. Moreover, an annotation of adjacency pairs' conversational dynamics is provided. Results of the analysis we applied on MoralConvITA shows that interesting patterns occur in the corpus, emerging from the intersection of moral studies and pragmatics that need to be generalized over larger corpora, and shedding some light on a novel promising perspective on the inter-user dynamics occurring in social media.

1. Introduction

The conversational nature of social media has been studied from several perspectives, among which, in the last years, community detection (Waseem and Hovy 2016; Lai et al. 2019; Vilella et al. 2020), and counter-speech analysis (Chung et al. 2019; Mathew et al. 2018; Fanton et al. 2021). Social media are indeed conversational environments where users and communities interact with each other, also producing conflictual situations, polarization, and sometimes toxic contents. That is the case of hate speech, that often affects the public online debate.

In particular, when people publicly debates about topics related to important societal challenges – those that trigger hatefulness – the conversation often takes the form of an exchange of moral values among social media users. In this context each single user can

* Dipartimento di Informatica - C.so Svizzera 185, 10149, Turin, Italy.
E-mail: marcoantonio.stranisci@unito.it

** Dipartimento di Studi Umanistici - via Sant Ottavio 20, 10124, Turin, Italy.
E-mail: michele.deleonard745@edu.unito.it

† Dipartimento di Informatica - C.so Svizzera 185, 10149, Turin, Italy.
E-mail: cristina.bosco@unito.it

‡ Dipartimento di Informatica - C.so Svizzera 185, 10149, Turin, Italy.
E-mail: viviana.patti@unito.it

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

indeed provide his/her values for confirming or contrasting those expressed by other users.

Nevertheless, despite the large variety of computational linguistic resources developed in the last few years for detecting hate speech and a wide range of related phenomena (Poletto et al. 2021; Fortuna and Nunes 2018; Schmidt and Wiegand 2017), to the best of our knowledge, the joint observation of conversational aspects and moral values involved has not been the major focus of any of their annotations. The only exception is the *Moral Foundations Twitter Corpus*¹, the large corpus for English described in (Hoover et al. 2020).

In this paper, following the research line started in our previous work about hate speech detection and the development of corpora (Sanguinetti et al. 2018) and benchmarks for this task (Basile et al. 2019; Sanguinetti et al. 2020), we want to investigate the relationship between conversation and moral values in contexts where users debate about topics that can trigger hate. Inspired by the above mentioned corpus created for English and aiming at developing a resource currently missing for Italian, we introduce a novel Italian social media corpus, where conversation dynamics are modeled in the perspective of the involved moral foundations. We especially focus on this level for observing the conversational interaction between users, with the main aim to shed light on the possible influence that the moral concerns expressed by the first message of the adjacency pair can exert on the second one.

For this purpose, we selected a discourse domain related to an issue that we know as especially relevant to moral values and a topic with sufficient popularity among Twitter users. We focused in particular on three categories of people especially vulnerable to hate speech, namely Roma, ethnic and religious minorities, and we drawn 1,724 tweets from TWITA (Basile, Lai, and Sanguinetti 2018) by using a keyword-based filtering. In addition, we also collected and organized data so that they can keep a record of the conversation dynamic where they were originally generated by users. The dataset consists indeed of adjacency pairs (Schegloff and Sacks 1973; Simpson 2005) of tweets that form micro-conversations where a tweet and a reply are generated by Twitter users self labeling as against discrimination by using the hashtag #*facciamorete* on their screen-name or user-description.

As far as the annotation of this corpus is concerned, the scheme mainly relies on the Moral Foundations Theory (MFT) categories. According to MFT, humans consistently rely on five moral concerns emerged as adaptive challenges: two individualizing foundations (Harm/Care and Fairness/Reciprocity) as they deal with the role of individuals within social groups, and three binding foundations (Ingroup/Loyalty, Authority/Respect, Purity/Sanctity), as they pertain to the formation and maintenance of group bonds (Weber and Federico 2013).

Going beyond the observation of moral concerns, we developed our annotation scheme also along two other directions that can better describe the inter-user interaction: the focus of the concern (on the violation or respect of the moral foundation expressed in the message) and the relation between the messages of the adjacency pair (whether the reply attacks, supports, or continues the conversation initiated by the tweet that is included in the same adjacency pair). An example annotated according to this scheme follows.

¹ The *Moral Foundations Twitter Corpus* is available at <https://osf.io/k5n7y/>

(1) **Tweet:** *Cara di #mineo sono 100mila euro al giorno per il business immigrazione, penso agli italiani in difficoltà... #portaaporta.*²

Concern expressed by the tweet: Ingroup-Loyalty/Betrayal

Focus: Prohibitive

Reply: *@user Paragonati ai 49 milioni di euro rubati dalla lega ancora pochi. A voglia di ospitare migranti.*³

Concern expressed by the reply: Fairness/Cheating

Focus: Prohibitive

Relation: Attack

As far as the suitability of the dataset within the context of applications, it has been observed (Kalimeri et al. 2019) that the detection of moral values together with other behavioral features of users might prove useful in general for designing more precise personalised services, communication strategies, and interventions, and can be used to sketch a portrait of people with similar worldview. Features based on moral concerns has been moreover proven to be useful in tasks related to sentiment analysis, see e.g. (Lai et al. 2021).

The paper is organized as follows. Section 2 briefly surveys related work, mostly focusing on MFT and its application in different contexts, and on pragmatics of conversation. Section 3 describes data collection and annotation, also discussing the inter-annotator agreement detected during the annotation process. Finally, Section 4 provides an analysis of moral and pragmatics features emerging from the gold standard corpus released. Section 5 concludes the paper and also addresses some future direction for the development of this research line.

2. Related Work and Theoretical foundations

According to the Moral Foundation Theory (MFT) individuals' moral beliefs are not universal, but reside on a plurality of "irreducible basic elements" that gives rise to many and sometimes conflicting moral configurations (Graham et al. 2013).

This theory unifies in five moral dyads the set of values originally proposed by Shweder (Shweder et al. 1997), i.e. *community*, *autonomy* and *sanctity*, and those discussed by Fiske (Fiske 1991), i.e., *communal sharing*, *authority tanking*, *equality matching* and *market pricing*. The moral dyads can be resumed as follows.

1. Care/Harm. Prescriptive concerns related to caring for others and prohibitive concerns related to not harming others.
2. Fairness/Cheating. Prescriptive concerns related to fairness and equality and prohibitive concerns related to not cheating or exploiting others.
3. Ingroup Loyalty/Betrayal. Prescriptive concerns related to prioritizing one's ingroup and prohibitive concerns related to not betraying or abandoning one's ingroup.

² Translation: #mineo's reception center they are 100thousands euros a day for the immigration business, I think to the Italians in distress ... #portaaporta

³ Translation: @user Compared with 49 millions euros stolen by the Lega party they are still a few. You can host a lot of immigrants ...

4. Authority/Subversion. Prescriptive concerns related to submitting to authority and tradition and prohibitive concerns related to not subverting authority or tradition.
5. Purity/Degradation. Prescriptive concerns related to maintaining the purity of sacred entities, such as the body or a relic, and prohibitive concerns focused on the contamination of such entities.

The morality of each individual is built upon a specific configuration of these concerns that are considered within the theoretical framework as partly innate, partly developed through experience and social relationships. This allows MFT's dyads to describe morality as organized in advance of experience, highly dependent on environmental influences collected during development within a particular culture, and to see moral judgments as intuitions that happen before the subject starts to reason.

Nevertheless, like, e.g., the list of basic emotions, whose definition and granularity meaningfully varies in different theories, also the MFT's list of basic foundations can be questioned and it cannot be in effect considered as the final list. MFT is a theory in motion, to be expanded but especially adequate for cross-disciplinary research, because it provides a common language for talking about the moral domain (Graham and Haidt 2012) also in different disciplinary contexts. For instance, several researches within this framework have been devoted to investigate relations between moral foundations and political ideology, referring in particular to the moral differences between liberals and conservatives (Graham, Haidt, and Nosek 2009), media studies (Winterich, Zhang, and Mittal 2012).

In recent years, MFT foundations in the online environment have been studied by some scholar together with its correlation with other topics, such as hate speech (Hoover et al. 2019), or political discourse (Johnson and Goldwasser 2018; Weber and Federico 2013). Concurrently, several resources to investigate this phenomenon have been released: corpora of annotated tweets (Hoover et al. 2020), dictionaries (Graham and Haidt 2012; Hopp et al. 2020), and knowledge graphs (Hulpus et al. 2020).

An especially interesting application of this theory is the Moral Foundations Twitter Corpus (MFTC) (Hoover et al. 2020). It is a large collection of English tweets annotated for moral sentiment built for advancing research at the intersection of psychology and Natural Language Processing. The collection focuses on seven distinct socially relevant discourse topics, among which that addressed in our dataset, i.e. hate speech and offensive language. The schema applied in the annotation separates the virtues from the vices of the moral dyads to consider the polarity of a message expressing a value.

Although our approach is inspired by the MFTC's major tenets, it addresses a different language, i.e. Italian, and adopts a revised version of the annotation schema which is multi-dimensional. Dyads are not splitted, and the user's focus on moral values is evaluated separately from the selected dyad. Moreover also conversational dynamics are evaluated, since the corpus consists of adjacency pairs of tweets, instead of single messages, for keeping a record of the conversation dynamic where they were originally generated by users.

Adjacency pairs are units of conversation consisting of sequences of two adjacent utterance length, produced by different speakers (Schegloff and Sacks 1973). The two messages are complementary: the first pair part assumes a specific kind of response (Levinson 1983). For instance, if the initial message contains a request, the reply will presumably express the function of an acceptance or a refusal.

Similarly, the Dialogue Act (DA) is a communicative activity with a certain commu-

nicative function, a semantic content, and an optional feedback dependence relation function (Bunt et al. 2010). A family of computational pragmatics models focuses on the identification of lexical, collocational, syntactic, or prosodic cues for DA detection in a message (Jurafsky 2004). Several annotation schemes derive from such models, among which the Dialog Act Markup in Several Layers (DAMSL) (Core and Allen 1997), implemented with some modification by (Stolcke et al. 2000), and ISO 24617-2 (Bunt et al. 2012). All of them list a set of function for DAs annotation. Recently, iLISTEN, a shared task for Italian consisting in automatically annotating dialogue turns with speech act labels, representing the communicative intention of the speaker, has been propose at the EVALITA evaluation campaign (Basile and Novielli 2018). The speech act taxonomy refines the DAMSL categories, based on two classes of functions (Cfr (Allen and Core 1997)): Forward Looking, the intended action expressed by the first pair part, and Backward Looking, which encodes how the reply is related with the original message. In our corpus, adjacency pairs internal structure often consists in a statement on immigration, accepted or rejected in the reply.

In our schema, replies are annotated with ‘attack’, that may imply rejection, ‘support’ and ‘same topic’, which can entail acceptance. However, these categories are not overlapping, since attacking, and supporting potentially fulfil other relevant functions to our work, such as outlining the moral or political stance of the speaker. Though, the two schema are mapped to support the qualitative analysis of conversational dynamics in Section 4.2.

3. MoralConvITA: A Corpus of Conversations with Annotated Moral Foundations

3.1 Data

In order to create the MoralConvITA corpus, a sample of 862 adjacency pairs of tweets were collected from January 2019 to June 2020. The data gathering process relied on the TWITA data set (Basile, Lai, and Sanguinetti 2018), and was structured as follows:

- all tweets generated by users self labeling as against discrimination with the hashtag *#facciamorete* on their screen-name or user-description were collected;
- the resulting selection was further filtered by using the Hate Speech corpus keywords (Sanguinetti et al. 2018);
- only reply messages were kept;
- first pair parts were retrieved through the Twitter Rest APIs.

In order to collect a meaningful amount of data where moral sentiment occurs, we choose a discourse domain related to an issue that we know as especially relevant to moral values and a topic with sufficient popularity among Twitter users. Nevertheless, considering that expressions of moral sentiment in one domain and about a specific topic might not generalize to data extracted from another domain, in future work we want to address other domains also.

3.2 Annotation

The task of annotating a corpus according to the MFT shares similarities with sentiment classification, but it also introduces notable challenges, such as the co-occurrence of many moral values in a message, their implicitness and subjectivity (Hoover et al. 2020). For addressing these challenges we discussed the design of the schema within the research group and we performed annotation trials on a small subset of the data before starting with the actual annotation process. Finally, for validating the schema, we carefully observed the behavior of each annotator and the agreement among the annotators, as reported in Section 4.

The schema we provided for MoralConvITA is centered on the MFT and under this respect inspired by the one applied in the Moral Foundations Twitter corpus for English. Moreover, in order to take into account the pragmatics of conversation, we defined also some other issue to be annotated for better representing the conversation dynamics. Three are the dimensions along which we annotated the adjacency pairs.

1. the most relevant **Moral Foundation** dyad, among the five pointed out by the MFT (Section 2);
2. the **Concern Focus** of the message, which may be prescriptive, if it highlights a virtue, or prohibitive, if it blames a misbehavior;
3. the **Conversational Relation** within the adjacency pair, representing whether the reply attacks, support or deals with the same topic of the first pair part.

Table 1 resumes the list of labels used for the annotation of each of these three categories.

Table 1
The labels annotated in the MoralConvITA

category	label
Moral Foundation	Care/Harm
	Fairness/Cheating
	Ingroup-Loyalty/Betrayal
	Authority/Subversion
	Purity/Degradation
Concern Focus	prescriptive
	prohibitive
Conversational Relation	attack
	support
	same topic
	no relation

Conversational Relation and Concern Focus dimensions were elaborated to better fit the annotation schema to the analysis of Twitter conversations. The former provides information about how the pairs of tweets relate to each other. The Concern Focus, instead, was introduced to mitigate the dichotomy between moral vices and virtues. In existing schemas a text can either express the respect for a moral concern or the stigmatization of its violation, but this distinction seems not to capture expressions that deliberately violate a moral value and may have a pragmatic effect. On this respect toxic speech, that often affects the conversation about migration, can be interpreted as a blatant violation of the Care/Harm dyad. Hence, we considered the Concern Focus as an independent dimension to annotate. For instance, instead of considering ‘care’, and ‘harm’ two separated labels, we treated them as a whole, and later evaluate their focus, that is ‘prescriptive’ if the message dwells on the moral rule to comply with, ‘prohibitive’ when its violation is reported by the user. Examples of tweets expressing moral dyads and their Concern Focus are listed in Table 2, while Conversational Relations are exemplified in Table 3.

It is worth highlighting some strategy we applied in the annotation. First, in addition to the dyads of the MFT we also used for the category Moral Foundation the label ‘no-moral’ when any moral concerns occurs in the message. Second, as far as the concern, it is annotated only in the messages where a moral foundation has been previously recognized by the annotator. Finally, the conversational relation is only annotated in the reply message for showing its link with the tweet that started the micro-conversation.

The annotation process involved a team composed of two skilled researchers, a man and a woman, and nine undergraduate university students, among which 3 men, and 6 women, aged 22-27. The skilled annotators were especially involved in designing and testing the schema, in tutoring the rest of the annotation and in solving the disagreement. Each of the nine students annotated at least 250 adjacency pairs along the three dimensions for building the corpus we actually released⁴ which includes 1,724 tweets, organized in 862 adjacency pairs.

The analysis of inter-annotator agreement (IAA), calculated using the the Fleiss’ Kappa IAA metrics and considering each of the categories annotated, is described in Table 4. It confirms the subjectivity of the task, which also results from the observations reported in (Hoover et al. 2020) for the Moral Foundations Twitter corpus for English. Considering that our corpus is organized in micro-conversations, we can report also some findings about the agreement detected in the perspective of the annotation of the conversations that compose MoralConvITA. In particular, the results provided in Table 4 highlight that the annotation of replies of the adjacency pairs has been affected by an also lowest agreement (0.17 for the Moral Foundation, and 0.18 for its Focus) with respect to the annotation of the tweets that initiate the conversation, while for the others it shows a fair agreement. The issue has been already pointed out by (Hoover et al. 2020) with respect to the development of the Moral Foundation Twitter Corpus. According to this study, the interpretation of morality in a text is subjective both for the annotators’ stance and for the lack of information about the author’s intention. Moreover, this low agreement among the annotators can be also motivated by the fact that the moral concern expressed in a message is often ambiguous because many values potentially coexist within the text. See for instance the following example.

⁴ <https://github.com/marcostranisci/MoralConvITA>

Table 2
Moral values annotated in the MoralConvITA corpus.

Moral Value	Example
Care/Harm	<p>@user Infatti lo dicevo perché entrambi erano cristiani! Concordo con lei che prima ci sono le Persone, che possono essere più o meno brave, cristiane o no, alte o basse...</p> <p>(@user In fact I said it because both were Christians! I agree with you that first of all there are the individuals, who can be more or less good, Christian or not, high or low...)</p>
Fairness/Cheating	<p>@user Ehm...e la crisi, la disoccupazione giovanile, sanità, strutture, l'istruzione. Queste non sono emergenze? No no.</p> <p>(@user Ehm...and the crisis, youth unemployment, health, facilities, education. These are not emergencies? No no.)</p>
Ingroup Loyalty/Betrayal	<p>Immagini esclusive di un gommone con 70 immigrati, scafista alla guida e motore potente, in acque maltesi. Qualcuno si degnierà di intervenire o li manderanno ancora una volta in direzione Italia???</p> <p>(Exclusive images of a dinghy with 70 immigrants, a driver and powerful engine, in Maltese waters. Will someone deign to intervene or) will they send them once again to Italy?</p>
Authority/Subversion	<p>A casa fanno la voce grossa e mostrano i muscoli con i disperati. A Bruxelles invece Salvini e company sono solo pecorelle di #Orban che è il primo nemico dell'Italia e che nega ogni giorno i nostri valori costituzionali.</p> <p>(At home they speak louder and show their muscles with the desperates. In Brussels instead Salvini and company are only sheeps of #Orban who is the first enemy of Italy and who denies every day our constitutional values.)</p>
Purity/Degradation	<p>#Iran, migliaia di prigionieri politici subiscono torture e maltrattamenti senza cure mediche. Libertà per #ArashSadeghi #FarhadMeysami #RajaeShahr e per tutti i dissidenti che non si arrendono al regime khomeinista.</p> <p>(#Iran, thousands of political prisoners suffer torture and ill-treatment without medical treatments. Freedom for #ArashSadeghi #FarhadMeysami #RajaeShahr and for all dissidents who do not belong to the Khomeinist regime.)</p>

(2) *Se io sono cittadino italiano non #Rom, allo Stato devo dire: dove abito, da quando ci abito, se sono sposata oppure no, quanti soldi ho in banca, devo pagare fino all'ultimo centesimo di tasse e se non faccio i vaccini mi denunciano. Scusate si può fare per tutti?*⁵

It could be intended as an instance of Ingroup-Loyalty/Betrayal, since it highlights a contrast between an ingroup (Italians) and an outgroup (Roma people). However, it

⁵ If I am an Italian citizen, and not a #Roma person, I must declare to my country: my place of residence, since when I live there, If I am married or not, my account balance, I have to pay every cent of taxes, and if I don't take the vaccine I am reported. Excuse me, this can be done for everybody?

Table 3
Relations linking tweets and replies annotated in the MoralConvITA corpus.

Conversational pattern	# Example
Support	<p>Tweet: <i>Oggi scopriamo dal Ministro Salvini che c'è una "questione rom" aperta. Ed io che pensavo che ci fosse invece una "questione MAFIA" aperta. O una "questione CORRUZIONE". E invece, dopo i migranti, si punta il dito contro un'altra minoranza. La miseria umana è tutta qui.</i></p> <p>Reply: <i>@user è il suo standard, prima i meridionali, poi gli emigranti e adesso i Rom... chissà chi punterà prossimamente..</i></p> <p>(Tweet: Today we discover from Minister Salvini that there is an open "Roma issue". And I thought that there was instead an open "MAFIA issue". Or a "question CORRUPTION". And instead, after the migrants, you point the finger at another minority. Human misery is all here.</p> <p>Reply: <i>@user it is his standard, first the southerners, then the emigrants and now the Roma...who knows who will point soon..)</i></p>
Attacks	<p>Tweet: <i>Da oggi anche l'Italia comincia a dire NO al traffico di esseri umani, NO al business dell'immigrazione clandestina. Il mio obiettivo è garantire una vita serena a questi ragazzi in Africa e ai nostri figli in Italia.</i></p> <p>Reply: <i>@matteosalvinimi NO alla propaganda fatta sulla pelle dei migranti! Più di seicento persone abbandonate in mare per scrivere un tweet? Vergognati per la tua disonestà.</i></p> <p>(Tweet: From today Italy too begins to say NO to human trafficking, NO to the business of the illegal immigration. My goal is to ensure a peaceful life for these children in Africa and our children in Italy.</p> <p>Reply: <i>@matteosalvinimi NO to the propaganda that negatively affect migrants! More than six hundred people banded at sea to write a tweet? Be ashamed of your dishonesty.)</i></p>
Same topic	<p>Tweet: <i>"Se si torna al voto come prima cosa dovremo costituire un fronte largo europeista da contrapporre al fronte anti-europeista di #Salvini e #DiMaio. L'Europa sarà la discriminante. #maratonamentana"</i></p> <p>Reply: <i>@user Da nord a sud le elezioni le vincerà di nuovo chi farà propaganda anti-migranti.. è questo il nodo fondamentale purtroppo!!</i></p> <p>(Tweet: "If we return to the vote as a first thing, we must form a broad pro-Europe front to join the anti-European front of #Salvini and #Dimaio. Europe will be the discriminating. #maratonamentana"</p> <p>Reply: <i>@user From north to south the elections will be won again by those who make anti-migrant propaganda.. this is the fundamental issue, unfortunately!!)</i></p>
No-relationship	<p>Tweet: <i>C'ho i parenti fasci e razzisti, mi vergogno tantissimo.</i></p> <p>Reply: <i>@user Fino a quando parli di diritti, migranti, accoglienza e amenità del genere nessuno cambierà idea. Se ai neosalviniani metteranno le mani in tasca, allora, potrai di nuovo discuterci. #SalviniDimettiti</i></p> <p>(Tweet: I have family members fascists and racists, I'm so ashamed.</p> <p>Reply: <i>@user As long as you talk about rights, migrants, reception and amenities of the generation nobody will change their mind. If the Neosalvinians get their hands in their pockets, then you can discuss it again. #Salvinigohome)</i></p>

Table 4

Fleiss’ Kappa for each label separately calculated for the tweet (which initiates the micro-conversation) and for the reply to the tweet.

label	Fleiss’ Kappa
Moral Foundation (tweet)	0.32
Moral Foundation (reply)	0.17
Concern Focus (tweet)	0.26
Concern Focus (reply)	0.18
Conversational Relation (reply only)	0.30

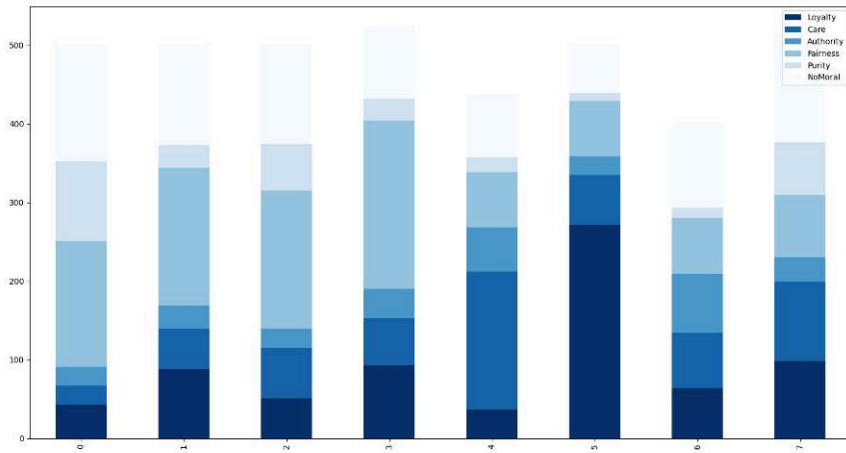


Figure 1

The distribution of moral foundations labels in the corpus (in tweets and in replies both) separately calculated for each of the 8 annotators (referred with numbers from 1 to 8).

could also express a concern on authority for its reference to the need of respecting country laws and therefore annotated with the label Authority/Subversion.

Separately calculating the distribution of the moral foundations for each annotator and putting together the tweets and the replies (as we did in Figure 1), we can see that some bias occurs and that some annotator used a very large amount of some label with respect of the average of the annotators. For instance annotator 9 used Ingroup-Loyalty/Betrayal more that twice that the other annotators.

A more general bias was moreover expected in our annotation, which depends on the involved annotators. Their age and skill is related in literature with a basically liberal vision, rather than to a conservative one, and to the exploitation of some specific moral foundation in the interpretation of messages. While conservative people tends to use all the moral foundations of the MFT spectrum, liberal people only rely its judgement

on the first ones, mostly Care/Harm, Fairness/Cheating and Ingroup-loyalty/Betrayal. This is confirmed by the analysis provided in the next section.

4. Analysis of the MoralConvITA Corpus

The final version of the MoralConvITA corpus consists of 1,724 tweets arranged in adjacency pairs annotated by at least three annotators, but only in 487 cases annotators reached a partial or total agreement on all the five dimensions of the annotation schema. Excluding all the tweet labeled as 'no-moral', the corpus reduces to 253 adjacency pairs. We thus chose to separately analyze moral foundations, moral focus, and conversational patterns.

The distribution of Moral Foundations in the adjacency pairs is discussed in Section 4.1, while an analysis of how foundations are shaped by the conversation is provided in Section 4.2.

4.1 Moral Foundations

The distribution of the labels annotated will be analyzed in this section according to two different perspectives, that is the moral dyads provided by the annotators and their occurrence in the tweets rather than in the replies, as shown in figure 2.

Two are the prevalent moral foundation dyads annotated in the corpus: **Fairness/Cheating** (408 occurrences), and **Loyalty/Betrayal** (255 occurrences). They both seem to be very specific to the topic of migration, since the latter draws a distinction between who is Italian and who is not, while the former is often used to report the hypocrisy of public players that deal with this topic.

In particular, the accuse of cheating follows two rhetorical patterns: the reception of asylum-seekers as a business, mainly occurring in original tweets, and the exploitation of migration for political propaganda, occurring in replies. This second case is more traditional in the Italian public debate, and most common in it. In fact, 67.7% of Fairness/Cheating labels occurs in replies, most of them focused on the anti-immigration proposals' inconsistency, and lack of actual effectiveness. For the same reason, the 9.6% of in-agreement adjacency pairs consists of a statement oriented to the Ingroup-loyalty/Betrayal value, and a response in which the Fairness/Cheating concern is present.

The 'immigration as a business' moral charge, more frequent in the first element of the pair, is a quite recent rhetorical argument, but its fast diffusion could be interpreted as a reshape of the traditional separation between liberals and conservatives (Haidt 2012). For instance, in the example (3), irregular migrants are depicted as victims of a foul game by pro-immigration organization, and the closing invective contains an exhortation to help them not only in words, but also in a concrete way.

(3) @user *Tutto inutile, lei sarà arrestata, la nave sequestrata ed i clandestini usati per il vostro sporco giochino, sparsi come buste di spazzatura sulle strade. Aprite le vs porte di casa invece che cavarvela con 15 euro a testa. Maledetti.*⁶

⁶ @user All for nothing, she will be arrested, the ship seized, and irregular migrants used for your foul game, scattered as trashbags on the streets. Instead of getting by 15 euros each, open your homes. You, damn.

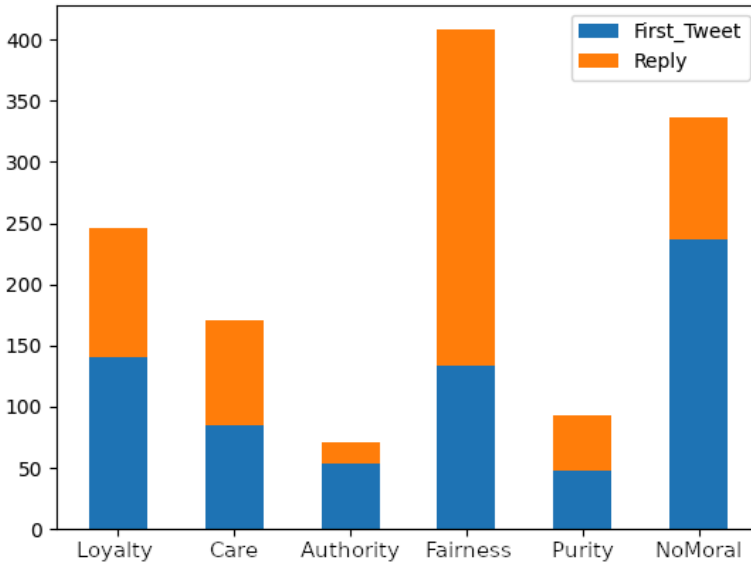


Figure 2

The distribution of moral concerns labels in the corpus (in tweets and in replies).

With 151 occurrences, **Care/Harm** is the third most prevalent moral concern in the corpus. Both when in the original tweet or in a reply, this value is almost always a pro-immigration stance signal as opposed to Fairness/Cheating or Loyalty/Betrayal. However, rare but interesting is the use of this moral concern to justify migrants rejections as a way to save their lives from human traffickers, as can be seen in (4).

(4) *@user i morti si moltiplicano per colpa di chi incoraggia il traffico di clandestini?*

In order to further understand whether there are linguistic clues signaling the correlation between the immigration topic and the three moral foundations more often annotated in the corpus, we calculated the **weirdness index** (Ahmad, Gillam, and Tostevin 1999; Florio et al. 2020), a technique that allows the retrieval of the most frequent and characterizing words within a specialized corpus of texts by contrasting it with a more general purpose dataset.

First, we calculated the relative frequency of each word in messages labeled with a given moral value, then we applied the same technique on the rest of the corpus. Finally, we computed the ratio between the two frequencies. This returned all the tokens that are frequent in messages annotated with a specific moral value, and occur less in other tweets of the corpus. In Table 5 a selection of the most specific words is listed.

Without forgetting the effect that the limited size of the corpus can have on the

⁷ @user deaths are multiplying due to people who encourage human traffick

validity of the index, some interesting signals can be drawn from this quantitative analysis. As expected, the words *reato*, *cattivo contribuente*, *corruzione* seem to correlate with the Fairness/Cheating domain, as well as *connazionali*, *sicure*, *tolleranzazero* with Loyalty/Betrayal, and *valori*, *restiamoumani*, *shoah* with Care/Harm.

We think that future work based on a larger dataset can provide the further and necessary evidence to these results, confirming the relationships between textual expressions and moral concerns.

Table 5

The most relevant words of MoralConvITA according to the weirdness-index calculation.

Fairness/Cheating	Loyalty/Betrayal	Care/Harm
reato	vivono	riace
fornero	bar	valori
dimaio	crimine	restiamoumani
cattivo	cambiato	raccontare
contribuente	stabile	shoah
inps	deliranti	passato
speso	connazionali	sanremo2019
corruzione	sicure	mediterraneo
stupro	buonisti	ponte
redditodicittadinanza	tolleranzazero	emigranti

4.2 Moral Foundations in Twitter Conversations

The application of the MFT framework for analyzing Twitter conversations resulted in the introduction of two additional dimensions to the annotation schema. The Concern Focus supports a more thorough investigation of how a message expresses the position of a user about a given moral foundation; the Conversational Relation allows to explore the conversational dynamic within the adjacency pair.

Concern Focus. For each tweet in the corpus expressing a moral dyad, the Concern Focus was annotated by choosing among the ‘prescriptive’ or ‘prohibiting’ label. Examples (5) and (6) were both annotated with the authority/subversion dyad, but the first with a ‘prescriptive’ focus, since it highlights respect for the law, while the second with a ‘prohibitive’ focus, as it is a critique to the government.

(5) *Chiunque sfrutta l’immigrazione clandestina per riempirsi le tasche va PUNITO in maniera esemplare, senza se e senza ma.*

*Complimenti a Carabinieri e Guardia di Finanza per l’operazione. Anche per gestori e cooperative in malafede, è finita la pacchia!*⁸

⁸ Anyone who uses illegal immigration to line their own pockets should be PUNISHED in an exemplary manner, no ifs or buts.

Congratulations to Carabinieri and Guardia di Finanza for the operation. Even for managers and cooperatives in bad faith, the free ride is over!

(6) *A casa fanno la voce grossa e mostrano i muscoli con i disperati. A Bruxelles invece Salvini e company sono solo pecorelle di #Orban che è il primo nemico dell'Italia e che nega ogni giorno i nostri valori costituzionali.*⁹

Unlike existing resources, which provide data-driven support for studying the psychological aspects of morality, Moral ConvITA mainly focuses on how this phenomenon is expressed in texts. The approaches are complementary and may lead to different interpretations of a message. For instance, (7) can be interpreted as a violation of the Loyalty/Betrayal principle since it highlights a conflict between Christians and Muslims. Conversely, in our corpus the tweet was annotated as conveying the Care/Harm dyad with a prescriptive focus, because the violation of the principle of care is not only present but also suggested, as it happens in many examples of HS. Similarly, (8) expresses Loyalty/Betrayal with a prohibitive focus. However, instead of being a stigmatization of somebody betraying her/his group, it reports the mediatic emphasis on crimes committed by migrants.

(7) - @matteosalvinimi *Pena di morte per i musulmani, TUTTI.*¹⁰

(8) @user *Aspetta che sia un immigrato preferibilmente di colore ad ammazzare la prossima donna e si scatena il #Capitonedastiera. L'omicida è un italiano? Quattro righe in cronaca, taglio basso e via la notizia dopo il primo lancio. Funziona così...*¹¹

The distribution of the focus is generally skewed on prohibition. According to the annotation, only 273 out of 1,179 focuses on the moral rule observance, which corresponds to 23%. The disproportion is more accentuated in replies, among which 81% of messages dwells on the violation of a moral rule.

The distribution differs when the intersection of the focus and the moral dyad is considered. While 86% of messages expressing Fairness/Cheating is also prohibitive (91% in replies), the annotated focus for the Care/Harm dyad is balanced. Finally, the presence of a prohibitive focus together with Loyalty/Betrayal values occurs in 76% of data, on average with the overall distribution, even if it is more rare in replies (71%). A deeper analysis should be performed in order to understand whether these numbers are the product of the topic, the contextual constraints of the social media where the conversations take place, or both. Moreover, a fine-grained annotation schema for this dimension is needed to capture a richer set of morally oriented communication functions.

Conversational Relation. The relation between two tweets in an adjacency pair could be either annotated as 'attack', 'support', 'continue' or 'no relation' (see Table 1). This dimension supports the analysis of the acceptance or rejection of messages expressing moral values (Section 2).

The quantitative analysis of these conversational patterns show a high prevalence of rejections to the original statement. 378 out of 786 in-agreement conversational pattern

9 At home they make a show of force and flex their muscles with desperate people. In Bruxelles, instead, Salvini and company are sheeps of #Orban, who he is Italy's first enemy and who negates constitutional rights every day.

10 - @matteosalvinimi Death penalty for Muslims, ALL.

11 @user Wait for a preferably black immigrant to kill the next woman and unleashes the #Snakefromthekeyboard. The murderer is an Italian? Four lines in the news, low profile and off the news after the first launch. It works like this...

was marked as an attack to the first element of the pair, while labels ‘support’, and ‘same topic’ collected together 360 annotations. The disproportion may also be larger because 33% of first pair elements in the corpus are replies themselves. Hence the adjacency pair may consist of two rejection responses to an original tweet which was not collected in the corpus (9).

(9) - @user1 @user2 *Impressionante superficialità. Più che Ministro...uno sceriffo. Caspita che cambiamento.* - @user3 @user4 @user5 *È diventato il #ministrodellimmigrazione. Altro non gli interessa...vedi #camorra #Ndrangheta #sacracoronaunita etc etc...*¹²

When the conversational relation and the moral dyad expressed by a reply are considered together, the number of adjacency pairs that can be usefully exploited for our analysis is reduced from 862 to 468, due to the low inter-annotator agreement (Table 6). In this subset the number of attacks increases by 8%, while the percentage of supporting replies is stable. As for the analysis of the concern focus, the distribution of conversational relations differs according to the moral dyad. More specifically, there are less messages annotated as expressing Loyalty/Betrayal and an attack at the same time.

The joint presence of a moral dyad in the first element of the pair and the conversational relation leads to a more important reduction of adjacency pairs that can be analyzed, since they are reduced to 417. In this subset it is worth mentioning the 78% of first elements expressing Loyalty/Betrayal and being attacked, that is more than 30% above messages conveying Care/Harm or Fairness/Cheating foundations.

Table 6

The joint distribution of moral dyads and conversational relations in the corpus.

	Care	Fairness	Authority	Loyalty	Purity	Total
1st tweet & attack	29	54	34	103	22	242
1st tweet & support	16	33	9	13	12	83
1st tweet & continue	27	33	8	15	9	92
reply & attack	43	144	5	44	28	264
reply & support	16	40	5	12	6	79
reply & continue	17	63	4	35	6	125

Questions as forms of Moral Rejection. Prosodic cues seem also to correlate with the presence of an attack in replies: 205 out of 378 rejection messages contain indeed a questions, while in 360 supporting responses there are only 138 questions.

Many of them appear to convey ironic statements, such as *quindi adesso i migranti possono*

¹² - @user1 @user2 *Impressive superficiality. More than Minister... a sheriff. Wow that change.*- @user3 @user4 @user5 *He has become the #MinisterofImmigration. He doesn't care about anything else...see #camorra #Ndrangheta #sacracoronaunita etc etc...*

*affogare in pace senza che nessuno li soccorra?*¹³. Others may be considered pragmatic rejections (Schlöder and Fernández 2015), namely utterances whose interpretation relies on information to be drawn from the context. For instance, the foundation expressed by the question *@giorgiameloni difendere da chi?*¹⁴ is recognizable only along with the exhortation in the first element of the pair: *Avanti insieme per difendere l'Italia!*¹⁵. Hence, the question conveys a stigmatization of the Loyalty/Betrayal dyad.

The interpretation of some message can be more problematic, like for instance *matteosalvinimi user con 49 milioni di euro sai quanti migranti ospito, matteo?*¹⁶, since external knowledge is needed to infer the Fairness/Cheating dyad from this question.

Finally, the detection of moral values expressed in a question may be supported by dialogical repetition (Bazzanella 2017). In *a lei il passato cosa ha insegnato?*, the repetition of the word 'passato/past' from the first message of the pair - *Il Governo sostiene tutte le iniziative in memoria della #Shoah, perché il passato ci insegna a combattere ogni forma di discriminazione e di odio*¹⁷ - is a cue of rejection. The first element of the pair, focused on the Care/Harm foundation, is challenged by a reply expressing Fairness/Cheating, since it seems to highlight the interlocutor's inconsistency.

The analysis of MFT in Twitter conversations shows some promising results. Considering the Concern Focus a separated dimension from foundations brought out a richer taxonomy of moral expressions that may be useful in understanding how specific moral stances interact with the spreading of toxic contents, as it emerges in the example (7). The conversational relation in adjacency pairs, especially when jointly investigated with dyads, appeared to show that some foundation are most likely to be rejected by the interlocutor, while others are more adopted to communicate disagreement. A preliminary analysis of questions as device for conveying a moral conflict emphasised the need of providing a fine-grained analysis of dyads are shaped within the conversation.

5. Conclusion and Future Work

This paper describes a novel Italian resource which is a collection of micro-conversations drawn from Twitter (adjacency pairs of messages, i.e. a tweet and its reply) and annotated for making explicit the occurrence of moral values and the conversational dynamics. The annotation scheme includes indeed moral concerns as categorized within the Moral Foundations Theory, the focus of each of the annotated moral concern and the relation that links the reply to a tweet in the conversation. As far the topic on which the corpus is focused, we selected a discourse domain related to an issue that we know as especially relevant to moral values and a topic with sufficient popularity among Twitter users, i.e. immigrants.

The main aim of making available this resource to the computational linguistics research community is at providing a missing dataset for Italian and at discussing some currently underrepresented phenomena that collocate at the intersection of social psychology, linguistics and conversational analysis.

13 So now migrants can drown in peace without anyone helping them?

14 *@giorgiameloni*, defend from whom?

15 Forward together to defend Italy!

16 *@matteosalvinimi @user* with 49 million euros do you know how many migrants I host, matteo?

17 What has the past taught you?

17 The Government supports all the initiatives in memory of the #Shoah, so that the past teaches us to fight all forms of discrimination and hatred

Nevertheless, considering that the expressions of moral sentiment in one domain and about a specific topic hardly generalize to data extracted from another domain, in future work we want to address other domains, e.g., misogyny, by collecting more data and by testing on them the scheme we propose in this paper.

References

- Ahmad, Khurshid, Lee Gillam, and Lena Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, MD, US, January. National Institute of Standards and Technology (NIST).
- Allen, James and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers.
- Basile, Pierpaolo and Nicole Novielli. 2018. Overview of the Evalita 2018 itaLIan Speech acT labEliNg (iLISTEN) Task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.
- Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 54–63, Minneapolis, Minnesota, US, June. "Association for Computational Linguistics".
- Basile, Valerio, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the University of Turin. In *5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–6, Turin, Italy, December. CEUR-WS.
- Bazzanella, Carla. 2017. Dialogic repetition. In *Dialoganalyse IV, Teil 1*. Max Niemeyer Verlag, pages 285–294.
- Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an ISO standard for dialogue act annotation. In *11th International conference on Language Resources and Evaluation (LREC 2010)*, pages 1787–1794, Valletta, Malta, May. European Language Resources Association (ELRA).
- Bunt, Harry, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R. Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 430–437, Istanbul, Turkey, May. European Language Resources Association (ELRA).
- Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July-August.
- Core, Mark G. and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35, Boston, MA, US, November.
- Fanton, Margherita, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online, August. Association for Computational Linguistics.
- Fiske, Alan P. 1991. *Structures of social life. The four elementary forms of human relations: communal sharing, authority ranking, equality matching, market pricing*. Free Press.
- Florio, Komal, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).
- Fortuna, Paula and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.
- Graham, Jesse and Jonathan Haidt. 2012. The moral foundations dictionary.
- Graham, Jesse, Jonathan Haidt, Matt Motyl, Sena Koleva, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47:55–130.

- Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.
- Haidt, Jonathan. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.
- Hoover, Joe, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.
- Hoover, Joseph, Mohammad Atari, Aida M Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, Drew Kogon, and Morteza Dehghani. 2019. Bound in hatred: The role of group-based morality in acts of hate, Jul. 10.31234/osf.io/359me.
- Hopp, Frederic R., Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2020. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, pages 1–15.
- Hulpus, Ioana, Jonathan Kobbe, Heiner Stuckenschmidt, and Graeme Hirst. 2020. Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80, Barcelona, Spain (Online), September. Association for Computational Linguistics (ACL).
- Johnson, Kristen and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia, July. Association for Computational Linguistics (ACL).
- Jurafsky, Daniel. 2004. Chapter 26: Pragmatics and computational linguistics. In *The handbook of pragmatics*. Wiley Online Library, pages 578–604.
- Kalimeri, Kyriaki, Mariano G. Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, 92:428–445, March.
- Lai, Mirko, Marco Antonio Stranisci, Cristina Bosco, Rossana Damiano, and Viviana Patti. 2021. HaMor at the Profiling Hate Speech Spreaders on Twitter. In *Notebook for PAN at CLEF 2021*, volume 2936 of *CEUR Workshop Proceedings*, Online, September. CEUR-WS.
- Lai, Mirko, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter. *Data Knowledge Engineering*, 124.
- Levinson, Stephen C. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Mathew, Binny, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv preprint arXiv:1812.02712*.
- Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.
- Sanguinetti, Manuela, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2@evalita2020: Overview of the evalita 2020 hate speech detection task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR.org.
- Sanguinetti, Manuela, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Schegloff, Emanuel A. and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.
- Schlöder, Julian J. and Raquel Fernández. 2015. Pragmatic rejection. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 250–260, London, UK, April. Association for Computational Linguistics (ACL).

- Schmidt, Anna and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics (ACL).
- Shweder, Richard A., Nancy C. Much, Manamohan Mahapatra, and Lawrence Park. 1997. The "big three" of morality (autonomy, community and divinity), and the "big three" explanations of suffering. In A. Brandt and P. Rozin, editors, *Morality and health*. Routledge, pages 119–169.
- Simpson, James. 2005. Conversational floors in synchronous text-based CMC discourse. *Discourse studies*, 7(3):337–361.
- Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.
- Vilella, Salvatore, Mirko Lai, Daniela Paolotti, and Giancarlo Ruffo. 2020. Immigration as a Divisive Topic: Clusters and Content Diffusion in the Italian Twitter Debate. *Future Internet*, 12(10):173.
- Waseem, Zeerak and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, San Diego, California, US, June. Association for Computational Linguistics (ACL).
- Weber, Christopher R. and Christopher M. Federico. 2013. Moral foundations and heterogeneity in ideological preferences. *Political Psychology*, 34(1):107–126.
- Winterich, Karen Page, Yinlong Zhang, and Vikas Mittal. 2012. How political identity and charity positioning increase donations: Insights from moral foundations theory. *International Journal of Research in Marketing*, 29(4):346–354.

Computational Grounding: An Overview of Common Ground Applications in Conversational Agents

Maria Di Maro*

Università di Napoli 'Federico II'

This work reports on the literature on grounding in conversational agents, as one of the pragmatic aspects adopted to ensure a better communicative efficiency in dialogue systems. The paper starts with a general description of the theory of grounding. As far as its computational implications are concerned, grounding phenomena are firstly framed in the common grounding processes described in terms of grounding acts. Secondly, they are considered in the argumentation-related framework within which already grounded information are processed. Open issues and application gaps are finally highlighted.

1. Introduction

In Stalnaker's words, "when speakers speak, they presuppose things and what they presuppose guides both what they choose to say and how they intend what they say to be interpreted. To presuppose something means to take it for granted as background information – as common ground among the participants during their conversation" (Stalnaker 2002, p. 701). In fact, communication is a joint activity in which two speakers must share information or, in other words, they must have a common ground, i.e., mutual knowledge, mutual beliefs, and mutual assumptions, as the foundation for mutual understanding (Clark and Brennan 1991). To coordinate on this process, speakers need to update, check, or revise their common ground with a process that constantly evolves through time. The importance of focusing on this communicative process reflects the need to bridge the gap left in the study and development of dialogue systems caused by the lack of insights into the application of pragmatics to conversational agents. Although pragmatics is very important in dialogue, as it is one of the aspects governing interpretation, understanding, and efficiency, its computational application is mainly focused on the study and identification of speech acts (Leech 2003). Furthermore, in the last ten years, semantics has been a more investigated topic within the dialogue systems field with respect to pragmatics, especially as far as the understanding of the correct recognition of the received intent was concerned, as shown in the publications on dialogue systems (Figure 1).

On the other hand, as far as pragmatics is concerned, in the last ten years, the research on Common Ground has started to see a thriving impulse (Figure 2). Nevertheless, a more in-depth analysis of pragmatic phenomena, such as Clarification Requests, related to Common Ground construction and consistency checks in human-machine interaction appears to be a missing spot in the research on dialogue systems.

* Interdepartmental Center for Advances in Robotic Surgery. Department of Electrical Engineering and Information Technology. University of Naples "Federico II", Italy. E-mail: maria.dimaro2@unina.it

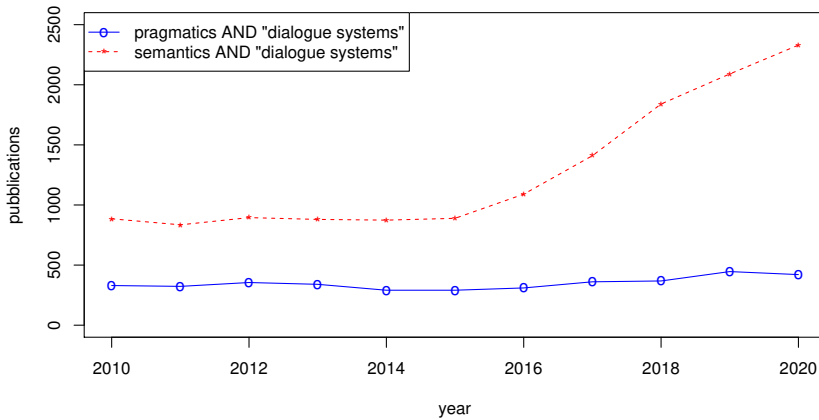


Figure 1
 Number of Google Scholar’s results on publications about dialogue systems applying semantics versus pragmatics from 2010 to 2020 [Retrieved on 30/04/2021].

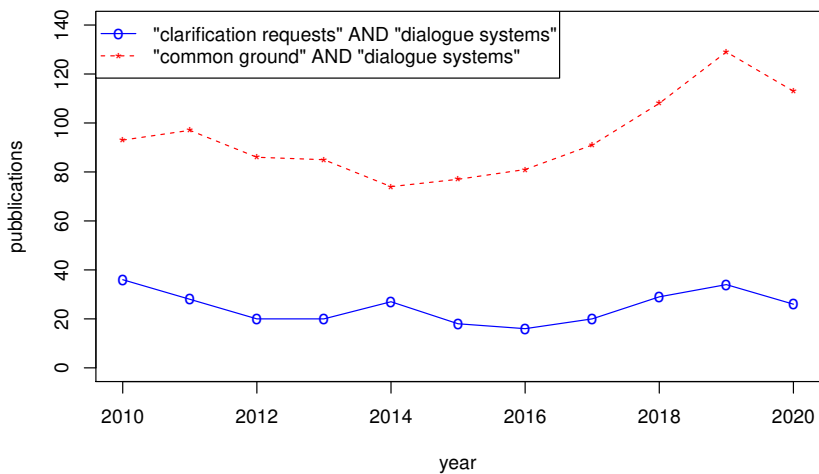


Figure 2
 Number of Google Scholar’s results on publications about clarification requests and common ground from 2010 to 2020 [Retrieved on 30/04/2021].

Different scholars (Bousquet-Vernhettes, Privat, and Vigouroux 2003; Beun and van Eijk 2004; Purver 2004a; Roque and Traum 2009; Hough, Zarri , and Schlangen 2017; M ller, Paul, and Li 2021) highlighted the urge of including pragmatic aspects in their systems to improve the communication process. This need resulted from the users’ need to interact with an agent capable of cooperating on the communicative actions.

This survey aims both at presenting a literature review on grounding theories and their application in dialogue systems, and at pointing out pragmatic aspects which still need to find a computational model. The paper is organised as follows: in the next section the theory of grounding is summarised; in section 2.1, the grounding acts reported in (Traum 1994) are explained; in section 3, computational applications of the theory are reported starting from the aforementioned grounding acts to more general works; finally, open issues concerned with the processing of grounding information are presented.

2. Grounding the Grounding Process

Stalnaker (2002) defined the notion of *common ground* as the sum of interlocutors' mutual, common, or joint beliefs and knowledge. Since Grice (1975), the importance of cooperation in a successful conversation was pointed out. In Grice (1989, p. 65), the term of *common ground* was introduced as related to communicative processes. In fact, participants in a conversation must have grounded knowledge in order to understand each other. Common ground, as Clark (2015) acknowledged, can be of four main types: personal, local, communal and specialised. *Personal Common Ground* is established collecting information over time through communicative exchanges with an interlocutor and it can be considered as a record of shared experiences with that person. A part of Personal Common Ground is *Local Common Ground* that is tied to a piece of information obtained from a single exchange with an unknown or known interlocutor. According to Clark (2015), information of this type can be, for instance, the opening hours for a shop, train timetables, and so on. *Communal Common Ground* refers to the amount of information shared with people belonging to the same community, that is to say, people that share general knowledge, knowledge about social background, education (schools attended, levels of education attained), religion, nationality, and language(s). Within a larger community, a smaller one can be found: *Specialised Common Ground* pertains to those people that share particular areas of expertise about some domain of knowledge, such as colleagues, friends, or acquaintances. It is marked by specialised vocabulary of that specific domain, such as medicine, law, and so on.

The process of *grounding* takes place in dialogue when the interlocutors update their common ground by accumulating information in the perceived common ground. In Clark and Schaefer (1989), the classical model of grounding is illustrated: dialogue participants reach their mutual belief by checking the mutual understanding. This is accomplished through *contributions*, that is the communicative actions collected through dialogue. Contributions can be divided into presentation phase and acceptance phase. During the presentation phase, the utterance is presented, whereas in the acceptance phase, the utterance is accepted by the interlocutor as understood. The utterance acceptance or refusal is signalled via diverse types of feedback. The refusal, for instance, can depend on different aspects, such as acoustic, semantic or intentional misunderstanding. According to Allwood et al. (1992, p. 4-5), feedback is indeed a linguistic mechanism which enables interlocutors to exchange information about four different basic communicative functions: i) contact (i.e., feedback expressing the will and/or ability to continue the interaction); ii) perception (i.e., feedback referring to the will and/or ability to perceive the message); iii) understanding (i.e., feedback about the will and/or ability to understand the message); iv) attitudinal reactions (i.e., feedback referring to the will and/or ability to react and respond appropriately). According to Clark and Brennan (1991), the first main form of positive evidence for acceptance are the acts of *acknowledgement* (the complete classification of *grounding acts* is detailed in

Section 2.1), in particular: i) *back-channel responses* that include continuers such as *uh*, *huh* or *yeah*, used to signal that the utterance has been understood and that there is no need to initiate a repair in the next turn; ii) *assessments* (i.e., *gosh*, *really*) that are usually produced without taking the turn. A second form of positive evidence is the initiation of the *relevant next turn*: suppose *A* is trying to ask *B* a question; if *B* understands it, the answer will be expected in the next turn. Questions and answers constitute adjacency pairs. In other words, once the first part of the adjacency pair is uttered, the second part is considered as conditionally relevant for the next turn. The third and most basic form of positive evidence is *continued attention* provided by an attentive listener. In conversation, people monitor their partner from time to time and immediately adapt to their feedback. If *A* utters something and notices that *B* was not paying attention, *A* could assume that *B* did not understand him. *B* must show that he is paying attention through different social signals, like eye gaze or other communicative feedback. *A* can, therefore, use phatic expressions (i.e., *Are you listening?*, *You know what I mean?*) to understand if *B* is following, or she can elicit attentive listener feedback in *B*. On the other hand, *B* could also want to show his attention by using communicative feedback. Positive evidence of understanding, thus, is provided by communicative feedback and comes with attention that is unbroken or undisturbed (Buschmeier 2018; Buschmeier and Kopp 2018). Furthermore, according to (Clark 1996, p. 147-148), these actions are processed following the concepts of *upward completion*, i.e., *in a ladder of actions, it is only possible to complete actions from the bottom level u through any level in the ladder*, and *downward evidence*, i.e., *in a ladder of actions, evidence that one level is complete is also evidence that all levels below it are complete*.

As argued by Clark and Schaefer (1989), the strength of evidence that *B* has understood *A* can depend on several factors, including the complexity of the presentation, the importance of its understanding, and the closeness among the participants. Moreover, since the acceptance phase can be recursive, as *B*'s acceptance to *A*'s presentation needs to be accepted as well, in Traum (1999) the *Strength of Evidence Principle*, introduced in Clark and Schaefer (1989, p. 268), is instead preferred to avoid recursion. This principle states that "*The participants expect that, if evidence e_0 is needed for accepting presentation u_0 , and e_1 for accepting presentation of e_0 , then e_1 will be weaker than e_0* " (Traum 1999, p. 2). In other words, the evidence is stronger when the need for acceptance is higher. The authors exemplified the principle as follows: *A* presents a book identification number, *f*, *six*, *two*, *B* accepts it by displaying it verbatim *f*, *six*, *two*; then *A* accepts the *B*'s acceptance by using a weaker evidence like *yes*. Lastly, *B* accepts the *A* evidence by proceeding to the next contribution. The traditional version of this principle exhorts speakers not to expend any more effort than they need to get their addressees to understand them with as little effort. Grice (1975) used two maxims of the cooperative principle to account for the communicative effort: according to the maxim of *quantity*, the speaker must not make their contribution more informative than is required, and, according to the maxim of *manner*, they must also be brief and avoid prolixity. In detail, the general principle of least collaborative effort introduced by Clark and Wilkes-Gibbs (1986) was used by the authors to criticise the general speaker economy principle (Brown 1958) which does not always represent the right strategy for grounding. As claimed by Clark and Wilkes-Gibbs (1986), there are three main problems with this principle: i) *time pressure*, speakers tend to limit the effort for planning an utterance which could result in incorrect productions; ii) *errors* that a speakers can make during speaking that need to be repaired; iii) *ignorance* of the personal knowledge and beliefs of the interlocutor can cause improper utterances. Instead, the authors focus on the *minimisation* of the collaborative effort, as "*speakers and addressees try to minimise collaborative effort, the work*

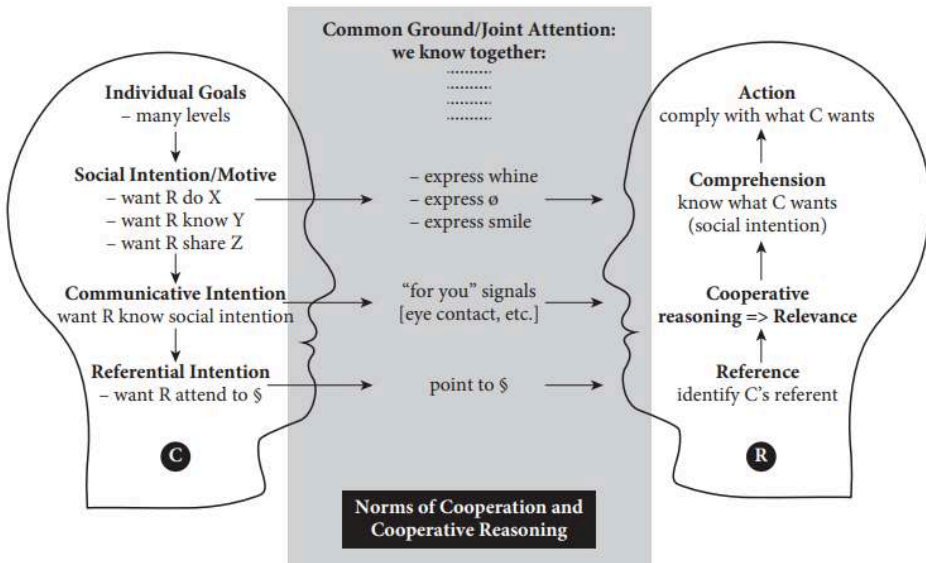


Figure 3
 Summary of cooperative model of human communication (C = communicator; R = recipient);
 Source: Tomasello (2010); All rights belong to their respective owners.

both speakers and addressees do from the initiation of the referential process to its completion” (Clark and Wilkes-Gibbs 1986, p. 26).

From a more cognitive point of view, grounding, referred to as an explicit signal of cooperation in dialogue, is also represented in the cooperation model of communication reported by Tomasello (2010) (Figure 2): the communicator *C* has individual goals, such as goals and values pursued in their life. If for any reason, *C* feels that the recipient *R* can be of any help in the achievement of some goals, *C* will produce specific acts which will bring *R* to do something, know something, or share something. This is represented by *C*'s social intention, which is expressed through communication. Therefore, a communication act (verbal or not verbal) is mutually manifested in the joint attentional frame. *C*'s communicative intention is consequently shared. *C* can also draw *R*'s attention to some referential situation in the external world (referential intention) designed to lead *R* to infer social intentions via processes of cooperative reasoning (Huang 2017, p. 282). On the other hand, *R* attempt to firstly identify the referent, typically within the space of the common ground, and secondly to infer the social intention, also by relating it to the common ground. Then, assuming that *R* understands *C*'s social intention, *R* can decide whether or not to cooperate as expected (Tomasello 2010; Huang 2017).

Whereas the cognitive and linguistic aspects of grounding are naturally clear, its computational applications can be prone to diverse difficulties. Pragmatics can sometimes be subjective, contextual, ambiguous, and its phenomena can be described through one-to-many and many-to-many relationships. Their computational modelling is, therefore, challenging, although different scholars worked on some aspects as it will be summarised in this work. In the next sections, we will focus on grounding

Table 1

Conversation Act Types (Traum 1999, *adapted*); *UU* and *DU* stand respectively for ‘utterance unit’ and ‘discourse unit’.

Discourse Level	Act Type	Sample Acts
Sub UU	Turn-taking	take-turn, keep-turn, release-turn, assign-turn
UU	Grounding	Initiate, Continue, Acknowledgement, Repair, Cancel, RequestRepair, RequestAcknowledgement
DU	Core Speech Acts	Inform, YesNoQuestions, Check, Evaluate, Suggest, Request, Accept, Reject
Multiple DUs	Argumentation	Elaborate, Summarise, Clarify, Q&A, Convince, Find-Plan

acts, as they were described in (Traum 1994), and how they could be mapped on research approaches described by different scholars. This work is, therefore, intended as a schematic literature review on some aspects of grounding that can function as a guide and lead to new studies, as research gaps are also highlighted.

2.1 Grounding acts

Traum (1994) provided a computational model of grounding. In his theory, he introduced a description of the so-called *grounding acts*, which are speech acts used to ground the traditional illocutionary speech acts (Austin 1975; Searle 1985). In other words, they correspond to “*the actions performed in producing particular utterances which contribute to this groundedness*” (Traum 1994, p. 31). In particular, he accounted for the protocol determining, for any sequence of grounding acts, whether the content of the communicated utterances is grounded or not. In table 1, its conversation acts are presented, among which the grounding acts are listed.

Each of the grounding acts considered is described as follows:

Initiate. This act is the initial utterance of a discourse unit and usually corresponds to the first utterance of the presentation phase (Clark and Schaefer 1989).

Continue. This represents the continuation of a previous act performed by the same speaker. A continue is expressed in a separate utterance unit, but is syntactically and conceptually part of the same act.

Acknowledgement. An act of acknowledgement is used to claim or demonstrate understanding of a previous utterance. It may be either a repetition or paraphrase of all or part of the utterance, an explicit signal of comprehension such as *ok* or *uh huh*, or an implicit signalling of understanding. Typical cases of implicit acknowledgement are answers to questions. Acknowledgements are also referred to by some as confirmations (Cohen and Levesque 1991) or acceptances (Clark and Schaefer 1989). Traum (1994) prefers the term ‘acknowledgement’ as a signal of understanding, whereas ‘acceptance’ is referred to a core speech act signalling agreement with a proposed domain plan.

Repair. A repair is used to change the content of the current discourse unit. This may correspond either to a correction, or it can concern the addition of material. Both solutions will change the interpretation of the speaker's intention. Repair actions should not be confused with domain clarifications. Repairs are concerned merely with the grounding of content. On the other hand, domain clarifications, which modify grounded content, are considered as argumentation acts (Traum 1994). As we will see in the next sections, this particular act that processes grounded information can have interesting computational applications.

Cancel. This act closes the current discourse unit as ungrounded. Rather than repairing the current unit, a Cancel leaves it; the speaker intention must, therefore, be possibly expressed in a new discourse unit.

RequestRepair. A request for a repair is, conversely, uttered by the interlocutor. This is equivalent to a next turn repair initiator or clarification request (Schegloff, Jefferson, and Sacks 1977). Often a RequestRepair can be distinguished from a Repair or Acknowledgement only by intonation. Implicit requests have also been studied (Schettino, Di Maro, and Cutugno 2020).

RequestAcknowledgment. The act is used as an attempt to elicit an Acknowledgement act in the other agent. This invokes a discourse obligation on the listener to respond with either the requested acknowledgement, or an explicit refusal or postponement (i.e., a followup repair or a repair request).

Starting from the description of grounding acts, in the next section, we will explore the studies that concentrated on their computational modelling, or of some of their aspects, in dialogue systems.

3. Computational Grounding

This section reports on pragmatics applied to dialogue modelling and automatic text processing. This branch of computational pragmatics, especially when applied to conversational agents, mostly deals with corpus data, context models, and algorithms for context-dependent utterance generation and interpretation (Huang 2017, p. 326). Nevertheless, conversational agents should be able not only to process local but also global structures of dialogues (Airenti, Bara, and Colombetti 1993). Whereas local structures are involved with linguistic rules (i.e., speech acts, turn-taking, etc.), which can be derived from corpus analysis, global structures refer to the conversation flow, that is the dialogue's action plan and how this is mutually known by dialogue participants (i.e., opening, closing, etc.). Cognitive pragmatics looks at these global structures derived from behavioural games, which in turn derive from grounding processes (Bara 1999). Different authors started including these processes in their dialogue systems architectures, especially as far as evaluating and updating common ground with their human partner. For instance, Roque and Traum (2009) have developed a dialogue system that tracks grounded information in the previous conversation. As a consequence, the dialogue system is capable of selecting its utterances using different types of evidence of the user's understanding (i.e., whether the dialogue system has just submitted material or the user has also acknowledged it, repeated it back, or even used it in a subsequent utterance) (Müller, Paul, and Li 2021).

Table 2
Computational Grounding acts state of the art

Grounding Act	References
Initiate	(Dahlbäck and Jönsson 1998)
Continue	(Schlangen and Skantze 2011) (Visser et al. 2012) (Visser et al. 2014)
Acknowledgement	(Skantze, House, and Edlund 2006) (Wang, Lee, and Marsella 2013) (Visser et al. 2012, 2014) (Eshghi et al. 2015) (Buschmeier 2018) (Buschmeier and Kopp 2018) (Schlangen 2019)
Repair	(Skantze 2008) (Swerts, Litman, and Hirschberg 2000) (Hough and Purver 2012) (Marge and Rudnicky 2015) (Purver, Hough, and Howes 2018) (Di Maro et al. 2019) (Marge and Rudnicky 2019)
Cancel	N/A
RequestRepair	(Gabsdil 2003) (Rodríguez and Schlangen 2004) (Purver 2004a) (Schlangen 2004) (Purver 2006) (Stoyanchev, Liu, and Hirschberg 2014) (Müller, Paul, and Li 2021)
RequestAcknowledgement	(Misu et al. 2011) (Buschmeier and Kopp 2014)

Using grounding strategies in conversational agents led to interesting implementations. One aspect which has not yet been investigated is concerned with the mechanisms of grounding between humans and dialogue systems. Experimental investigations have mostly studied “*how users evaluate the interaction, instead of studying interaction mechanisms*” (Müller, Paul, and Li 2021, p. 3). For instance, Roque and Traum (2009) performed a user study in which subjects interacted with their system and rated how much they felt the system understood them, put effort into understanding them, and gave appropriate responses. Conversely, what most studies do not ask is how a specific dialogue principle, such as the use of a particular type of request, is used by a system to affect user behaviours. Therefore, to learn more about human–machine dialogues mechanisms, it is important to turn to more basic experimental research methods (Müller, Paul, and Li 2021).

With the purpose of providing a structured view concerning the application of grounding in dialogue systems, we start with the classification presented in Traum (1994), and summarised in section 2.1, as a point of departure to understand which aspects of grounding has been modelled over time. As we will see, some of them are more investigated than others, while new other aspects have been considered. In table 2, the studies in which grounding acts are modelled are reported.

Initiate. In the LINLIN dialogue model (Dahlbäck and Jönsson 1998), the initiative is defined as the move whose aim is to introduce a goal. It can have different functions: update, question, answer, discourse opening, discourse continuation, discourse ending. The initiation act in dialogue systems is described in terms of presentation phase, in which form it is presented and which function it shows. This act, as reported in Clark

and Schaefer (1989), introduces something that has to be grounded, via implicit and explicit feedback, to proceed with the exchange. Since this act can also, from other aspects, overlap with other type of grounding acts, such as acknowledgement, as also reported in (Clark and Schaefer 1989), specific details are given in this section, when the corresponding grounding acts are dealt with more in detail.

Continue. For the continue act, defined as the continuation of a previous act by the same speaker, we can account for the studies on incrementality in dialogue. Dialogue processing is, indeed, *incremental*: the processing starts before the input is completed (Kilger and Finkler 1995). Systems designed for incremental processing can process the user inputs, with or without intermediate feedback, before the system output is generated. Incrementality is, for this reason, a research aspect comparable to what has been studied for continue grounding acts. Here, the aspect of grounding is referred to the fact that the previous act is considered as understood and grounded, in that no repair is needed, and therefore the current speaker can go on with the contribution which it refers to. In (Schlangen and Skantze 2011), a model for incremental processing architecture is presented. In their model, this act corresponds to the *incremental unit*, which is the “minimal amount of characteristic input”. The incremental processing is composed of a left buffer, a processor, and a right buffer, as represented in Figure 4. The authors also point out for future application the necessity to connect such model for incremental processing and grounding of interpretations in previous processing with models of dialogue-level grounding in the information-state update tradition (Larsson and Traum 2000). For example, the study of self-correction could be a starting point in the connection of sub-utterance processing and discourse-level processing (Ginzburg, Fernández, and Schlangen 2007). Visser et al. (2012, 2014) define incremental understanding in terms of pairs of frames generated every 200 milliseconds, where the first frame is a prediction of the meaning of the complete user utterance, although not yet fully uttered, whereas the second frame is the sub-frame of what the user said so far. Here, feedback of different kinds are analysed before the completion of the utterance.

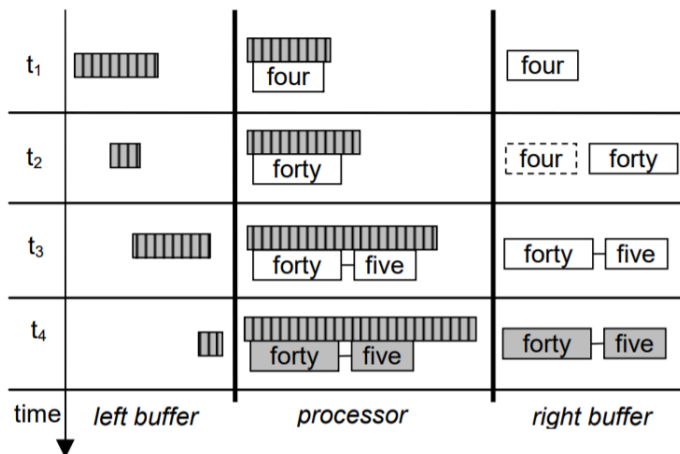


Figure 4 Speech recognition as an example of incremental processing; Source: Schlangen and Skantze (2011); All rights belong to their respective owners.

Acknowledgment. The use of acknowledgement feedback in human-machine interaction has been deeply investigated. In Skantze et al. (2006), for instance, investigate feedback produced both by the user and the system. These were used along with other types of feedback categorising the subjects' responses based on pragmatic meaning. In Wang et al. (2013), a Listener Feedback Model for virtual agents in multi-party conversations is presented, for which the importance of using such systems is underlined. The use of understanding feedback were also studied in incremental models as signals used to update the grounding state (Visser et al. 2012, 2014; Eshghi et al. 2015). In (Buschmeier and Kopp 2018; Buschmeier 2018), acknowledgement acts are studied as attentiveness markers: "*artificial conversational agents should have the capability to use such a mechanism, too, because it would allow them to approach potential or upcoming problems in understanding (and other listening related communicative functions) before they become more serious and require costly repair actions*" (Buschmeier and Kopp 2018, p. 1220). Acknowledgement acts are important for collaborative goals, as also pointed out in Schlangen (2019), and more generally also in Benotti and Blackburn (2021).

Repair. The repair act is aimed at grounding information which may not be clear to either the user or the system. Purver et al. (2018, p. 426), indeed, describe repair as one of "*the primary strategies by which interaction participants achieve and maintain shared understanding*". This set of strategies is specifically used to highlight and/or resolve miscommunications or potential miscommunications. In Di Maro et al. (2021), different miscommunication scenarios are listed. Starting from Allwood et al. (1992) four basic communicative functions, the communication levels *contact*, *perception*, *understanding*, *intention* were defined. At each level, one or many problems can occur, which are triggered by specific linguistic or informational issues. According to the type of problem, a specific repair strategy can be used to ensure the grounding process to be successful. For instance, in Marge and Rudnick (2015, 2019), recovery strategies were studied in three different scenarios: referential ambiguity (more than one possible action), impossible-to-execute (zero possible actions), and executable (one possible action). In Hough and Purver (2012, p. 143), repair acts are incrementally generated "*in line with psycholinguistic evidence of preference for locality and the availability of access to the semantics of repaired material*". Prosodic features of repairs are also investigated both from the perspective of users (Di Maro et al. 2019) and of machines (Swerts, Litman, and Hirschberg 2000).

Cancel. Among the investigated breakdown recovery strategies, the cancel act appears to be not so explored. In fact, although the speaker could leave the interaction without giving any further explanation or without trying to repair, thus without modifying their common ground, when the system does not understand them, it is unlikely to find studies which focus on how a system does not try to recover the interaction. This act, in fact, could be more interestingly investigated when its adoption is caused by the analysis of multiple discourse units. In this scenario, when the modification of the common ground interests a previous discourse unit, the repair could imply a higher cost or effort. As a consequence, the dialogue can go on without accepting the last utterance and a re-planning is therefore needed. A parallelism can be drawn, for example, with a car's satellite navigation system that prefers to recompute the route, when repairing the misunderstood action could be too difficult or impossible for the driver. Nevertheless, actions on multiple discourse units are studied in Traum (1994) as argumentation acts where negotiation is important. On the other hand, as far as the last action is concerned, the problem is usually not followed by a cancellation, but by a repair.

RequestRepair. This act is investigated in dialogue systems especially in terms of clarification requests. Clarification requests (CRs) are one of the pragmatic tools used in conversation to prove, ensure, and maintain the mutual understanding of the communicated message between the interlocutors. Purver (2004a, 2004b, 2006) stated that interlocutors initiate a CR when a problem in processing the previous utterance occurs. For this reason, they are also called *anaphoric feedback*, as they refer to what has previously been uttered. Furthermore, CRs are considered as meta-communicative tools as well, since they function as an acknowledgement of the level of understanding of an utterance (Ginzburg and Macura 2005). The use of CRs is also described in terms of cognitive-pragmatic instruments adopted for grounding purposes. As pointed out by scholars, such as Clark (1996), to pursue the aim of succeeding in their joint activity, interlocutors need to ground what has been communicated. Among the scholars who pursued the intent of categorising different types of CRs, Purver (2004b) classified CRs according to form and reading, where *form* refers to the surface form, such as when i) an element from the previous utterance is used in the request (*reprise*), ii) an element from the previous utterance is used in combination with a *wh*-interrogative pronoun (*wh-reprise*), or iii) when a reformulation or a generic question is adopted (*non-reprise*). *Reading*, conversely, refers to the compromised item that the request questions about, such as a constituent or a clause. This classification established a precise way to describe how CRs can be automatically recognised or selected by a system and opened the way to further investigations, also concerning the causes and problems triggering the initiation of such requests. For instance, Rodriguez and Schlangen (2004) introduce the notion of *problem*, causing the instantiation of a CR. In fact, different kinds of problems, such as acoustic or lexical ones, can determine the adoption of a different informative CR.

Clarification is then a fundamental part of the grounding process. Through the pragmatic tool of CRs the interlocutors can maintain the mutual understanding of the communicated message during a conversation. Clarifications are usually uttered in a context of miscommunication. Following Hirst et al. (1994), miscommunication can be partitioned into three different types: Misunderstanding, non-understanding, and misconception.

Misunderstandings are not immediately detected, since the hearer thinks that what has been understood is the right message, but it is not the one the speaker intended to convey.

The second type of miscommunication is non-understanding that occurs when the hearer finds the message uttered by the speaker ambiguous, or, as Gabsdil (2003) noticed, when the hearer is uncertain about the interpretation given to the message. In this case, even the form in which the requests are formulated can vary. Uncertain interpretations can coarsely be associated with single polar questions, whereas ambiguous understanding is more likely to result in alternative questions or *wh*-questions. Furthermore, non-understanding in general can occur on several different communicative levels, ranging from establishing contact among the dialogue partners to the intended meaning or function of the utterance in context, as previously also pointed out. Clark (1996) listed four basic levels of communication in a framework that represents the interaction as a joint activity of the dialogue participants: i) execution/attendance, ii) presentation/identification, iii) signal/recognition, iv) proposal/consideration. As Gabsdil (2003) pointed out, on the lowest level, dialogue participants establish a communication channel, which is then used to present and identify signals on level two. On level three, these signals are interpreted before their communicative function is evaluated on the proposal/consideration level. The framework of joint actions requires that dialogue participants coordinate their individual actions on all of those levels.

Gabsdil (2003) combined the cause of non-understanding with Clark's four levels of communication, giving some examples and organising a coarse-grained classification of clarifications. Connected to these levels, two main readings for clarifications were proposed by Ginzburg and Cooper (2001). Their "clausal reading" can be related to the presentation/identification level and their "constituent reading" to the signal/recognition level. Clausal readings are used "to confirm the content of a particular sub-utterance" (Ginzburg and Cooper 2001, p. 1), and it can roughly be paraphrased as "Are you asking/asserting that X?" or "For which X are you asking/asserting that Y?". Constituent readings, on the other hand, "elicit an alternative description or ostension to the content (referent or predicate etc.) intended by the original speaker of the reprised sub-utterance." (Ginzburg and Cooper 2001, p. 1).

Misconceptions, finally, occur when the "hearer's most likely interpretation suggests that beliefs about the world are unexpectedly out of alignment with the speaker's" (Hirst et al. 1994). Clarifications in response to misconceptions usually convey extra-linguistic information like surprise or astonishment.

As already anticipated, CRs can occur in different forms and readings. The correlation between form and function of CRs has also been investigated by Rodriguez and Schlangen (2004), who presented a multidimensional classification of CRs forms and a fine-grained correlation between them and their functions. The study has been carried out in a corpus of German task-oriented dialogues, the "Bielefeld Corpus"¹, which contains 22 dialogues consisting of about 3962 turns, and 36,000 words. In the experimental setup, a dialogue participant was supposed to give instructions to the interlocutor to build a model plane. The authors pointed out some features used to describe the surface form of CRs. Concerning the attribute *Mood*, the possible values are declarative, polar questions, alternative questions, wh-questions, imperative and others; for *Completeness* are particles (*Pardon?*), partial fragments or complete sentences; for *Relation* are literal repetition of the unclear part, the addition of a part to the repetition, reformulation of the problematic utterance, or independent (i.e., no part of the utterance are repeated or reformulated); finally, for *Boundary tone* are rising or falling intonation.

Rodriguez and Schlangen (2004) posed the foundation for the identification of problems that could cause misunderstanding, taking into account the CRs readings proposed earlier, but trying to define them in a more fine-grained way. The authors devised a multidimensional classification scheme where form and function are meta-features taking sub-features as attributes. They start from the models of Clark (1996) and Allwood (1995) concerning the four levels of communication mentioned before, adding other types of sub-levels. Each of those levels is a possible locus for communication problems. This dimension specifies the extent and severity of the problem. The extent, as the authors argued, describes whether a specific CR points to a problematic element in the problematic utterance or not. The severity, on the other hand, describes which action the CR initiator requests from the interlocutor: the CR initiator can ask for a reformulation of the problematic utterance, probably triggered by a complete understanding failure, or they can ask for a confirmation of the previous hypothesis of which they are not certain. The scholars also classified the answers to CRs that can be i) yes/no answers, ii) repetitions or reformulations of the unclear element, iii) elaborations of the problematic utterance with the addition of new elements, iv) word definitions, or, lastly, v) no reaction at all. As a consequence, the satisfaction of CR initiators to the reaction

1 <http://www.sfb360.uni-bielefeld.de/>

of the CR addressee can be classified as happy or unhappy, according to the right or wrong interpretation of the CR.

Stoyanchev et al. (2014) point out how important it is for the communicative efficiency in human-machine interaction to have clarification requests which are not generic but *targeted*, in that they are based on contextual information. For instance, in Müller et al. (2021), rephrasing strategies are used to ask for correctness before grounding the received information.

As it will be pointed out at the end of this section, other types of misunderstanding repair strategies to be considered are more classifiable as related to argumentation acts of grounded information, a field of research that is becoming worth exploring.

Request Acknowledgment. In the “Media Equation” (Reeves and Nass 1996), it is hypothesised that “people will give more spontaneous back-channels to a spoken dialogue system that makes more spontaneous back-channel-inviting cues than a spoken dialogue system that makes less spontaneous ones”. Based on this hypothesis, Misu et al. (2011) presented the basis for a dialogue system capable of eliciting back-channels from users. For this purpose, they constructed a dialogue-style TTS which makes use of back-channel-inviting cues, whose application resulted in the more user’s spontaneous back-channels, informative for the system. Similarly, Buschmeier and Kopp (2014) defined when the system should elicit feedback in the user in order to avoid undesirable dialogue states. In fact, the system needs feedback when i) its belief about the user’s mental state is not informative enough; ii) its belief about the user’s mental state has not changed in a long time; iii) its belief about the user’s mental state is different from a desired one deriving from a previous communicative action by the agent. In Buschmeier and Kopp (2018), the same result as in Misu (2011) was reported: participants provided more feedback with an attentive listener agent, that is with agent capable of a) interpreting communicative listener feedback from users, b) adapting their production to the users’ needs, whose interpretation is based on the processed feedback, and c) eliciting feedback through feedback elicitation cues when needed. The use of such feedback is moreover important to other grounding acts, such as *initiate*.

In this section, we focused on defining a parallel between theory and application, by describing some works on dialogue systems which explicitly applied grounding acts in the dialogue. As a take-home message, it can be pointed out that theory was diversely adapted to the available technology and different new methodologies were implemented. A perfect mapping between theory and application has not yet been reached. Some aspects of grounding were therefore more investigated than others, and some others became crucial. In general, the importance of the grounding process has been variously highlighted, starting from uncertainty signalisation (Fernández et al. 2007; Hough and Schlangen 2017), to different degrees of grounding (Roque and Traum 2008; Roque 2009; Petukhova et al. 2015), to the use of grounding in dialogue systems evaluation (Curry, Hastie, and Rieser 2017; Zou 2020). The research on dialogue systems, in fact, has always underlined the need to test and evaluate their functionality and performances. Nevertheless, the evaluation of dialogue systems has always been a problematic task to carry out. When Turing (1950) suggested the imitation game as a possible evaluation of the intelligence a machine can show, he was thinking of replacing the question whether a machine is able to think with its imitation capabilities. The concept of thinking has always been difficult to define. Instead, the imitation game could actually be a valid and answerable question to pose. To answer this question positively, the evaluator should not be able to tell the difference between the machine

and the human interlocutor, in that the machine succeeded in imitating intelligent human-like behaviour. Here, the concept of intelligence needs some in-depth consideration. Gottfredson (Gottfredson 1997, p. 13) defined *intelligence* as the “ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience”. As we may easily comprehend, this definition is far from the possibility a machine can have to imitate some behaviours. If the aim is not only to reproduce, but also to evaluate some intelligent aspects a machine could have, we may need to adopt different tests. Therefore, the Turing test, although sometimes still used, can conversely represent the desirable goal of an intelligent agent, which shows behaviours that are human-like, rather than an evaluation tool for system performances. At the same time, the question that could here be raised is whether we really want a system to be completely indistinguishable from human beings and why we want that. Conversely, we might want systems capable of showing their specific intelligent features which might be suitable for *artificial beings* only. Similarly to Turing, Schatzmann et al. (2005) evaluate two aspects: i) human-like system’s responses; ii) how well the user models cover the variety of the user population in the training data. Even here, what is missing is a shareable framework to carry out this evaluation and an in-depth description of how the system is actually working.

Whichever is the way we imagine our dialogue system to be, the evaluation should rather consider some specific traits of what we call intelligence, or better, in this case, of interactional intelligence. With *interactional intelligence*, we mean the ability to recognise intentions, beliefs, and aptitudes towards the dialogic exchange and the ability to respond appropriately (Levinson 1995; Buschmeier 2018). As we will see in section 4, this capability makes the system argumentation-skilled. For goal-oriented dialogue systems, the completeness of the task, dialogue length, and user satisfaction are usually taken into account. On the other hand, for general purpose dialogue systems, approaches like next utterance classification and word perplexity are preferred (Serban et al. 2018). To the present day, fully satisfactory automatic classification metrics for dialogue systems do not exist. Nevertheless, the combination of different methodologies could lead to better results. Grounding acts can, in this sense, also be used as a methodology to evaluate dialogue system’s performances. More specifically, Curry et al. (2017) report a comparison between systems using explicit feedback and systems using implicit feedback. In Zou (2020), on the other hand, evaluation techniques are compared and faults are highlighted in that not “all aspects of dialogue from naturalness and coherence to long-term engagement and flow” are captured. One possible evaluation metrics could consider usability principles (Dix et al. 2003), namely learnability, flexibility, robustness. Specifically, as far as the robustness principle is concerned, that is the level of support that the system provides to the user in completing and assessing a task successfully, dialogue systems can make their internal states observable through verbal or non-verbal interaction, thus via grounding acts. More in detail, when problems occur in information processing, the observable character of such states can be utilised to recover the problems. In section 4, the need for this type of analysis will be better detailed.

3.1 Latest datasets for grounding acts

In order to make the process of grounding modelling possible, dialogue datasets are needed. In the past years, many dialogue datasets have been collected to study grounding and grounding-related problems (Serban et al. 2018). Nevertheless, the latest corpora collections are particularly important to mention as they are mostly concerned with collecting large amount of data in order to be used to train dialogue systems

with machine learning, which indeed need more data when compared to past collections. Different techniques can be used to model and train dialogue systems: whereas some use online learning (Liu and Mazumder 2021), reinforcement learning (Pietquin 2007; Young et al. 2010), probabilistic reasoning (Skantze 2007; Stoyanchev, Lison, and Bangalore 2016; Rossignol, Pietquin, and Ianotto 2010), or graph representations (Liu and Mei 2020; Mi et al. 2020; Chaudhuri, Rony, and Lehmann 2021), many grounding phenomena are learned and modelled in conversational agents via machine learning algorithms. It is important to point out that grounding can be better observed in spontaneous conversation, as eliciting it can be easier for some aspects (i.e., feedback) rather than for others. For this purpose, in the past, there have been works on agents interacting with humans applying improvisation (Bruce et al. 2000; Martin, Harrison, and Riedl 2016; Winston and Magerko 2017). Nevertheless, there are not so many corpora collecting such spontaneous dialogues, and the ones available are also far too small for machine learning purposes (Busso and Narayanan 2008). In the SPOLIN corpus (Cho and May 2020), 6,760 English Improv dialogues, comprising 90,000 turn pairs, have been collected. The *improvisational theatre* dialogues considered here are important for grounding purposes, as in this form of theatre everything is performed without a script, a scenery, or other established environment; for this reason, everything must be grounded via interactions. The specific aim of this dataset was to study *yes-and*s turns, where an acknowledgement act was combined with a new next relevant contribution. Similarly, different common grounding phenomena, like the ones described in Traum (1994), are observable in the collection presented in Udagawa and Aizawa (2019, 2020), comprising 6,760 dialogues, and whose aim is to be adopted in the training of end-to-end dialogue systems. End-to-end dialogue systems, in fact, are usually based on neural networks (Shang, Lu, and Li 2015; Vinyals and Le 2015; Sordoni et al. 2015; Dodge et al. 2016; Serban et al. 2016) and need large amount of data. For the same purposes, Chen et al. (2021) collected 10K human-to-human dialogues containing 55 distinct user intents. The few amount of appropriate dialogue corpora for grounding applications in dialogue systems in various languages can be still considered as the Achilles' heel of the data-driven research, like the machine learning-based one.

4. "What the heck are you saying?" Corrective dialogues and grounded information

As reported in the previous sections, different scholars highlighted the urge of including grounding processing in their systems, for which argumentation of grounded information needs more investigation. In this section, the attention will be focused on grounding-related corrective dialogues. In this context, the argumentative nature of some of such dialogues, in the form of Common Ground Inconsistencies, will also be taken into account.

Among the most investigated grounding aspects, corrective dialogues have drawn much attention as their adoption improves the communication process. This resulted from the users' need to interact with an agent capable of cooperating with the communicative actions. Human interlocutors always contribute with questions, answers, and feedback (Beun and van Eijk 2004). For instance, a corrective dialogue is a particular type of dialogue occurring when: i) the user notices an error in the system and corrects it; ii) the user changes their mind; iii) the user's beliefs are in contradiction with the system's beliefs and expectations. In the first two cases, the corrective dialogue is initiated by the user (it corresponds to the grounding act of Repair), whereas, in the last case, it is initiated by the system (it corresponds to the grounding act of RequestRepair) (Bousquet-Vernhettes, Privat, and Vigouroux 2003). One example of corrective dialogue

in human-machine interaction is the one presented in Beun and van Eijk (2004). The authors focused on a particular communicative problem related to conceptual discrepancies between a computer system and its user. Starting from the assumption that both the system and its user have a mental representation of a domain, the *mental* representation of the system, e.g., the ontology, contains conceptualisations that are made explicit in a formal language. Despite their possible incompleteness and inaccuracy, this information can be used to trace the system's reasoning about concepts, items, and their properties. Most importantly, this representation also allows the detection of conceptual discrepancies, arising when the system observes that the user applies an incorrect action to a particular object. The authors also stated that, although feedback of different kinds are now generally used in such systems, there is still no accurate *mathematical theory* for natural communicative behaviours and their computational model to human-machine interaction, especially as far as conceptual discrepancies are concerned. What is still missing is, therefore, a reference model guiding the adoption of a specific type, content, and form of the feedback that has to be generated in a particular situation (Beun and van Eijk 2004).

While conceptual discrepancies can be concerned with the last dialogue state whose inaccuracy can lead to a RepairRequest + Repair or directly to a Repair act, some inconsistency can also refer to a previous stage of the interaction, as in Khouzaimi et al. (2015). In this case, Traum (1994) considers the consequent acts as argumentation ones, as already grounded information are now being negotiated. The linguistic activity of argumentation is pragmatically regulated by a sequence of purposive speech acts in conflict (Walton and Godden 2006), as it represents the discussion of opposing ideas to find the truth, namely dialectics. Dialectics in dialogue systems can be framed in the field of formal and computational argumentation, where two main research topics are listed: argumentation-based inference and argumentation-based dialogue. Argumentation-based inference concentrates on establishing what conclusions can be drawn starting from incomplete or inconsistent information. Argumentation-based inference models work similarly to Hegel's dialectic, since they investigate statements from a logical point of view without considering multiple participants. Historically, the first one who described an Abstract Argumentation Framework was Dung (1995). On the other hand, Pollock (1987) first established the basis for formal argumentation-based inference.

Argumentation-based inference is different from argumentation-based dialogue, in that the former is a formal method which is applied to a single entity to decide about the truth of an argument. On the other hand, argumentation-based dialogue considers problems arising from dialogues among different agents. In such cases, information is, in fact, distributed among the agents, who may or may not be willing to share it at different points in time due to individual strategies and goals. A solid argumentation-based dialogue theoretical framework is, in fact, still missing because of the complexity of the phenomenon in question: *"the study of argumentation-based dialogue consists of a variety of different approaches and individual systems, all exciting work but with few unifying accounts or general frameworks"* (Prakken 2017, p. 53). Among the types of dialogue that are studied in argumentation-based studies, we mention persuasion, negotiation, information seeking, deliberation, inquiry, and quarrel (Walton 1984; Walton and Krabbe 1995). These classes, however, are not meant to be absolute, as multiple goals may be present during a single dialogue. Among the ones listed, persuasion dialogues appear to have been studied the most (Yuan, Moore, and Grierson 2004; Prakken 2008). As far as deliberation dialogues are concerned, the collaboration, here, takes place to find an optimal solution to a problem for which the involved agents have not yet a solution.

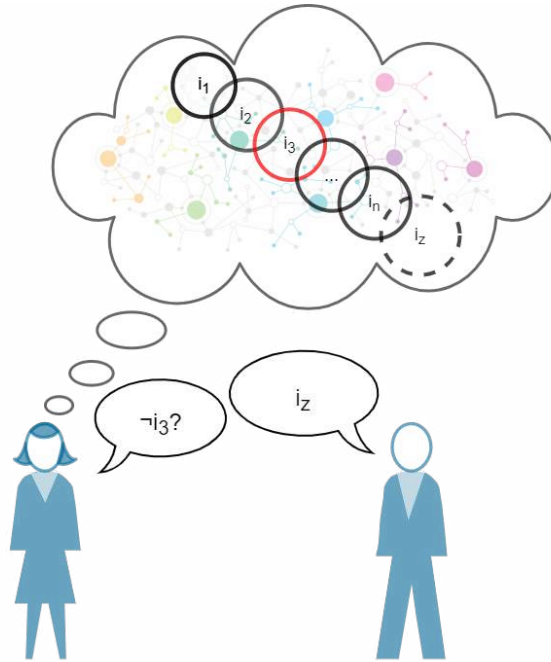


Figure 5
Representation of the Common Ground CRs elicitation scenario

For this type of dialogue, an interesting result was found. In case of a two-agents system adhering strictly to the communication protocol, forming their claims on the basis of their knowledge and adopting a collaborative attitude, it was demonstrated that the agreed solution is always acceptable to both parties (Black and Atkinson 2010). This results from employing argumentation, whose usefulness in dialogue systems, designed for deliberation, was demonstrated in Kok et al. (2010).

The problem that characterises argumentation-based dialogue with respect to argumentation-based inference is, therefore, the presence of different agents in the setting. This introduces multiple, not necessarily aligned, knowledge and, possibly, conflicting goals in the pursuit of a solution to a problem. Pragmatic strategies adopted in such situations are to be investigated, as they are generally concerned with grounded information. Based on the analysis of map-tasks, whose structure can be compared to deliberation dialogue for their goals, an argumentation-based act trigger was identified (Di Maro et al. 2021), namely Common Ground Inconsistencies, which can lead the interlocutor to the adoption of clarification requests, as its corresponding argumentation-based act. Similarly to the aforementioned conceptual discrepancies, Common Ground Inconsistencies refer to problems with grounded information.

In Figure 5, a Common Ground Inconsistency scenario eliciting a Common Ground clarification request (Common Ground CR) is displayed. With Common Ground CR we refer to clarification requests with an argumentative function. In fact, they do not help the speaker ground a piece of information, but they refer to previous discourse units, where that piece of information was already grounded. In the current state of the dialogue, a new evidence clashes with the grounded one, and, therefore the Common

Ground CR is uttered² (Di Maro 2021; Di Maro, Origlia, and Cutugno 2021a). As in Figure 5, in the mind of the female agent *A*, the Communal Common Ground is stored to guide the process of accumulating information in the Personal Common Ground. The information $(i_1, i_2, i_3, \dots, i_n)$ are communicated by the male agent *B* to *A*, and sequentially stored in her Personal Common Ground. When *B* utters a new information i_z , this is represented as a new item candidate to be part of the Personal Common Ground. This representation generates a bias/evidence conflict (Domaneschi, Romero, and Braun 2017), in that the presence of the new item i_z in the Personal Common Ground clashes with the presence of another item i_3 , whose validity is now questioned. This conflict represents a Common Ground Inconsistency and is translated in the Common Ground CR $\neg i_3?$, whose form, function and illocutive effect are reported in Di Maro et al. (2021b). As also highlighted in Di Maro et al. (2021), polar questions are especially important to express Common Ground Inconsistencies, in that their epistemic stance is clearly expressed compared to other types of questions (Domaneschi, Romero, and Braun 2017). Finally, differently from other CRs, Common Ground CRs do not necessarily refer to the immediately previous utterance, but to previously - correctly or wrongly - grounded information.

5. Conclusion

In Human-Machine interaction, the study and application of pragmatic aspects has covered few phenomena, although their importance was recognised in various linguistic studies. On the one hand, error handling and requests for clarification have always had a central role, since the correct understanding and the consequent task completion of the system are the desired goals. On the other hand, back-channels and acknowledgement feedback have also been investigated to ensure grounding. If commercial systems try to identify possible mistakes which can be caused by users or by technology limits, their ability to understand the real cause of problems to adequately signal them and let the human user correct them is still a frontier not exhaustively explored. The complexity of possible misunderstanding and conflicting situations makes it necessary to study the communicative strategies used to efficiently handle the related interaction problems.

As mentioned at the beginning of this work, the aim pursued here was also to stimulate further investigations and applications of pragmatics, and especially grounding, in conversational agents, by underlying application gaps. In fact, whereas semantics has been a more investigated topic within the dialogue systems field with respect to pragmatics, where speech acts modelling drew more attention. Furthermore, although CRs and corrective dialogues are widely studied in linguistics, their application in dialogue systems is still limited, especially when referred to already grounded information. Further investigations on grounding-related problems concerning dialogue states which do not necessarily correspond to the current dialogue state but to previous steps of the dialogue history are therefore needed. This could, moreover, expand the study on argumentation-based dialogues leading to the foundation of a shared theoretical framework.

² We are aware that Clarification Requests are generally used to correctly update the common ground. Nevertheless, the term *Common Ground CR* refers here, as in the mentioned studies, to requests used to check what is already stored in the common ground.

Acknowledgments

The author would like to thank the reviewers for their comments and help. This work was also supported by the Italian PON I&C 2014-2020 within the BRILLO project, no. F/190066/01-02/X44.

References

- Airenti, Gabriella, Bruno Giuseppe Bara, and Marco Colombetti. 1993. Conversation and behavior games in the pragmatics of dialogue. *Cognitive Science*, 17(2):197–256.
- Allwood, Jens. 1995. An activity based approach to pragmatics. In *Abduction, belief and context in dialogue: Studies in computational pragmatics*. John Benjamins.
- Allwood, Jens, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.
- Austin, John Langshaw. 1975. *How to do things with words*, volume 88. Oxford University Press.
- Bara, Bruno Giuseppe. 1999. *Pragmatica cognitiva: i processi mentali della comunicazione*. Bollati Boringhieri.
- Benotti, Luciana and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online, April.
- Beun, Robbert-Jan and Rogier M. van Eijk. 2004. Conceptual discrepancies and feedback in human-computer interaction. In *Proceedings of the conference on Dutch directions in HCI*, page 13. Association for Computing Machinery, New York, NY, United States, June.
- Black, Elizabeth and Katie Atkinson. 2010. Agreeing what to do. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 12–30, Toronto, Canada, May. Springer.
- Bousquet-Vernhettes, Caroline, Régis Privat, and Nadine Vigouroux. 2003. Error handling in spoken dialogue systems: toward corrective dialogue. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Château d’Oex, Vaud, Switzerland, August.
- Brown, Roger. 1958. How shall a thing be called? *Psychological review*, 65(1):14.
- Bruce, Allison, Jonathan Knight, Samuel Listopad, Brian Magerko, and Illah R. Nourbakhsh. 2000. Robot improv: Using drama to create believable agents. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 4, pages 4002–4008, San Francisco, CA, USA, August. IEEE.
- Buschmeier, Hendrik. 2018. *Attentive Speaking. From Listener Feedback to Interactive Adaptation*. Ph.D. thesis, Faculty of Technology, Bielefeld University, Bielefeld, Germany.
- Buschmeier, Hendrik and Stefan Kopp. 2014. When to elicit feedback in dialogue: Towards a model based on the information needs of speakers. In *International Conference on Intelligent Virtual Agents*, pages 71–80, Boston, August. Springer.
- Buschmeier, Hendrik and Stefan Kopp. 2018. Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS ’18*, pages 1213–1221, Stockholm, Sweden. International Foundation for Autonomous Agents and Multiagent Systems.
- Busso, Carlos and Shrikanth S. Narayanan. 2008. Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the iemocap database. In *Ninth annual conference of the international speech communication association (ISCA)*, Brisbane, Australia, September.
- Chaudhuri, Debanjan, Md Rashad Al Hasan Rony, and Jens Lehmann. 2021. Grounding dialogue systems via knowledge graph aware decoding with pre-trained transformers. In *European Semantic Web Conference*, pages 323–339, Online, August. Springer.
- Chen, Derek, Howard Chen, Yi Yang, Alex Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, June 6–11.
- Cho, Hyundong and Jonathan May. 2020. Grounding conversations with improvised dialogues. page 2398–2413, Online, July 5 - 10.
- Clark, Eve V. 2015. Common ground. In *The Handbook of Language Emergence*. Wiley, Chichester, UK, page 328–353.
- Clark, Herbert H. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

- Clark, Herbert H. and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 222–233, Washington, DC, USA. American Psychological Association.
- Clark, Herbert H. and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.
- Clark, Herbert H. and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.
- Cohen, Philip R. and Hector J. Levesque. 1991. Confirmations and joint action. In *Proceeding of the International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, pages 951–959, Sydney, Australia, August.
- Curry, Amanda Cercas, Helen Hastie, and Verena Rieser. 2017. A review of evaluation techniques for social dialogue systems. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*, pages 25–26, Glasgow, UK, 13 November.
- Dahlbäck, Nils and Arne Jönsson. 1998. A coding manual for the linköping dialogue model. *unpublish manuscript*.
- Di Maro, Maria. 2021. "Shouldn't I use a polar question?" Proper question forms disentangling inconsistencies in dialogue systems. *Unpublished Dissertation, Università degli Studi di Napoli Federico II*.
- Di Maro, Maria, Hendrik Buschmeier, Stefan Kopp, and Francesco Cutugno. 2021. Clarification requests negotiating personal common ground. In *Proceedings of the XPRAG.it (Poster)*, Online, July. URL = <https://osf.io/tmjrf/>.
- Di Maro, Maria, Antonio Origlia, and Francesco Cutugno. 2021a. Common ground inconsistencies in dialogue systems: conflict patterns implied by polar question forms (submitted). *Speech Communication*.
- Di Maro, Maria, Antonio Origlia, and Francesco Cutugno. 2021b. Polarexpress: Polar question forms expressing bias-evidence conflicts in Italian. *International Journal of Linguistics*, 13(4):14–35.
- Di Maro, Maria, Jana Voße, Francesco Cutugno, and Petra Wagner. 2019. Perception breakdown recovery in computer-directed dialogues. In *Proceedings of the First International Seminar on the Foundations of Speech (SEFOS 2019)*, Sønderborg, Denmark, December.
- Dix, Alan, Alan John Dix, Janet Finlay, Gregory D. Abowd, and Russell Beale. 2003. *Human-computer interaction*. Pearson Education.
- Dodge, Jesse, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating prerequisite qualities for learning end-to-end dialog systems. In *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico, May 2 - 4.
- Domaneschi, Filippo, Maribel Romero, and Bettina Braun. 2017. Bias in polar questions: Evidence from English and German production experiments. *Glossa: a Journal of General Linguistics*, 2(1).
- Dung, Phan Minh. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.
- Eshghi, Arash, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 261–271, London, UK, April.
- Fernández, Raquel, Andrea Corradini, David Schlangen, and Manfred Stede. 2007. Towards reducing and managing uncertainty in spoken dialogue systems. In *Proceedings of the 7th International Workshop on Computational Semantics (IWCS07)*, Tilburg, Netherlands, January.
- Gabsdil, Malte. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35, Palo Alto, California, March.
- Ginzburg, Jonathan and Robin Cooper. 2001. Resolving ellipsis in clarification. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 236–243, Toulouse, France, July.
- Ginzburg, Jonathan, Raquel Fernández, and David Schlangen. 2007. Unifying self-and other-repair. In *Decalog 2007: proceedings of the eleventh Workshop on the Semantics and Pragmatics of Dialogue*, Rovereto, Italy, May 30 - June 1. Rotooffset Paganella.
- Ginzburg, Jonathan and Zoran Macura. 2005. The emergence of metacommunicative interaction: Some theory, some practice. In *Proceedings of the 2nd International Symposium on the Emergence*

- and *Evolution of Linguistic Communication*, pages 35–40, Hatfield, UK, April.
- Gottfredson, Linda S. 1997. Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24:13–23.
- Grice, Herbert Paul. 1975. Logic and conversation. In *Speech acts*. Brill, pages 41–58.
- Grice, Paul. 1989. *Studies in the way of words*. Harvard University.
- Hirst, Graeme, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. 1994. Repairing conversational misunderstandings and non-understandings. *Speech Communication*, 15(3-4):213–229.
- Hough, Julian and Matthew Purver. 2012. Processing self-repairs in an incremental type-theoretic dialogue system. In *Proceedings of the 16th SemDial workshop on the semantics and pragmatics of dialogue*, pages 19–21, Paris, France, September.
- Hough, Julian and David Schlangen. 2017. It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 274–282, Vienna, Austria, March. IEEE.
- Hough, Julian, Sina Zarrieß, and David Schlangen. 2017. Grounding imperatives to actions is not enough: A challenge for grounded nlu for robots from human-human data. In *GLU 2017 International Workshop on Grounding Language Understanding*, pages 88–91, Stockholm, Sweden, 25 August.
- Huang, Yan. 2017. *The Oxford Handbook of Pragmatics*. Oxford University Press.
- Khouzaimi, Hatim, Romain Laroche, and Fabrice Lefevre. 2015. Turn-taking phenomena in incremental dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1890–1895, Lisbon, Portugal, September.
- Kilger, Anne and Wolfgang Finkler. 1995. Incremental generation for real-time applications. *Tech. rep. RR-95-11. Saarbrücken, Germany: Deutsches Forschungszentrum für Künstliche Intelligenz*.
- Kok, Eric M., John-Jules Ch. Meyer, Henry Prakken, and Gerard A.W. Vreeswijk. 2010. A formal argumentation framework for deliberation dialogues. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 31–48, Toronto, Canada, May. Springer.
- Larsson, Staffan and David R. Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural language engineering*, 6(3-4):323–340.
- Leech, Geoffrey. 2003. Pragmatics and dialogue. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.
- Levinson, Stephen C. 1995. Interactional biases in human thinking. In *Social intelligence and interaction*. Cambridge University Press, pages 221–260.
- Liu, Bing and Sahisnu Mazumder. 2021. Lifelong and continual learning dialogue systems: learning during conversation. *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-2021)*, February.
- Liu, Bing and Chuhe Mei. 2020. Lifelong knowledge learning in rule-based dialogue systems. *arXiv preprint arXiv:2011.09811*.
- Marge, Matthew and Alexander Rudnicky. 2015. Miscommunication recovery in physically situated dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 22–31, Prague, Czech Republic, September.
- Marge, Matthew and Alexander I. Rudnicky. 2019. Miscommunication detection and recovery in situated human-robot dialogue. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9(1):1–40.
- Martin, Lara J., Brent Harrison, and Mark O Riedl. 2016. Improvisational computational storytelling in open worlds. In *International Conference on Interactive Digital Storytelling*, pages 73–84, Los Angeles, CA, USA, November. Springer.
- Mi, Jinpeng, Jianzhi Lyu, Song Tang, Qingdu Li, and Jianwei Zhang. 2020. Interactive natural language grounding via referring expression comprehension and scene graph parsing. *Frontiers in Neurobotics*, 14.
- Misu, Teruhisa, Etsuo Mizukami, Yoshinori Shiga, Shinichi Kawamoto, Hisashi Kawai, and Satoshi Nakamura. 2011. Toward construction of spoken dialogue system that evokes users' spontaneous backchannels. In *Proceedings of the SIGDIAL 2011 Conference*, pages 259–265, Portland, Oregon, June.
- Müller, Romy, Dennis Paul, and Yijun Li. 2021. Reformulation of symptom descriptions in dialogue systems for fault diagnosis: How to ask for clarification? *International Journal of Human-Computer Studies*, 145:102516.
- Petukhova, Volha, Harry Bunt, Andrei Malchanau, and Ramkumar Aruchamy. 2015. Experimenting with grounding strategies in dialogue. In *Proceedings of the 19th Workshop on the*

- Semantics and Pragmatics of Dialogue (SEMDIAL 2015 goDIAL)*, Gothenburg, Sweden, August.
- Pietquin, Olivier. 2007. Learning to ground in spoken dialogue systems. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV-165, Honolulu, HI, USA, June. IEEE.
- Pollock, John L. 1987. Defeasible reasoning. *Cognitive science*, 11(4):481-518.
- Prakken, Henry. 2008. A formal model of adjudication dialogues. *Artificial Intelligence and Law*, 16(3):305-328.
- Prakken, Henry. 2017. Historical overview of formal argumentation. *IfCoLog Journal of Logics and their Applications*, 4(8):2183-2262.
- Purver, Matthew. 2004a. Clarie: The clarification engine. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 77-84, Barcelona, Spain, July.
- Purver, Matthew. 2004b. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, University of London, London, UK.
- Purver, Matthew. 2006. Clarie: Handling clarification requests in a dialogue system. *Research on Language and Computation*, 4(2-3):259-288.
- Purver, Matthew, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in cognitive science*, 10(2):425-451.
- Reeves, Byron and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, UK.
- Rodríguez, Kepa Joseba and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue*, Barcelona, Catalonia, Spain, July.
- Roque, Antonio. 2009. *Dialogue management in spoken dialogue systems with degrees of grounding*. University of Southern California.
- Roque, Antonio and David Traum. 2008. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54-63, Columbus, Ohio, USA, June.
- Roque, Antonio and David Traum. 2009. Improving a virtual human using a model of degrees of grounding. In *Twenty-First International Joint Conference on Artificial Intelligence*, Pasadena, California, USA, July.
- Rossignol, Stéphane, Olivier Pietquin, and Michel Ianotto. 2010. Simulation of the grounding process in spoken dialog systems with bayesian networks. In *International Workshop on Spoken Dialogue Systems Technology*, pages 110-121, Shizuoka, Japan, October. Springer.
- Schatzmann, Jost, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*, Lisbon, Portugal, September.
- Schegloff, Emanuel A, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361-382.
- Schettino, Loredana, Maria Di Maro, and Francesco Cutugno. 2020. Silent pauses as clarification trigger. In *Laughter and Other Non-Verbal Vocalisations Workshop: Proceedings (2020)*, Bielefeld, Germany, October.
- Schlangen, David. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136-143, Cambridge, Massachusetts, USA, April 30 - May 1.
- Schlangen, David. 2019. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings. *Computing Research Repository Journal*, August.
- Schlangen, David and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83-111.
- Searle, John R. 1985. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.
- Serban, Iulian Vlad, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1-49.
- Serban, Iulian Vlad, Alessandro Sordani, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Conference of the Association for the Advancement of Artificial Intelligence*, volume 16, pages 3776-3784, Phoenix, Arizona, USA, February.
- Shang, Lifeng, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

- Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, July.
- Skantze, Gabriel. 2007. Making grounding decisions: Data-driven estimation of dialogue costs and confidence thresholds. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 206–210, Antwerp, Belgium, September.
- Skantze, Gabriel. 2008. Galatea: A discourse modeller supporting concept-level error handling in spoken dialogue systems. In *Recent Trends in Discourse and Dialogue*. Springer, pages 155–189.
- Skantze, Gabriel, David House, and Jens Edlund. 2006. User responses to prosodic variation in fragmentary grounding utterances in dialog. In *Ninth International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, September.
- Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, May–June.
- Stalnaker, Robert. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.
- Stoyanchev, Svetlana, Pierre Lison, and Srinivas Bangalore. 2016. Rapid prototyping of form-driven dialogue systems using an open-source framework. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–219, Los Angeles, September.
- Stoyanchev, Svetlana, Alex Liu, and Julia Hirschberg. 2014. Towards natural clarification questions in dialogue systems. In *AISB symposium on questions, discourse and dialogue*, volume 20, April.
- Swerts, Marc, Diane Litman, and Julia Hirschberg. 2000. Corrections in spoken dialogue systems. In *Sixth International Conference on Spoken Language Processing*, Beijing, China, October.
- Tomasello, Michael. 2010. *Origins of human communication*. MIT press.
- Traum, David R. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.
- Traum, David R. 1999. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems: Papers from the AAAI Fall Symposium*, pages 124–131, North Falmouth, MA, USA, November.
- Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 10.
- Udagawa, Takuma and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 01, pages 7120–7127, Hilton Hawaiian Village, Honolulu, Hawaii, USA, January 27 – February 1.
- Udagawa, Takuma and Akiko Aizawa. 2020. An annotated corpus of reference resolution for interpreting common grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 05, pages 9081–9089, Hilton New York Midtown, New York, USA, February.
- Vinyals, Oriol and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.
- Visser, Thomas, David Traum, David DeVault, and Riëks op den Akker. 2012. Toward a model for incremental grounding in spoken dialogue systems. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, Santa Cruz, CA, USA, September.
- Visser, Thomas, David Traum, David DeVault, and Riëks op den Akker. 2014. A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8(1):61–73.
- Walton, Douglas and David M. Godden. 2006. The impact of argumentation on artificial intelligence. *Considering Pragma-Dialectics: A Festschrift for Frans H. Van Eemeren on the Occasion of his 60th Birthday*, pages 287–299.
- Walton, Douglas and Erik C.W. Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.
- Walton, Douglas N. 1984. *Logical Dialogue-Games*. University Press of America, Lanham, Maryland.
- Wang, Zhiyang, Jina Lee, and Stacy Marsella. 2013. Multi-party, multi-role comprehensive listening behavior. *Autonomous Agents and Multi-Agent Systems*, 27(2):218–234.
- Winston, Lauren and Brian Magerko. 2017. Turn-taking with improvisational co-creative agents. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, 1, Snowbird, Utah, USA, October.

- Young, Steve, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.
- Yuan, Tangming, David Moore, and Alec Grierson. 2004. Human-computer debate, a computational dialectics approach. *unpublished doctoral dissertation, Leeds Metropolitan University*.
- Zou, Yiqian. 2020. An experimental evaluation of grounding strategies for conversational agents. *Master Thesis. Department of Philosophy, Linguistics and Theory of Science. University of Gothenburg*.

Cutting melted butter? Common Ground inconsistencies management in dialogue systems using graph databases

Maria Di Maro*
Università di Napoli ‘Federico II’

Antonio Origlia**
Università di Napoli ‘Federico II’

Francesco Cutugno†
Università di Napoli ‘Federico II’

In this work, a spoken dialogue system architecture capable of dealing with Common Ground inconsistencies is proposed. Specifically, attention will be drawn upon the Conflict Search Graph, with insights on its ability to recognise problems and make them explicit via polar questions. Appropriate question forms are, indeed, adopted for the occurring type of common ground conflict, based on previous experiments, which showed that providing automatic dialogue systems with such grounding capabilities can lead to improved usability and naturalness. The described system architecture is, thus, able to detect conflicts and to use argumentation-based pragmatic strategies to signal them consistently with previous observations.

1. Introduction

Dialogue systems, also referred to as conversational agents, are nowadays in the spotlight in different commercial, academic and industrial sectors: suffice to consider the success and popularity of tools like Amazon Alexa and Google Home (López, Quesada, and Guerrero 2017), or widespread in-car dialogue systems (Becker et al. 2006; Kousidis et al. 2014). Conversational agents are computer systems capable of interacting with humans through verbal signals. They are one of the most currently investigated field of Artificial Intelligence, since the ability to communicate inferences and one’s understanding by means of language is one possible way to manifest intelligence (Sperber and others 1994). While a shared opinion of how *intelligence* can be defined is far from being widely accepted (Warner 2002), one possible definition is proposed in (Legg and Hutter 2007), which define it, despite all the criticism, as “the capacity for knowledge, and knowledge possessed.”. In this definition, one concept draws particular attention: ‘knowledge’, as knowledge bases are a crucial aspect for dialogue systems to appear *intelligent*. Concerning the approaches used in such systems, these appear to be distributed in a continuum where we find, at the extremes, systems using deterministic rules to react to specific signals (McGlashan et al. 1992), and end-to-end dialogue systems which do not make any distinction in the abilities the system should perform at different levels, but are rather trained with data from which tendencies are statistically extracted (Ritter, Cherry, and Dolan 2010; Vinyals and Le 2015; Serban et al. 2016; Bordes, Boureau,

* Interdepartmental Center for Advances in Robotic Surgery E-mail: maria.dimaro2@unina.it

** URBAN/ECO Research Center E-mail: antonio.origlia@unina.it

† URBAN/ECO Research Center E-mail: cutugno@unina.it

and Weston 2016). In the middle, there are hybrid systems using either statistical or deterministic approaches to implement different modules dedicated to the management of specific strategies and tools. Overall, in the field of language understanding and generation, the corpus-driven approach is becoming increasingly important to infer, with the application of machine learning algorithms, knowledge and communicative strategies (Serban et al. 2018). Nevertheless, beyond pattern recognition capabilities provided by machine learning algorithms, decision making in dialogue management still benefits from the design of appropriate knowledge representation, which supports both the efficiency and the interpretability of a technological system.

Knowledge representation dedicated to dialogue management is very close to the concept of Common Ground, that is mutual knowledge, beliefs, and assumptions, as the foundation for mutual understanding in conversation (Clark and Brennan 1991). Common ground, as Clark (Clark 2015) acknowledged, can be of four main types: personal, local, communal and specialised. *Personal Common Ground* (PCG) is established collecting information over time through communicative exchanges with an interlocutor and it can be considered as a record of shared experiences with that person. A part of PCG is *Local Common Ground* that is tied to a piece of information obtained from a single exchange with an unknown or known interlocutor. According to Clark (Clark 2015), information of this type can be, for instance, the opening hours for a shop, train timetables, and so on. With *Communal Common Ground* (CCG), it is intended an amount of information shared with people that belong to the same community, that is to say, people that share general knowledge, knowledge about social background, education (schools attended, levels of education attained), religion, nationality, and language(s). Within a larger community, a smaller one can be found: *Specialised Common Ground* pertains to those people that share particular areas of expertise about some domain of knowledge, such as colleagues, friends, or acquaintances, and it is marked by specialised vocabulary of that specific domain, such as medicine, law, and so on. For the purposes of this work, only PCG and CCG are going to be considered, where CCG defines the rules of the cooking domain, for which it is common knowledge that, for instance, butter is an ingredient, and where PCG stores the given information concerning the recipes steps.

This work aims at investigating the following research questions:

1. Is it possible to design a knowledge representation module hosting, at the same time, both the CG and the dialogue state?
2. Is it possible to use CG inconsistencies detection as an argumentation-based dialogue system metrics?

The general objective of this work is to investigate how inconsistencies in the knowledge stored in the CG can be efficiently detected with as much detail as possible to support error reporting in a dialogue system. Specifically, we propose the use of graph databases as an integrated solution to dialogue state tracking, knowledge representation and conflict detection as a fundamental building block for dialogue systems with argumentation capabilities.

The paper is organised as follows: in Section 2, we summarise the theories underlying our approach and motivating the proposed system's architecture, while in Section 3, we report similar previous works closely related to the one presented here. Section 4 describes the proposed system's architecture and the materials used to test its conflict detection capabilities. Section 5 describes how the graph structure representing the CCG

was assembled using freely available resources. Section 6 describes, instead, how the PCG is built, as commands are issued from a simulated user, and how consistency checks are performed, at each iteration, to verify that the PCG consistency is not compromised by the last command. The same Section describes the procedure used to extract inconsistency details after a conflict is detected. Lastly, Section 7 reports the results obtained using simulated dialogues together with error analysis.

2. Background theory

As anticipated, dialogue systems are interactive devices. *Interacting* refers to actions that have some effect on others. The mutual influence agents can have on one another is built through communicative processes, both verbal and not verbal. On the other hand, *communicating* means to transmit information. According to the Shannon-Weaver model of communication, mostly applicable to machines' interaction, communication deals with the transmission of signals from one system to another, where the system communicating can be of the same nature or not (Shannon 1948). According to this model, the transmitter encodes a message which is sent through a channel to the receiver who decodes it. The communication channel is also called *noise* because it can be loaded with noise of different kind. Nevertheless, communication is more than just transmitting information, as information must be processed in order to enable the receiving agent to produce a coherent output. Moreover, as stated by Allwood (Allwood 2013), communication includes not only the sharing of information, but also of cognitive content or understanding with varying degrees of awareness and intentionality. In fact, *A* and *B* communicate if and only if *A* and *B* share a cognitive content as a result of *A* influencing *B*'s perception, understanding and interpretation and vice versa. Despite its little applicability in human conversation, Shannon and Weaver's model is useful to understand how communication works in terms of processes' states. This model can indeed be compared with the one described by Jakobson about the functions of language (Jakobson 1956). According to the author, in fact, the elements interacting in communication are i) the addresser, who sends a message to the addressee; ii) the message, which is connected and interpretable because of the presence of a context it can refer to; iii) a code, common to the addresser and addressee, used to codify the message; iv) a contact, which is the physical channel and the psychological connection between the addresser and the addressee, enabling both of them to enter and stay in communication. To each item of the communication circuit corresponds a specific language function.

Directly connected to communication is *dialogue*, seen as the prototypical form of language use and communicative exchange (Bazzanella 1994, 2005). Dialogue is a communicative process which requires two or more interlocutors, who coherently transmit pieces of information in one or more dialogue turns. The importance of focusing on such topics reflects the need to bridge the gap in the study and development of dialogue systems left by the lack of insights into the application of pragmatics to conversational agents. Although pragmatics is the level of language analysis strongly depending on dialogue, its computational application is mainly focused on the study and identification of speech acts (Leech 2003). In more detail, in the field of pragmatics, in the last ten years, research on Common Ground has seen a thriving impulse. Nevertheless, despite the fact that Clarification Requests are one of the grounding tools used by interlocutors while conversing, their study and application in dialogue systems have not yet seen a boost. All in all, a more in-depth analysis of pragmatic phenomena related to Common Ground construction and consistency checks in human-machine interaction, such as the

use of Clarification Requests, appears to be a missing spot in the research on dialogue system, and whose necessity needs to be confirmed in terms of efficiency increase with the support of the here presented study.

Clarification Requests are an important pragmatic device adopted to establish and maintain successful communication (Clark 1996; Allwood 1995). Among the different types of Clarification Requests, one class is used in specific contexts, namely when Common Ground Inconsistencies occur. With *Common Ground Inconsistencies* we refer to the incompatibility between the listener belief and the new evidence provided by the speaker. In other words, given a domain D , we define a set of sequential actions A as a number of different a_i . Each a_i is associated with a set of states S_i composed of verifiable pre-conditions s_{pre} and post-conditions s_{post} . D is inconsistent when an action a_i exists, associated with its S_i , where either s_{pre} and/or s_{post} are incompatible with respect to the S set belonging to another a_j , as they cannot co-exist. When this conflict takes place an inconsistency occurs. This conflict can depend on i) a s_{pre} which is incompatible with the rules of the Communal Common Ground (i.e., *cut the milk*) ii) the incompatibility of s_{pre} of the current a with s_{post} resulting from a preceding a , saved in the set of shared knowledge - the Personal Common Ground. Although both Common Ground Inconsistencies can cause corrective feedback, only the second type is linked to the adoption of Clarification Requests. As it will be described in the next section, polar questions are particularly important in these conflicting scenarios, since they clearly express the presuppositional stance of the listener when compared to other types of questions.

As far as clarification in dialogue is concerned, the act of clarifying succeeds the grounding request (CR) generated when facing understanding problems and constitutes an argumentation act (Traum 1994, p. 28). *Argumentation acts* are defined as “sequences of core speech acts, with constraints on the timing and content” (Traum 1999), i.e., an answer actually providing information asked for by the question. Concerning argumentation, there is a solid tradition in Artificial Intelligence concerning argumentation based inference starting with (Dung 1995), which formally described an abstract argumentation framework AF as a pair $(AR, attacks)$ where AR represents a set of arguments and $attacks$ is a binary relation in $AR \times AR$. Argumentation-based inference is a formal method for a single entity to decide about the truth of an argument and, therefore, does not consider the problems arising from dialogues among different agents.

Argumentation-based Dialogue (ABD) refers to the modelling of the verbal interaction aimed at the resolution of conflicts of opinions via the adoption of specific strategies. This field of study consists of a variety of different approaches and individual systems, with few unifying accounts or general frameworks (Prakken 2017).

In ABD, information is distributed among different agents, who may or may not be willing to share it at different points in time due to individual strategies and goals. This poses a problem both from the point of view of communication protocols, to ensure fairness and efficiency and from the point of view of behaviour. Adopting a goal-oriented perspective, dialogues have been classified as (Walton 1984; Walton and Krabbe 1995):

- Persuasion: aimed at solving a difference of opinion;
- Negotiation: aimed at solving a conflict of interest by reaching a deal;
- Information seeking: aimed at information exchange;

- Deliberation: aimed at reaching a decision or at establishing a course of action;
- Inquiry: aimed at growth of knowledge and agreement *per se*;
- Quarrel: aimed at winning a verbal fight or a contest.

Among the types of ABD, we concentrate on deliberation dialogue. Specifically, we consider the specific case of user-initiative dialogues where a human *leader* plans a series of operations to be later performed by an automatic *follower* whose only task is to check the consistency of the instructions sequence, very similarly to what happens, for example, in the map task (Baker and Hazan 2011). An important aspect of deliberation dialogue we focus on, in this paper, consists of the capability of the system to identify possible inconsistencies and to signal them with proper explanations.

2.1 Conflict-related Correcting Feedback in Conversational Agents

Classic approaches to ABD adopt the same setting that has been successfully used for argumentation based inference: that is, inference rules are derived to establish a course of action that is deterministic given a system configuration. Structural relationships among *claims* and various kinds of *replies* are established in a formal protocol dedicated to establishing whether a speech act is legal or not. This allows to provide a formal description of situations when a dialogue terminates or, in the case of competitive settings, is *won*. Since persuasion is the most studied situation in ABDs, a typical example of formal communication language is the one described in (Prakken 2005). In this type of setting, a *claim* provided by an agent *A* is supported by *data*, constituting an argument that can be explicitly put forward as a reply to a *why* move made by an agent *B*, which explicitly requests the speaker to explain the reasons why a statement should be accepted. Claims can be *attacked* by counter-arguments, which are other claims aimed at proving previous statements as false. *Conceding* and *retracting* moves respectively declare the acceptance of a statement or a change of attitude towards it, from commitment to non-commitment. Note that this does not imply a change of *belief*, as it is usually specified that the publicly declared position of an agent may not reflect what the agent actually believes.

An interesting result is found in the framework of deliberation dialogues, when collaboration is assumed on the task of finding an optimal solution to a problem for which none of the involved agents has a solution, yet. In the case of a two-agents system adhering strictly to the communication protocol, forming their claims on the basis of their knowledge bases and adopting a collaborative attitude, (Black and Atkinson 2010) demonstrated that the agreed solution is always acceptable to both parties. The usefulness of argumentation in dialogue systems designed for deliberation was, instead, demonstrated in (Kok et al. 2010).

The problem that characterises ABD with respect to argumentation based inference is the presence of different agents in the setting. This introduces multiple, not necessarily aligned, knowledge bases and, possibly, different/conflicting goals in the pursuit of a solution to a problem. There are attempts to deal with the partial knowledge each agent has concerning the others' goals and knowledge using rule-based systems: (Dunne and Bench-Capon 2006) examines the consequences of having *suspicious of hidden agenda* in the case of negotiation based dialogues while, in (Kok 2013), the strategic usefulness of reinforcing an agent's own claims versus the usefulness of undermining the other agents' claims is considered. These approaches, however, have been recently surpassed

by more flexible, probabilistic approaches, modelling opponents in terms of probability distributions over their possible beliefs and goals and using these to compute the utility of each legal dialogue move depending on their own goals and beliefs (e.g. (Hadjinikolis et al. 2013; Rienstra, Thimm, and Oren 2013)). Moreover, recent work put forward the need to model the *degree* or strength of an agent's belief towards a statement, modelled as the probability of the statement being true, rather than assuming it to either be or not be true (Hunter and Thimm 2016, 2017).

In this context, polar questions can serve as an argumentation tool. They usually encode in themselves not only a mere request but also presuppositions, agendas and preferences. Furthermore, when the questioner is closer to a K+ position, the use of a polar question can also implicate a disaffiliation. In this case, we refer to epistemically biased questions. According to the literature, one way of expressing disaffiliation is through the use of *Reversed Polarity Questions*, that are questions that convey bias towards the opposite valence than the utterance (Koshik 2002, 2005). For example, negative interrogatives can also function as positive assertions challenging the recipient's position (Heritage 2002). Criticisms and challenges can also be expressed through declaratives (i.e. *You shouldn't have done that*), imperatives (i.e. *Don't do that to me again*), or exclamations (i.e. *How dare you?*), which are perceived more confrontational and explicit and can be therefore face-threatening (Hayano 2013; Sidnell and Stivers 2012, 411). Among non-standard communications, conflicting representations (Huang 2017) are listed as interactions taking place when a discrepancy between what is communicated and what is believed by the agent occurs. In these scenarios, polar questions can, therefore, serve as a knowledge challenging tool.

Different authors pointed out how either the original bias of the speaker or the contextual evidence bias could influence the syntactic form of polar questions.

Original speaker bias. Belief or expectation of the speaker that p is true, based on his epistemic state prior to the current situational context and conversational exchange (Ladd 1981, 166).

Contextual evidence bias. Expectation that p is true (possibly contradicting a prior belief of the speaker) induced by evidence that has just become mutually available to the participants in the current discourse situation (Buring and Gunlogson 2000, 7).

Following (Domaneschi, Romero, and Braun 2017), possible combinations of the original bias of the speaker (where $B(p)$ is positive, $B(-)$ is neutral, and $B(\neg p)$ is negative) and the contextual evidence (where $E(p)$ is positive, $E(-)$ is neutral, and $E(\neg p)$ is negative) were investigated, in order to point out the influence they may have on the choice of polar question forms. This contrast represents, indeed, the conflict existing between the presupposed knowledge of the questioner and the one of the answerer.

The experiment illustrated in (Domaneschi, Romero, and Braun 2017; Di Maro, Origlia, and Cutugno 2021) pointed out the importance speakers give to the syntactic form with respect to the pragmatic needs. Results showed that the use of high negation polar questions better suits the pragmatic need of referring to a specific type of conflict between an original bias and an opposing contextual evidence. Namely, the conflict is between a strong presupposition of the speaker and a piece of information stored in the *Personal Common Ground* in a previous step of the interaction clashing with a contextual evidence given by the interlocutor. The same principles can, therefore, be applied when modelling human-machine dialogues. For this reason, even an apparently marginal difference, like the use of a negated form against its positive one, can express a specific speaker's stance and have a strong impact on the conversation efficiency.

Corrective dialogues are an important negotiation phase to build a coherent CG in the communication process. Human interlocutors always contribute with questions, answers, and feedback (Beun and van Eijk 2004): these are typical of corrective dialogues, occurring when : i) the user notices an error in the system and corrects it; ii) the user changes their mind; iii) the user's beliefs are in contradiction with the system's beliefs and expectations. Among these cases, only the third one is characterised by system initiative (Bousquet-Vernhettes, Privat, and Vigouroux 2003). For example, in (Beun and van Eijk 2004), the authors focused on a particular communicative problem related to conceptual discrepancies between a computer system and its user. In their final report, the authors stated that, although feedback is now used in such systems, there is still no accurate *mathematical theory* for natural communicative behaviours and their computational model to human-machine interaction, especially as far as conceptual discrepancies are concerned. What is still missing is, therefore, a reference model guiding the adoption of a specific type, content, and form of the feedback that has to be generated in a particular situation (Beun and van Eijk 2004).

In this work, a type of corrective dialogue is investigated, in which the system has a non-expert role and adjusts its grounded knowledge when conceptual discrepancies occur because of an inaccuracy, which causes an inconsistency, in the sequence of actions uttered by the user. Presenting a system architecture that includes the capability of detecting inconsistencies and reporting them to the user using adequate linguistic strategies is the goal of this work.

3. Related works

When pragmatics is applied to dialogue modelling, we talk about computational pragmatics, especially as far as the development of dialogue systems is concerned. In fact, computational pragmatics mostly deals with corpus data, context models, and algorithms for context-dependent utterance generation and interpretation (Huang 2017, p. 326). Nevertheless, conversational agents should be able not only to process local but also global structures of dialogues (Airenti, Bara, and Colombetti 1993). Whereas local structures are involved with linguistic rules (i.e., speech acts, turn-taking, etc.), which can be derived from corpus analysis, global structures refer to the conversation flow, that is the dialogue's action plan and how this is mutually known by dialogue participants (i.e., opening, closing, etc.). Cognitive pragmatics looks at these global structures derived from behavioural games, which in turn derive from grounding processes (Bara 1999). Different authors started including these processes in their dialogue systems architectures, especially as far as evaluating and updating common ground with their human partner, which is also the main topic of this work. For instance, Roque and Traum (Roque and Traum 2009) have developed a dialogue system that tracks grounded information in the previous conversation. As a consequence, the dialogue system is capable of selecting its utterances using different types of evidence of the user's understanding (i.e., whether the dialogue system has just submitted material or the user has also acknowledged it, repeated it back, or even used it in a subsequent utterance) (Müller, Paul, and Li 2021).

Using grounding strategies in conversational agents brought to interesting implementations. One aspect which has not yet been investigated is concerned with the mechanisms of grounding between humans and dialogue systems. Experimental investigations have mostly studied *how users evaluate the interaction, instead of studying interaction mechanisms* (Müller, Paul, and Li 2021, 3). For instance, Roque and Traum (Roque and Traum 2009) performed a user study in which subjects interacted with

their system and rated how much they felt the system understood them, put effort into understanding them, and gave appropriate responses. Conversely, what most studies do not ask is how a specific dialogue principle, such as the use of a particular type of request, is used by a system to affect user behaviours. Therefore, to learn more about human-machine dialogue mechanisms, it is important to turn to more basic experimental research (Müller, Paul, and Li 2021), like the one presented in this work.

The use of graph databases for dialogue systems, on the other hand, is also acquiring importance. In (Pichl et al. 2020), for example, an RDF-based conversational knowledge graph is used in the pipeline. Here, objective and subjective knowledge are represented. The advantage of using a graph database, like the one that is presented in this work, instead of an RDF structure, like the one used by the authors, lies in the fact that such databases are optimised for path search operations (i.e., the path that links the entity with a certain label to the action that caused the entity to get that specific label) and that they perform their operations in a much faster way. Others, such as (Axelsson and Skantze 2020), also adopt knowledge graphs, generated from Wikidata, connected to a behavioural tree that guide the grounding process of the items in the graph via feedback interpretation.

4. Methodology

In this Section, the system architecture proposed to implement Embodied Conversational Agents using graph databases as knowledge representation support to identify conflicting instructions is presented. Also, the materials used come from a previous experiment and were selected because they were found to be well balanced during the calibration phase. In this work, we submit the semantic representations of human commands, involved in the selected recipes, to the system to evaluate its conflict detection capabilities.

4.1 System architecture

The system presented here is intended as one of the possible applications of the framework FANTASIA by (Origlia et al. 2019), whose architecture is shown in Figure 1¹. FANTASIA's aim is to integrate different modules, such as a graph database, a dialogue manager, a game engine, and a voice synthesis engine for the development of social interactive systems. Integration efforts are, indeed, an important issue to overcome when a research group, for instance, shares the same theoretical framework but needs ad-hoc solutions for different applications. Approaches found in the literature to address this issue typically concentrated on communication layers, to which different actors in an interactive system must subscribe to exchange data. In such approaches, developing low-level code is still necessary to implement the application. Contrary to these, the high-level development languages provided by game engines, but also by other specialised solutions, offer an important chance to simplify the process when directly integrated in a proposed framework, as in FANTASIA.

The application of interest in this work is concerned with natural interaction. Specialised frameworks have dealt with this kind of interaction and focused mainly on virtual human management. In these frameworks, when game engines are adopted, they

¹ Figure 1 shows an improved version of the architecture of the one displayed in the reference paper (Origlia et al. 2019)

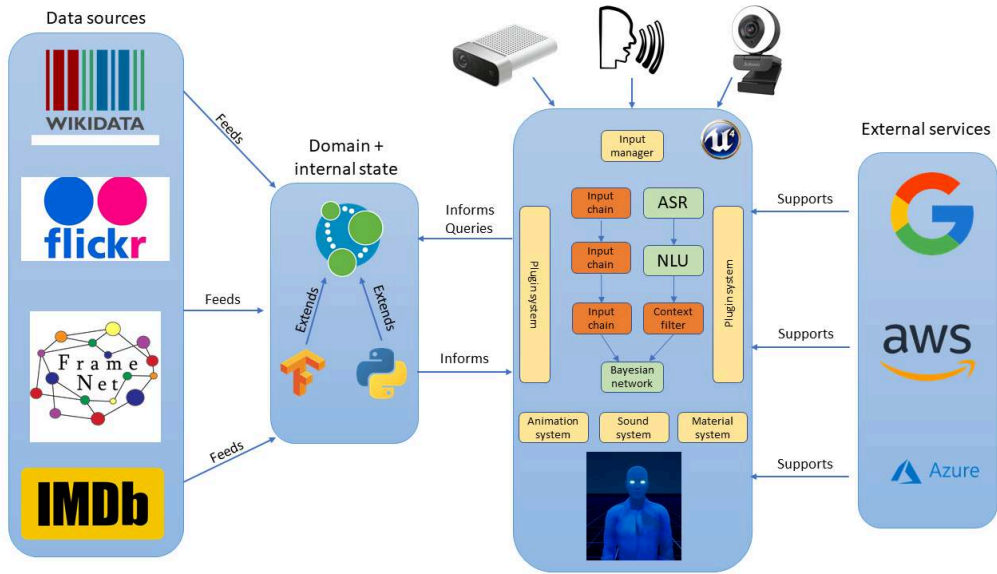


Figure 1

The FANTASIA architecture. Data coming from multiple sources, like Linked Open Data, are combined in a graph database, where further processing can be applied. Interaction management for Embodied Conversational Agents can be implemented in the Unreal Engine 4, also using third party AI services and multiple types of controllers.

have usually been used only as rendering modules. However, modern game engines are interesting candidates to host most of the behavioural logic and realisation modules in an integrated solution. In FANTASIA, as shown in Figure 1, a high industry-grade game engine such as the Unreal Engine 4 (UE4) is (Sanders 2016) adopted to control the virtual environment and an Embodied Conversational Agent. The engine manages communication with the human user, but it is also used to integrate language processing pipelines using informational data represented in graph format.

The knowledge base was represented in a graph database using Neo4j. Neo4j (Webber and Robinson 2018) is an open source graph database manager that has been developed over the last 16 years and applied to a high number of tasks related to data representation. It can be deployed in server mode and queried over a specific port using a standard HTTP or the dedicated Bolt protocol. It can also be embedded in Java applications through dedicated APIs. In Neo4j, nodes and relationships may be assigned *labels* that describe the type of object they are associated with. Neo4j is characterised by high scalability, ease of use and its proprietary query language, namely Cypher. Cypher is designed to be a *declarative* language that highlights patterns' structure using an SQL-inspired *ASCII-art syntax*. The increasing importance of graph databases is also pointed out in the *Gartner Top 10 Trends in Data and Analytics for 2020* where graph analytics and algorithms are considered important to improve AI and ML initiatives². Furthermore, The increasing importance of Neo4j is also demonstrated by the fact that this tool is able to detect conflicts and to use argumentation strategies to signal them consistently

² <https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020/> [last consultation on 19th January 2021]

with previous observations. This means that such graphs can be employed not only for rule-based reasoning but also for machine learning approaches.

Neo4j allows to combine data coming from different sources under a single, graph-based representation; for instance, sources of information other than textual and Linked-Open Data (LOD) can be integrated in the representation, like DBpedia and Wikidata, of interest in this work. The knowledge hosted by the aforementioned database is customisable according to the domain of application. In this work, the sources integrated in this tool are FrameNet (Baker, Fillmore, and Lowe 1998) and Wikidata³. Domains are indeed described through the set of basic actions extracted from FrameNet. Each domain element is, furthermore, represented with its characteristics retrieved from Wikidata. Wikidata serves as a human and machine-readable source containing structured data. The Wikidata project has become relevant, to the point that it is being employed as a connecting resource for many different dataset (e.g. the Thesauri collected from the Getty Research Institute, such as the Art & Architecture Thesaurus⁴ and the Library of Congress⁵).

The domain of interest chosen for this work is the cooking domain. Therefore, all structure-related explanations will be framed in this conceptual area. Details on the structure of the knowledge base, whose peculiarities are employed to search for conflicts, are given in the next section. Using this domain, pragmatic-related reasoning skills were implemented and tested, whose results are reported and discussed.

4.2 Materials

The materials used to test the conflict detection capabilities of the presented system architecture consist of a set of 10 recipes, extracted from the Italian cooking recipes website *GialloZafferano*⁶ and manually segmented into a series of steps, each corresponding to a single action. Although, currently, automatic systems capable of executing such tasks are only developed for research purposes⁷, it is reasonable to assume that most people know the basics of cooking and can therefore participate to our experiments. Actions and their involved parameters were annotated using FrameNet as a basis, so that each action is an instance of a frame and involve entities assume the role of frame elements. This way, steps identified in all recipes can be connected to a shared, standardised structure. This is enriched by adding pre-conditions, namely boolean checks to be performed on the PCG to verify its stability after accepting a new action, and post-conditions, namely updates to the PCG after a new action is accepted.

To simulate the occurrence of conflicting situations, for whose resolution a consistency recovery strategy had to be employed, an inconsistent action a_x was inserted in A . The inconsistency emerges when the pre-conditions of a later action are not verified because of post-conditions applied after accepting a_x . The conflicting inconsistency, representing a positive bias versus negative evidence contrast, was determined by the opposition of some aspects of a_x and some aspects of the consecutive a_n . The goal of the system, in this case, is to detect conflicts causing pre-conditions checks to fail and to identify the cause of the inconsistency in any previous action declared in the sequence. Actions causing conflicts, a_x , are found at variable distance from the action where the

3 https://www.wikidata.org/wiki/Wikidata:Main_Page [last consultation on 19th January 2021]

4 <https://www.getty.edu/research/tools/vocabularies/aat/> [last consultation on 19th January 2021]

5 <https://www.loc.gov/librarians/controlled-vocabularies/> [last consultation on 19th January 2021]

6 www.giallozafferano.it/

7 <http://www.rodyman.eu>

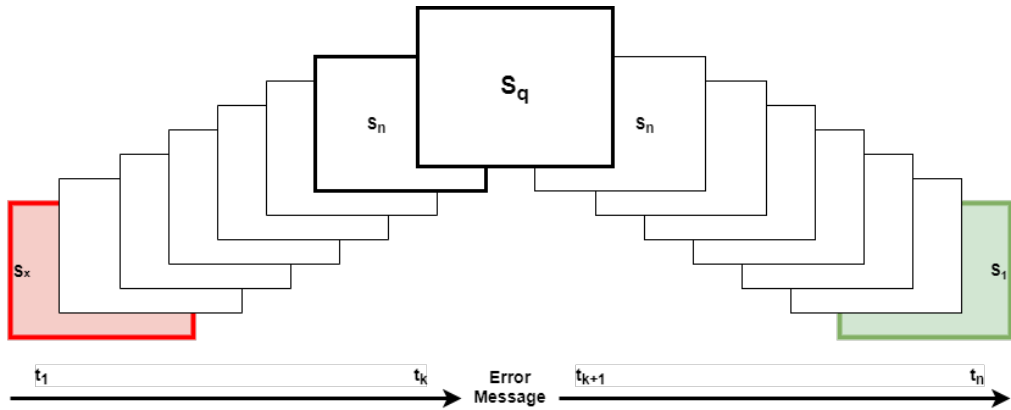


Figure 2
Experiment structure

conflict actually emerges, a_n , so that no assumption is made about how far in the past the conflict is rooted. Also, there are at least five actions between a_x and a_n so that the possibility that a_x is found in human subjects' short term memory, thus making the conflict easier to detect, will be reduced.

To evaluate how realistic the selected recipes are, a preliminary experiment was conducted to check if the chosen situations were either too easy or too difficult to detect for a human subject. Using a series of slides (Appendix C), $[s_1, \dots, s_n]$, for each recipe, which visually represented the steps involved in the recipe, we elicited spoken commands from a group of 36, gender balanced, subjects. While this was part of another experiment focusing on polar question forms, it allowed us to verify that the chosen sequences were understandable by human subjects, that it was possible for the artificially constructed conflicts to be detected by human participants and that none of the considered sequences was trivial. Regardless of the linguistic condition considered in the experiment, once the presence of the conflict was reported, participants could go back in the recipe in order to look for the conflict. The experiment, which made use of slides, was constructed in a way that, once the subject requested to go back after the prompt, the experimenter went instead forth, where the previous slides were presented backwards, as shown in Figure 2. Here, the conflicting slide was substituted with the correct one. This way, the identification of the conflict and the speaker's self-correction could be guided.

The goal of this experiment consisted of establishing whether a high negative polar question was more informative than a positive polar question to help the interlocutor to resolve a conflict.

A preliminary result of this experiment, presented in (Di Maro 2021), also relevant for this work, is that no recipe included in the considered dataset contained a conflict that was either too simple or too hard to identify. Table 1 shows that all conflicts were found, by human subjects, at least one time while no conflict was systematically detected. Given the results obtained during the calibration phase, in this work we used the same materials.

In addition to these recipes, 10 more were collected and annotated after the preceding dialogue modelling phase. This allowed to verify that the annotation process could

Table 1
Percentage of conflicts found per recipe

Recipe	Code	Conflict Found
Besciamella (<i>B��l'chamel</i>)	R01	66,67%
Carbonara	R02	50%
Cestini ripieni (<i>Oat yoghurt baskets</i>)	R03	33,33%
Crocchette di patate (<i>Potato croquettes</i>)	R04	50%
Pancakes	R05	75%
Patate al forno (<i>Baked potatoes</i>)	R06	66,67%
Piadina	R07	62,5%
Polpette (<i>Tuna meatballs</i>)	R08	50%
Tiramis��	R09	66,67%
Pizzette rosse (<i>Small pizzas</i>)	R10	33,33%

be applied to recipes different from the starting ones. An example recipe, divided into action, is reported here, with the conflicting input highlighted in bold.

```
Apply_heat Food:burro;Container:pentola;
Grinding Patient:noce moscata;
Cause_to_be_included New_member:noce moscata;Existing_member:burro;
Cause_to_be_included New_member:part#latte;Existing_member:burro;
Cause_to_amalgamate Parts:burro;
Cause_to_be_included New_member:farina;Existing_member:composto;
Cause_to_amalgamate Parts:composto;
Apply_heat Food:latte;Container:pentolino;
Cause_to_be_included New_member:noce moscata, sale;Existing_
member:latte;
```

In this case, the last action requires the nutmeg (noce moscata) to be added to the milk (latte). However, the nutmeg had already been added to the butter (burro), as the system assumes that, when no quantity is specified, all the available quantity of a named item is used. As a consequence, it is impossible to perform the last action because of the preceding one.

5. The Conflict Search Graph

In this work, we propose the use of a graph structure to represent state configurations at any time during a deliberation dialogue. Our model allows to represent dialogue history (i.e., the PCG) together with domain knowledge (i.e., the CCG), so that CG stability checks and dialogue state tracking can be represented in the form of graph queries. For the sake of simplicity, we assume that the items included in the CCG are known to both interlocutors but, in a wider view, the CCG only represents what an interlocutor *believes* to be known in the community they are part of.

From a formal point of view, dialogue states are defined by extending the concept of D as a sequence of actions, as presented in Section 2, to the joint representation of dialogue actions and domain knowledge, to support inconsistency detection. This is represented as a graph $D = \langle V, E \rangle$ where V is a set of vertices and E is a set of edges among the vertices in V . Edges are defined as functions between v_1 and v_2 where

$v_1, v_2 \in V$. The edge is assumed to be oriented from v_1 to v_2 . Vertices in V are divided in subgroups representing different roles in the CG:

- $A \subset V$ represents the set of actions incrementally added to the dialogue and accepted in the CG;
- $F \subset V$ is a set of *frames* describing available action *types* in the domain, their pre-conditions and their post-conditions;
- $I \subset V$ is a set of domain *items* whose features that are relevant for the domain are known;
- $L \subset V$ is a set of *frame elements* that describe the role domain items cover when involved in an action $a \in A$;
- $N \subset V$ is a set of named entities referring to items inside a specific action and assigning them to a specific frame element;
- $C \subset V$ is a set of *constraints* used to link *frame elements* to their admissible *items*

Edges in E are constituted by a series of two-parameter functions:

- $followed_by(a_i, a_j)$ states that $a_i \in A$ immediately precedes $a_j \in A$ in the sequence of actions accepted in the CG;
- $is_a(a_i, f_j)$ states that $a_i \in A$ implements the frame $f_j \in F$;
- $has_fe(f_i, l_j)$ states that $f_i \in F$ has a frame element $l_j \in E$;
- $names(a_i, n_j)$ states that $a_i \in A$ refers to the named entity $n_j \in N$;
- $assigned_to(n_i, l_j)$ states that the named entity $n_i \in N$ assigns the role described by $l_j \in L$;
- $refers_to(n_i, i_j)$ states that the named entity $n_i \in N$ refers to the item $i \in I$;
- $constrains(c_i, l_j)$ states that the frame element $l_j \in L$ can only accept specific items, linked to $c_i \in C$;
- $accepts(c_i, i_j)$ states that the constraint $c_i \in C$ allows the item $i_j \in I$ to be assigned to the frame elements c_i is linked to;

A *stable* CG is defined as a graph G where a set of stability checks, also based on frames pre-conditions, are all verified. For example:

$$check_m(G) \Leftrightarrow \forall c_i \in C, l_j \in L, n_k \in N, i_z \in I \mid \quad (1)$$

$$accepts(c_i, i_z) \implies constrains(c_i, l_j) \wedge assigned_to(n_k, l_j) \wedge refers_to(n_k, i_z)$$

states that, if an item is linked to a frame element through a named entity, then the item is also accepted by the constraints posed on the frame element. Therefore, G is considered stable using the following rule:

$$stable(G) \Leftrightarrow \forall check_i \mid check_i(G) \quad (2)$$

A new candidate action to be included in the CG can be defined as a tuple $X = \langle a_n, \bar{N}, \bar{E} \rangle$ containing a new action a_n , a set of named entities \bar{N} and a set of new edges \bar{E} . At any given time t , G_t represents the common ground configuration at t . Updating G by accepting X means creating a new graph $G' = \langle V', E' \rangle$ where $V' = V \cup a_n \cup \bar{N}$ and $E' = E \cup \bar{E}$. G' can be accepted as an updated version of G only if G' is stable, so that:

$$G_{t+1} = G' \text{ if } stable(G') \text{ else } G \quad (3)$$

Graph-based representations of the CG also allow the use of path-search queries to extract details about conflicts causing stability checks failures to guide the generation of confirmation requests. This theoretical model is implemented, in the presented system in the form of a *Conflict Search Graph*.

The *Conflict Search Graph* is the crucial module of the system, where knowledge is dynamically stored and checked during the interaction, and where reasoning processes occur. The aim of this module is to have a structured resource where the knowledge domain (i.e., part of the CCG) is stored, and whose conflict search module can be used to signal which input does not respect the rules of the CCG and cannot, therefore, become part of the PCG. In fact, the graph is not just used to represent the domain and its rules: it also supports the automatic process of recognising Common Ground Inconsistencies. Other than detecting unverified pre-conditions, the graph is used to store the dialogue history so that inconsistencies caused by post-conditions applied by previous actions let the system identify the potential source of the current inconsistency. Pre-conditions of an action describe, in general, the configurations of the CG that are compatible with action instancing. On the other hand, post-conditions are the resulting values assigned to an entity after the action has been processed. When a post-condition resulting from a previous action clashes with a pre-condition of the current action and inconsistency occurs. Whereas the pre-conditions make aware of the possible presence of a conflict, the post-conditions help identify the conflicting action. The check-related process guides the adoption of Clarification Requests.

The application described in this section implements a virtual agent, called Bastian, that accepts commands given in the cooking domain and checks their validity. To build the knowledge base of this application, two main resources were comprised, as previously introduced: Wikidata and FrameNet. From Wikidata, domain elements are retrieved to collect labels and characteristics of the single items involved in the cooking domain. From FrameNet, the set of basic actions involved in the domain is extracted and detailed to support the specific dialogue application. Here, the definition of the domain elements, expressed as SPARQL queries, is presented, together with the frames set and the connecting structure representing the dialogue domain specific for the application. For the cooking domain, represented in the application, specific frame elements were selected, such as semantic roles mainly conveyed by Ingredients, Tools and similar, and connected to Wikidata classes. Besides the data extracted from the aforementioned resources, additional information was added in the graph, namely pre-conditions and post-conditions of specific actions, as it will be illustrated. At the present, we rely on hard-coded rules to test out hypothesis, but data can be theoretically automatically

learned from structured data, like Wikipedia - now still incomplete, especially as far as pre- and post-conditions information are concerned. In this way, whereas from Wikidata not only Italian translation but also item states could be retrieved, from FrameNet action structures are derived. In addition, in the graph, these resources were combined and enriched with pre-conditions rules, as to represent the rule-based structure of the CCG. For example, as a first step, each element labelled as *Ingredient* was defined as an instance of a class descending from the concept *Food* (Q2095) in Wikidata. The set of items representing potential ingredients was obtained using Query 1, in Appendix A.

Subsequently, the tree-like structure rooted in *Food* was represented in Neo4j and Italian labels were recovered. These steps were performed in separated queries as the number of results was significantly high and timeout errors occurred at the endpoint in this situation. For the representation of other elements of the domain, *Tools* were defined as classes of objects descending from *Kitchen_Utensil* (Q3773693) as shown in Query 2, in the Appendix A.

Differently from the previous query, instances of classes were not considered as they cover specific objects, like single knives belonging to collections or commercial products. In addition, as the number of results of this query was lower, it was possible to obtain the Italian labels and the tree-like structure in a single query without risking timeout errors. Similarly, *Containers*, were defined as classes descending from the *Tableware* class (Q851782: glasses, plates, etc...), *Cooking Instruments* descended from the concept *Cookware_and_Bakeware* (Q154038: cooking pots, casseroles, etc...) while *Cooking appliances* descended from the concept *Cooking_Appliance* (Q57583712: stoves, ovens, etc...). In Neo4j, the relationships between Wikidata nodes reflect the original ones, as shown in Table 2. All imported nodes are provided with the Wikidata ID, the list of English labels, and the list of Italian ones.

Table 2
Neo4j nodes and relationships

Source node	Relationship	Destination Node
INGREDIENT_INSTANCE	BELONGS_TO	INGREDIENT_CLASS
INGREDIENT_CLASS	SUBCLASS_OF	INGREDIENT_CLASS
TOOL	SUBCLASS_OF	TOOL
CONTAINER	SUBCLASS_OF	CONTAINER
COOKING_APPLIANCE	SUBCLASS_OF	COOKING_APPLIANCE
COOKING_INSTRUMENT	SUBCLASS_OF	COOKING_INSTRUMENT

Concerning FrameNet, the entire structure of the resource was modelled in Neo4j following the same labels and relationships available in the original resource. To access the most recent version of FrameNet, online data were collected, rather than using periodic dumps. This was necessary because the dumps offer old versions of FrameNet with no updates. The main Neo4j labels representing the FrameNet structure are *FRAME*, and *FRAME_ELEMENT*, which were connected to each other by a *BELONGS_TO* relationship. For each *FRAME* and *FRAME_ELEMENT*, their name was imported, together with frame definitions and related examples.

Table 3

Structure of the sub-graph related to ACTIONs

Source node	Relationship	Destination Node
USER	DECLARES	ACTION
ACTION	IS_A	FRAME_INSTANCE
ACTION	REFERS_TO	ENTITY
ENTITY	REFERS_TO	PERCEIVED_ENTITY
ENTITY	ASSIGNED_TO	FRAME_ELEMENT

5.1 Domain specific knowledge representation

After organising the base resources in the database, the specific domain structure was established. This served both to connect the original resources and to represent the application-dependent dialogue constraints. First of all, the root of the application-specific domain was represented by a `DIALOGUE_DOMAIN` node, containing a *name* property to identify the domain. For each of the domain elements recovered from Wikidata, a `DOMAIN_ELEMENT` node was created, where a *name* property identifies the domain element. In the considered case, `DOMAIN_ELEMENT` nodes were *Ingredient Tool*, *Container*, *Cooking appliance* and *Cooking Instrument*. `DOMAIN_ELEMENT` nodes were connected to the `DIALOGUE_DOMAIN` node by `BELONGS_TO` relationships. `DOMAIN_ELEMENT` nodes were, then, connected to the Wikidata nodes retrieved using the presented SPARQL queries. As a result, the application-specific domain was connected to Wikidata.

Information coming from Natural Language Understanding and environment perception systems were defined in a specific way to allow standardisation of common ground consistency checks. In the case of deliberation dialogue, a `USER` node was defined for each human participant. One peculiarity of this kind of dialogue is that more than two agents can be involved in the exchange; that is also one of the reasons why argumentation-based inference theories cannot be always applied to dialogue and, therefore, a dedicated framework is needed. This node thus allows for the representation of each human interlocutor recognised by the systems. `ACTION` nodes represent declarations from a `USER`, which is connected to them by `DECLARES` relationships. Since `ACTION`s are always related to `FRAME_INSTANCES`, a `IS_A` relationship was established between `ACTION`s and `FRAME_INSTANCES` they represent. For each recognised `ACTION`, the linguistic entities recognised in the user utterance were represented by `ENTITY` nodes coherently with NLU responses. `ACTION`s were linked to `ENTITY` nodes by `REFERS_TO` relationships. Moreover, `ENTITY` nodes were linked to `FRAME_ELEMENT` nodes, according to the role NLU assigns to the recognised entities, by `ASSIGNED_TO` relationships. Lastly, objects perceived by the agent in the environment are represented by `PERCEIVED_ENTITY` nodes, which were linked to `DOMAIN_ELEMENT` nodes by `IS_A` relationships. The different types of node separating what is being said from what is perceived are necessary to support *grounding* approaches, where linguistic entities are linked to perceived objects. This also allows to detect inconsistencies between entities present in user utterances and perceived reality. In this case, a simple strategy based on string similarity was used to perform grounding, as the main interest is on conflict detection. The structure of the sub-graph related to `ACTION`s is shown in Table 3.

Once an ACTION is declared, the related ENTITY nodes are created and linked to the ACTION node by a REFERS_TO relationship. ENTITY nodes are then linked to the PERCEIVED_ENTITY nodes on the basis of the Sorensen-Dice coefficient (Sorensen 1948) obtained for every possible pairing between the value property of the ENTITY node and the name property of all the available PERCEIVED_ENTITY nodes. This way, plurals, derivative forms, or non-standards forms could be included to be linked to PERCEIVED_ENTITY nodes comprised in the knowledge graph. These are linked to the corresponding PERCEIVED_ENTITY nodes by the relation REFERS_TO. Nodes and relationships were generated using Query 3, in the Appendix A.

To connect the dialogue domain to FrameNet, a similar strategy was adopted. In total, 10 frames were used in the presented application: for each of these frames, a FRAME_INSTANCE node was created and connected to the original FRAME by an INSTANCE_OF relationship. Also, for each frame, a subset of FRAME_ELEMENT nodes was considered for the application domain. To represent this, a USES relationship was established between the FRAME_INSTANCE node and the FRAME_ELEMENT node of interest. To indicate which domain elements can be associated with a FRAME_ELEMENT in the application domain, CONSTRAINT nodes were established. First of all, FRAME_INSTANCE nodes were connected to CONSTRAINT nodes by a HAS_CONSTRAINT relationship. Then, the CONSTRAINT node was connected to the FRAME_ELEMENT node it was applied to by a REFERS_TO relationship and to a DOMAIN_ELEMENT node that can be associated to the FRAME_ELEMENT by another REFERS_TO relationship. CONSTRAINT nodes can, therefore, be used to describe which DOMAIN_ELEMENTS can be associated to fill a slot based on a FRAME_ELEMENT in a dialogue management system. While CONSTRAINT nodes are not relevant for conflict detection, they are included to support more advanced checks in the future.

Since Framenet does not provide pre-conditions and post-conditions for the application of the related actions, these must be defined at application level: in this case, pre- and post-conditions are represented as properties of the FRAME_INSTANCE nodes and contain Cypher queries designed to verify, given the way the specific application manages common ground updates, that the necessary checks are performed before accepting a user-declared action. To be interpreted by a single function, in the application logic, the results format is constrained to a table containing a row for each pre-condition to be tested. Each row consists of the following columns:

- Eval: the truth value of the pre-condition;
- ConflictingAction: the ID of the ACTION node causing a pre-condition to be violated, if present
- NLEExplanation: a fragment of text providing an explanation, in natural language, of the violated pre-condition;
- ConflictingFrame: the name property of the FRAME instanced by the FRAME_INSTANCE causing the conflict;
- OriginalEntity: the name property of the PERCEIVED_ENTITY involved in the ACTION causing the violation.

As a pre-condition example, consider the *Grinding* frame. As showed in Listing 4 in Appendix A, the FRAME_ELEMENT *Patient* is checked with the UNION of three separated sub-queries, each considering a different pre-condition, to verify that it is not

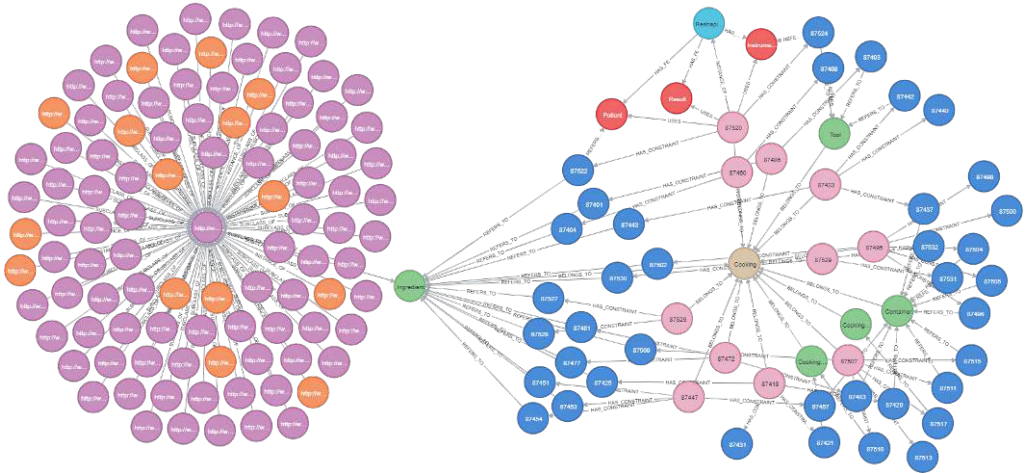


Figure 3
 The application level dialogue domain connecting Wikidata and FrameNet. The structure of the original resources is preserved in this schema while the dialogue domain structure and constraint inform the served application. Purple and orange nodes represent Wikidata instances and classes, green nodes represent DOMAIN_ELEMENTS, blue nodes represent CONSTRAINTs, red nodes represent FRAME_ELEMENTS, pink nodes represent FRAME_INSTANCES. For illustration purposes, only one FRAME node (in cyan) is reported. The brown node represent the DIALOGUE_DOMAIN node.

populated with an entity, whose quantity is no longer available, or with an entity which is is neither liquid or already in a powder form.

Running this query on a graph representing the common ground configuration is, thus, important to check whether the last ACTION can be accepted or not, in that it is verified that the updated graph does not violate the pre-conditions set by the activated FRAME_INSTANCE. Figure 3 shows the application level dialogue domain as an intermediate graph structure connecting the knowledge provided by Wikidata and FrameNet.

If all pre-conditions are verified, the declared ACTION can be accepted and post-conditions can be applied. For the case of the FRAME_INSTANCE related to the FRAME *Grinding*, the PERCEIVED_ENTITY related to the ENTITY assigned to the *Patient* FRAME_ELEMENT becomes a new version of itself, which acquires the POWDER label. The *Grinding* post-conditions are declared as in Listing 9 in Appendix A. The pre-conditions defined before would not be verified now, for the most recent version of the involved PERCEIVED_ENTITY. This is because it cannot be assigned to the *Patient* FRAME_ELEMENT for an ACTION related to the FRAME_INSTANCE referring to the FRAME *Grinding*. The Neo4j graph representing a user utterance and its role in the common ground is shown in Figure 4.

6. Conflict detection

To connect the internal knowledge representation hosted in Neo4j with the interaction management system implemented in UE4, the FANTASIA framework is used. To test the capability of the system to keep track of the dialogue state, commands are sent to the system one at a time. This way, the system can either accept or reject statements by



Figure 4

The graph representing the relationship between data coming from an NLU system in the common ground given the user utterance *Trita la noce moscata* (Grind the nutmeg). A USER (green) DECLARES an ACTION (purple), which IS_A FRAME_INSTANCE (pink) of the FRAME (cyan) *Grinding*. The ACTION REFERS_TO an ENTITY (grey), that is assigned to the FRAME_ELEMENT (red) *Patient* of *Grinding* and REFERS_TO a PERCEIVED_ENTITY (yellow). According to the *Grinding* post-conditions, a second PERCEIVED_ENTITY is CREATED_FROM the original one representing the *noce moscata*. The new PERCEIVED_ENTITY is also CREATED_BY the ACTION and it has the POWDER label.

updating the graph and rolling back changes, if necessary, by using graph projections in open database transactions to test pre-conditions. When a command is accepted, post-conditions are used to commit the transaction. The system used to test the conflict detection capabilities of the system can easily be extended to a fully interactive approach to involve human participants, in the future.

The Neo4j module provides access to the graph-based representation of the CG and to the dialogue history. UE4 manages the interaction using the 3D interface and the information provided by the other modules. UE4 also hosts the application logic, generating the virtual agent's behaviour using an underlying model based on the results presented before. To allow updates to the domain representation to be reflected in UE4, the system first queries the graph database to obtain the list of FRAME_INSTANCES and their CONSTRAINTs, dynamically initialising internal data structures to match the ones obtained from Neo4j. These are used in UE4 to support the creation of appropriate queries once user utterances are analysed. After obtaining a structured representation of the user's utterance from the an NLU backend, the CG manager matches the intents and entities detected by this module with, respectively, frames and FRAME_ELEMENTS, as described in the previous subsection. To simulate the process of *hypothesising* the situation after accepting the ACTION resulting from the analysis of the user utterance, the CG manager opens a transaction in Neo4j, adding the ACTION and its related structure without committing changes. This way, it is possible to work with a volatile version of the updated database that can be easily rolled back, should the ACTION be rejected. In this way, a *hypothetical common ground* is created to check for consistency based on the rules defined in the graph. Since multiple transactions can be opened in Neo4j, it is also possible, if necessary, to support the simultaneous existence of multiple *hypothetical*

common grounds. Pre-conditions are, therefore, checked inside the open transaction and the graph database compiles a report following the structure previously described. The CG manager, using this information, commits the changes together with post-conditions if all pre-conditions are verified and generates an acknowledgement utterance to be synthesised by the TTS system. If a pre-condition is not verified inside the transaction, the changes are rolled back and the data included in the Neo4j report are used to generate an appropriate feedback message: in this case, a negative polar question. In other words, given the sequence of frames activated by user utterances $F = \{f_1, \dots, f_k\}$, for each argument of the current predicate evoking a specific instantiated semantic frame f_k , and given the pre-conditions $s_{pre_k} = \{p_1, \dots, p_n\}$ of the k -th frame, when p_i of the semantic role of that argument is verified for $1 \leq i \leq n$, no conflicts arise.

If a conflict occurs, it must be signalled in order to enable subsequent repair. The fact that pre- and post-conditions are explicitly reported in the graph is not only useful to find the conflict, but also to explain why an action cannot be accepted, possibly indicating the source of the error. Before highlighting the conflicting action with a polar question, the system explains why the action cannot be performed. For instance, if the user asks the system to grind an ingredient which was already ground in a previous action, Bastian will reply with *I can't. X is ground* followed by the question *Didn't I have to grind X?* The data building the explanation are retrieved from a Cypher query and specifically from the aforementioned NLEExplanation column. The explanation given here is of the type *why*-explanation, which is used to convey the underlying, hidden reasons for an action or event (Stange and Kopp 2020). While explanations are found to increase the understandability and desirability of agents' behaviours (Stange and Kopp 2020), they can be cause of failures in case of inconsistencies. Although explanations are useful in the interaction, as they undo the devastating consequences of logical inconsistencies, they are not sufficient to detect the conflict (Khemlani and Johnson-Laird 2012). As demonstrated in (Domaneschi, Romero, and Braun 2017), the form used in verbal productions having the function of a Clarification Request is influenced by the type of conflict detected between bias and contextual evidence. The combination of both explanations and clarification requests can, therefore, consistently improve the interaction. If, on the other hand, the ACTION can be accepted, the NL feedback generated is a simple feedback with an *Acknowledgement* pragmatic function (Savy 2010). The system logic flow, as designed for a fully interactive agent, is summarised in Figure 5.

7. Results

Starting from the sequences of frames activated by actions depicted in the considered recipes, a dedicated task was used to test the conflict detection capabilities of the machine and its abilities to identify the sources of such conflicts. As reported in (Di Maro 2021), the level at which communicative failures can occur are of four different types, hierarchically ordered: Contact, Perception, Understanding, and Intention. When a problem at the contact level occurs, all the other levels fail, as they are entailed in the first one; when a problem does not occur at contact level, it can occur at the perception level, and the following ones are, therefore, failing too, and so on. Before analysing how the Common Ground is stored and how inconsistencies are found, it is important to point out what happens at the preceding levels, i.e., speech and intent recognition, where for the first one the acoustic signal is recognised, whereas for the second one the semantic analysis is carried out. For the goals of this study, we do not consider the potential communicative failures occurring at levels higher in the hierarchy presented in (Di Maro 2021). This is plausible because speech recognition and intent recognition modules have

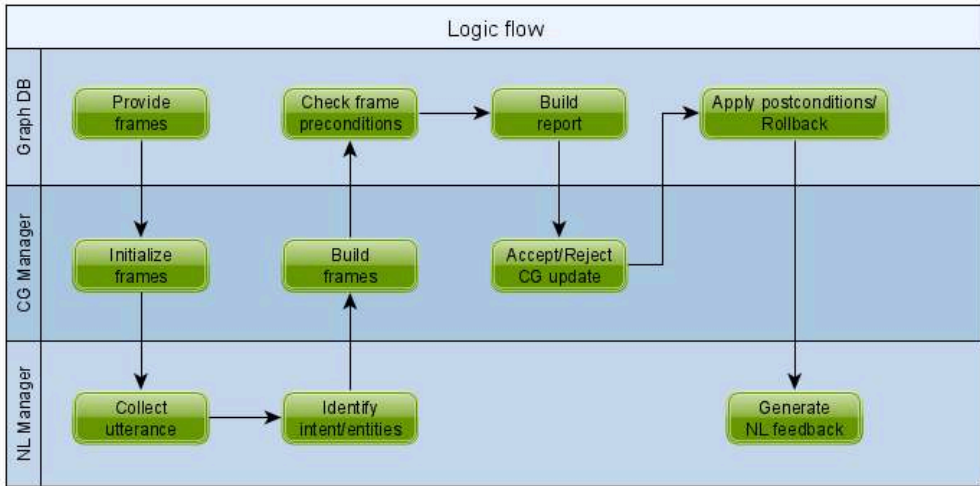


Figure 5
The logic flow in the cooking domain for a fully interactive agent.

reached good reliability. As a reference, on our materials, speech recognition reached a word error rate of 0,14, while intent recognition reached an average F-score of 0,74.

On the other hand, to test the Conflict Search Graph, for each actions sequence describing a recipe, the interpreted frames were submitted to the graph, iteratively. At each step, the system considers the PCG configuration should the last action be accepted and identifies the relevant pre-conditions. It, then, applies the selected pre-conditions and it verifies if the resulting graph is stable. If stability is verified, the graph is updated using the action’s post-conditions, otherwise changes are rejected and dialogue history is analysed to find a possible cause for the detected conflict. This process is included in the pre-conditions check, as offending patterns can be used to further detail the problem, as in traditional inference engines. From the application point of view, we specify that, while pre-conditions and post-conditions were specified at database level, the query to perform the stability checks and to recover the details was always the same. This represents an efficient way to separate application logic from PCG management.

In Table 5, the test results are displayed. The system always detected the conflicts and was able to correctly identify the conflict in most of the cases, with three exceptions, namely *Pancakes*, *Piadina Romagnola*, and *Polpettine*. In these cases, the conflict was detected but the expected conflicting action did not correspond to the one selected by the system. By analysing the errors, however, the system choices do have an acceptable explanation.

For the *Pancakes* recipe, the expected conflict corresponded to *melt butter in a pan*, where no quantity was specified although only part of the butter should have been used in this action. The conflict is triggered when the action *put the butter in the pan* is received in input, as the butter is no longer available. Nonetheless, the conflict was found at *add milk and butter to the yolks*. Although the error was to use the whole butter quantity in the action of melting it, it is also true that it actually becomes impossible to put the butter in the pan when this is added to other ingredients.

Similarly, in the *Piadina* recipe, the conflict was inserted by replacing *put part of the flour in the bowl* with *put the flour in the bowl*. The conflict is triggered when the operation

dust the work surface with flour is received in input, as the flour is no longer available. The system found the conflict at *add lard, salt, baking soda and little water to the flour*. As before, the only constraint required is the usability of flour, which stopped being usable after being mixed with other ingredients. Furthermore, it had not yet undergone any change of status. The next action, corresponding to *Cause_to_amalgamate*, is identified as the conflicting action because it is there that any possible reference to the flour is lost.

Finally, for the *Polpettine di tonno* recipe, the ingredient *ricotta* (*add parmesan, tuna, eggs, and anchovies to the ricotta*) was replaced by *breadcrumbs* (*add parmesan, tuna, eggs and anchovies to breadcrumbs*). The conflict was found by the system in the action where other ingredients were added to the *breadcrumbs*, making the breadcrumbs no longer available. This ingredient was, in fact, needed in a subsequent action, where meatballs had to be dunked in it.

Summarising, the presented results show that architecture based on the Conflict Search Graph was able to analyse pre-conditions rules correctly in a simulated scenario. In those cases where system responses were not equal to the ones expected at design time, response analysis indicated that, still, an acceptable logical explanation was provided by the system.

8. Conclusions

Dialogue systems' architectures designed for argumentation are often tailored on specific tasks, making the approaches harder to generalise and less oriented towards the definition of theoretical models of Argumentation Based Dialogue. These have been reported to be less investigated than the ones developed for Argumentation Based Inference. In this paper, we have proposed an architecture, based on the FANTASIA framework, leveraging on the capabilities of graph databases to store different kinds of information related to the Common Ground to support dialogue management tasks that involve argumentation features. We have shown a procedure to collect and organise data coming from widely accessible information sources and we integrated these data with a separated representation to manage application-specific knowledge. This way, both the domain and the incrementally built interaction history concur in determining how dialogue evolves using the same structure. Nevertheless, a separation between domain knowledge (the CCG) and the application specific knowledge (the PCG) is still present, so that the system is flexible and easily adaptable to new application domains.

From the client application level, the operational cycle is abstracted in a sequence of steps that do not depend on the characteristics of the applications itself: after an intent is recognised, stability checks can be performed using a general query retrieving and testing pre-conditions and returning a fixed structure, independent from the intent itself. Also, conflict details are retrieved as part of this process, similarly to what happens with inference engines. *Hypothesising* processes are managed using database transactions and commit/rollback mechanisms linked, when necessary, to post-conditions application. This answers the first research question by showing that graph databases indeed allow to represent, in a single, performance oriented, structure both dialogue state and CG.

To test the approach, we considered the specific case of *deliberation* dialogues and a specific type of conflict between previously acquired information (bias) and the implications of the last utterance (evidence). We have shown that specific conflict patterns in the dialogue domain (post-conditions of previously accepted actions colliding with the pre-conditions of incoming new actions) can be described in the form of path-search queries to the graph database, which always detected the conflicts and provided a *plausible* solution even in the cases where the obtained answer was different from the expected one.

This representation inherits the advantages coming from performance-oriented graph technologies while also providing many of the services offered by inference engines, thus constituting a powerful platform to develop a general view of Argumentation Based Dialogue using graph representations. The presented test constitutes the basis for argumentation-based dialogue systems centred on the concept of conflict detection for interaction management, providing indications for future developments aimed at fully answering the second research question.

References

- Airenti, Gabriella, Bruno Giuseppe Bara, and Marco Colombetti. 1993. Conversation and behavior games in the pragmatics of dialogue. *Cognitive Science*, 17(2):197–256.
- Allwood, Jens. 1995. An activity based approach to pragmatics. In *Abduction, Belief and Context in Dialogue*. John Benjamins.
- Allwood, Jens. 2013. A framework for studying human multimodal communication. *Coverbal synchrony in human-machine interaction*, 17.
- Axelsson, Nils and Gabriel Skantze. 2020. Using knowledge graphs and behaviour trees for feedback-aware presentation agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, Online, October.
- Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August.
- Baker, Rachel and Valerie Hazan. 2011. Diapixuk: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods*, 43(3):761–770.
- Bara, Bruno Giuseppe. 1999. *Pragmatica cognitiva: i processi mentali della comunicazione*. Bollati Boringhieri.
- Bazzanella, Carla. 1994. *Le facce del parlare. Un approccio pragmatico all'italiano parlato*, volume 17. La nuova Italia Collana: Biblioteca di Italiano e oltre.
- Bazzanella, Carla. 2005. *Linguistica e pragmatica del linguaggio. Un'introduzione*. Laterza.
- Becker, Tilman, Nate Blaylock, Ciprian Gerstenberger, Ivana Kruijff-Korbayová, Andreas Korthauer, Manfred Pinkal, Michael Pitz, Peter Poller, and Jan Schehl. 2006. Natural and intuitive multimodal dialogue for in-car applications: The sammie system. *Frontiers in Artificial Intelligence and Applications*, 141:612.
- Beun, Robbert-Jan and Rogier M. van Eijk. 2004. Conceptual discrepancies and feedback in human-computer interaction. In *Proceedings of the conference on Dutch directions in HCI*, page 13. Association for Computing Machinery, New York, NY, United States, June.
- Black, Elizabeth and Katie Atkinson. 2010. Agreeing what to do. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 12–30, Toronto, Canada, May. Springer.
- Bordes, Antoine, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.
- Bousquet-Vernhettes, Caroline, Régis Privat, and Nadine Vigouroux. 2003. Error handling in spoken dialogue systems: toward corrective dialogue. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Chateau d'Oex, Vaud, Switzerland, August.
- Buring, Daniel and Christine Gunlogson. 2000. Aren't positive and negative polar questions the same? In *Presented at Linguistic Society of America*.
- Clark, Eve V. 2015. Common ground. In *The Handbook of Language Emergence*. Wiley, Chichester, UK, pages 328–353.
- Clark, Herbert H. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.
- Clark, Herbert H. and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 222–233, Washington, DC, USA. American Psychological Association.
- Di Maro, Maria. 2021. "Shouldn't I use a polar question?" Proper question forms disentangling inconsistencies in dialogue systems. *Ph.D. Dissertation*.
- Di Maro, Maria, Antonio Origlia, and Francesco Cutugno. 2021. Polarexpress: Polar question forms expressing bias-evidence conflicts in Italian. *International Journal of Linguistics*.
- Domaneschi, Filippo, Maribel Romero, and Bettina Braun. 2017. Bias in polar questions: Evidence from English and German production experiments. *Glossa: a Journal of General*

- Linguistics*, 2(1).
- Dung, Phan Minh. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.
- Dunne, Paul E. and T.J.M. Bench-Capon. 2006. Suspicion of hidden agenda in persuasive argument. *Frontiers in Artificial Intelligence and Applications*, 144:329.
- Hadjinikolis, Christos, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter McBurney. 2013. Opponent modelling in persuasion dialogues. In *Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, China, August.
- Hayano, Kaoru. 2013. 19 question design in conversation. *The handbook of conversation analysis*, page 395.
- Heritage, John. 2002. The limits of questioning: Negative interrogatives and hostile question content. *Journal of Pragmatics*, 34(10-11):1427–1446.
- Huang, Yan. 2017. *The Oxford Handbook of Pragmatics*. Oxford University Press.
- Hunter, Anthony and Matthias Thimm. 2016. On partial information and contradictions in probabilistic abstract argumentation. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*, Cape Town, South Africa, April.
- Hunter, Anthony and Matthias Thimm. 2017. Probabilistic reasoning with abstract argumentation frameworks. *Journal of Artificial Intelligence Research*, 59:565–611.
- Jakobson, Roman. 1956. Metalanguage as a linguistic problem. *Selected writings*, 7:113–121.
- Khemlani, Sangeet S. and Philip N. Johnson-Laird. 2012. Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica*, 139(3):486–491.
- Kok, Eric M. 2013. *Exploring the practical benefits of argumentation in multi-agent deliberation*. Ph.D. thesis, Utrecht University.
- Kok, Eric M., John-Jules Ch Meyer, Henry Prakken, and Gerard AW Vreeswijk. 2010. A formal argumentation framework for deliberation dialogues. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 31–48, Toronto, Canada, May. Springer.
- Koshik, Irene. 2002. A conversation analytic study of yes/no questions which convey reversed polarity assertions. *Journal of Pragmatics*, 34(12):1851–1877.
- Koshik, Irene. 2005. *Beyond rhetorical questions: Assertive questions in everyday interaction*, volume 16. John Benjamins Publishing.
- Kousidis, Spyros, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. 2014. A multimodal in-car dialogue system that tracks the driver’s attention. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 26–33, Istanbul, Turkey, November. ACM.
- Ladd, Dwight Robert. 1981. A first look at the semantics and pragmatics of negative questions and tag questions. In *Papers from the Regional Meeting. Chicago Ling. Soc. Chicago, Ill*, volume 17, pages 164–171.
- Leech, Geoffrey. 2003. Pragmatics and dialogue. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.
- Legg, Shane and Marcus Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds Mach.*, 17(4):391–444, December.
- López, Gustavo, Luis Quesada, and Luis A. Guerrero. 2017. Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*, pages 241–250, Los Angeles, USA, July. Springer.
- McGlashan, Scott, Norman Fraser, Nigel Gilbert, Eric Bilange, Paul Heisterkamp, and Nick Youd. 1992. Dialogue management for telephone information systems. In *Proceedings of the third conference on Applied natural language processing*, pages 245–246, Trento, Italy, 31 March - 3 April. Association for Computational Linguistics.
- Müller, Romy, Dennis Paul, and Yijun Li. 2021. Reformulation of symptom descriptions in dialogue systems for fault diagnosis: How to ask for clarification? *International Journal of Human-Computer Studies*, 145:102516.
- Origlia, Antonio, Francesco Cutugno, Antonio Rodà, Piero Cosi, and Claudio Zmarich. 2019. Fantasia: a framework for advanced natural tools and applications in social, interactive approaches. *Multimedia Tools and Applications*, 78(10):13613–13648.
- Pichl, Jan, Petr Marek, Jakub Konrád, Petr Lorenc, Van Duy Ta, and Jan Šedivý. 2020. Alquist 3.0: Alexa prize bot using conversational knowledge graph. *3rd Proceedings of Alexa Prize*.
- Prakken, Henry. 2005. Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, 15(6):1009–1040.

- Prakken, Henry. 2017. Historical overview of formal argumentation. *IfCoLog Journal of Logics and their Applications*, 4(8):2183–2262.
- Rienstra, Tjitze, Matthias Thimm, and Nir Oren. 2013. Opponent models with uncertainty for strategic argumentation. In *Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, China, August.
- Ritter, Alan, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California, June. Association for Computational Linguistics.
- Roque, Antonio and David Traum. 2009. Improving a virtual human using a model of degrees of grounding. In *Twenty-First International Joint Conference on Artificial Intelligence*, Pasadena, California, USA, July. Citeseer.
- Sanders, Andrew. 2016. *An introduction to Unreal engine 4*. CRC Press.
- Savy, Renata. 2010. Pr. A. Ti. D: A coding scheme for pragmatic annotation of dialogues. In *The seventh international conference on Language Resources and Evaluation*, Malta, May.
- Serban, Iulian Vlad, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.
- Serban, Iulian Vlad, Alessandro Sordani, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Conference of the Association for the Advancement of Artificial Intelligence*, volume 16, pages 3776–3784, Phoenix, Arizona, USA, February.
- Shannon, Claude Elwood. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Sidnell, Jack and Tanya Stivers. 2012. *The Handbook of Conversation Analysis*, volume 121. John Wiley & Sons.
- Sorensen, Thorvald. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34.
- Sperber, Dan et al. 1994. Understanding verbal understanding. *What is intelligence*, 179:98.
- Stange, Sonja and Stefan Kopp. 2020. Effects of a social robot’s self-explanations on how humans understand and evaluate its behavior. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 619–627, Cambridge, United Kingdom, March.
- Traum, David R. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.
- Traum, David R. 1999. Speech acts for dialogue agents. In *Foundations of rational agency*. Springer, pages 169–201.
- Vinyals, Oriol and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.
- Walton, Douglas and Erik C.W. Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.
- Walton, Douglas N. 1984. *Logical Dialogue-Games*. University Press of America, Lanham, Maryland.
- Warner, Michael. 2002. Wanted: A definition of intelligence. *Studies in Intelligence*, 46:9, 01.
- Webber, Jim and Ian Robinson. 2018. *A pragmatic introduction to neo4j*. Addison-Wesley Professional.

Appendix A

```

SELECT DISTINCT ?item ?itemLabel (group_concat(DISTINCT
?altEN;separator="|") as ?altENs) ?type
{
  {
    ?item wdt:P31 ?class .
    ?class wdt:P279* wd:Q2095 .
    ?item rdfs:label ?itemLabel .

    FILTER(LANG(?itemLabel) = "en")

    OPTIONAL{
      ?item skos:altLabel ?altEN.
      FILTER (lang(?altEN) = "en")
    }

    BIND("instance" AS ?type)
  }
  UNION
  {
    ?item wdt:P279* wd:Q2095 .
    ?item rdfs:label ?itemLabel .

    FILTER(LANG(?itemLabel) = "en")

    OPTIONAL{
      ?item skos:altLabel ?altEN.
      FILTER (lang(?altEN) = "en")
    }
    BIND("class" AS ?type)
  }
}
GROUP BY ?item ?itemLabel ?altENs ?type

```

Listing 1

SPARQL query used to retrieve the set of possible ingredients from Wikidata.

```

SELECT ?item ?parent ?itLabel ?enLabel
(group_concat(DISTINCT ?altEN;separator="|") as ?altENs)
(group_concat(DISTINCT ?altIT;separator="|") as ?altITs) {
?item wdt:P279* wd:Q3773693.
?item wdt:P279 ?parent.
?parent wdt:P279* wd:Q3773693.

OPTIONAL {
  ?item rdfs:label ?enLabel .
  FILTER(LANG(?enLabel) = "en")
}

OPTIONAL {
  ?item rdfs:label ?itLabel .
  FILTER(LANG(?itLabel) = "it")
}

FILTER ( bound(?itLabel) || bound(?enLabel) )

OPTIONAL{
  ?item skos:altLabel ?altEN.

```

```

    FILTER (lang(?altEN) = "en")
  }

  OPTIONAL{
    ?item skos:altLabel ?altIT.
    FILTER (lang(?altIT) = "it")
  }
}
GROUP BY ?item ?parent ?itLabel ?enLabel ?altENs ?altITs

```

Listing 2

SPARQL query used to retrieve tools from Wikidata.

```

MATCH (a:ACTION) WHERE NOT (a)-[:IS_FOLLOWED_BY]->()
WITH a
MATCH (pe1:PERCEIVED_ENTITY), (e:ENTITY)-[:REFERS_TO]- (a)
OPTIONAL MATCH (pe1)-[:CREATED_FROM]- (pe2:PERCEIVED_ENTITY)
WITH pe1, a, e, pe2, COLLECT(pe2)[0] AS successor
WHERE successor IS NULL OR NOT successor.name = pe1.name
UNWIND split(apoc.text.replace(e.value, "\\[[\\.\\d]+\\]", ""), ",") AS
  names
WITH pe1.name AS name, COLLECT(names) AS names, apoc.text.
  soresenDiceSimilarity(names, pe1.name) AS score, a
WITH MAX(score) as maxValue, a
MATCH (pe1:PERCEIVED_ENTITY), (e:ENTITY)-[:REFERS_TO]- (a)
OPTIONAL MATCH (pe1)-[:CREATED_FROM]- (pe2:PERCEIVED_ENTITY)
WITH maxValue, pe1, a, e, pe2, COLLECT(pe2)[0] AS successor WHERE
  successor IS NULL OR NOT successor.name = pe1.name UNWIND split(
  apoc.text.replace(e.value, "\\[[\\.\\d]+\\]", ""), ",") AS names
WITH pe1.name AS bestMatch, COLLECT(names) AS names, COLLECT(apoc.text.
  soresenDiceSimilarity(names, pe1.name)) AS score, maxValue, a
WITH bestMatch, apoc.coll.zip(names, score) AS pairs, maxValue, a
WITH bestMatch, MAX([pair IN pairs WHERE pair[1] = maxValue])[0][0] AS
  entityName, a
WHERE entityName IS NOT NULL
WITH entityName, bestMatch, a
MATCH (pe1:PERCEIVED_ENTITY), (e:ENTITY)-[:REFERS_TO]- (a) WHERE pe1.
  name = bestMatch AND e.value CONTAINS(entityName)
OPTIONAL MATCH (pe1)-[:CREATED_FROM]- (pe2:PERCEIVED_ENTITY)
WITH entityName, a, pe1, e, pe2, COLLECT(pe2)[0] AS successor WHERE
  successor IS NULL OR NOT successor.name = pe1.name
CREATE (pe1)-[:REFERS_TO {label: entityName}]- (e)

```

Listing 3

Cypher query linking linguistic ENTITY nodes to the corresponding PERCEIVED_ENTITY nodes after NLU

```

// Pre-conditions for Grinding
//Condition 1: Verify that there is enough of the involved
  PERCEIVED_ELEMENTS to perform the ACTION
// Get the last ACTION, the ENTITY nodes it refers to, the
  PERCEIVED_ELEMENTS they REFER_TO
// and the FRAME_ELEMENTS of type "Patient" ENTITY node are ASSIGNED_TO
.
MATCH (a1:ACTION)-[:REFERS_TO]->(e:ENTITY)-[:ASSIGNED_TO]->(fe:
  FRAME_ELEMENT),
(e)-[r1:REFERS_TO]->(pe1:PERCEIVED_ENTITY)
WHERE NOT (a1)-[:IS_FOLLOWED_BY]->() AND fe.name IN ['Patient']

```



```

// If available, get the PERCEIVED_ENTITY nodes CREATED_FROM each
  PERCEIVED_ENTITY
// ENTITY nodes REFER_TO.
OPTIONAL MATCH (pe1) <- [r2:CREATED_FROM] - (pe2:PERCEIVED_ENTITY)

// Compute the PERCEIVED_ELEMENTS quantity used by the last ACTION. 0
  means "all the available quantity".
// If the available quantity is infinite, default to 1 to avoid
  Infinity - Infinity = NaN
WITH pe1, r2, a1,
CASE
  WHEN toFloat(apoc.text.regexGroups(e.value, r1.label + "\{(\d+)\}")
    [0][1]) = 0 AND gds.util.isFinite(pe1.quantity) THEN pe1.quantity -
    SUM(r2.quantity)
  WHEN toFloat(apoc.text.regexGroups(e.value, r1.label + "\{(\d+)\}")
    [0][1]) = 0 AND gds.util.isInfinite(pe1.quantity) THEN 1
  ELSE toFloat(apoc.text.regexGroups(e.value, r1.label + "\{(\d+)\}")
    [0][1])
END AS newQuantity

// If the available quantity is more than 0 and subtracting the
  declared quantity is at least 0 the pre-condition is verified
WITH pe1.quantity - SUM(r2.quantity) - newQuantity >= 0 AND pe1.
  quantity - SUM(r2.quantity) > 0 AS Eval,

// Builds the explanation concatenating "Non ho abbastanza" with the
  label of the insufficient PERCEIVED_ENTITY
"Non ho abbastanza " + pe1.name + ". " AS NLEExplanation, pe1, a1

// If the conflict is caused by a preceding ACTION, get the necessary
  data to build the HNPQ
// (ID of the conflicting ACTION, name of the conflicting FRAME, list
  of Ingredients involved in the conflicting ACTION)
OPTIONAL MATCH (pe1) <- [:CREATED_FROM] - (pe2:PERCEIVED_ENTITY) - [:
  CREATED_BY] -> (a2:ACTION) - [:REFERS_TO] -> (:ENTITY) - [:REFERS_TO] -> (pe3
  :PERCEIVED_ENTITY),
(a2:ACTION) - [:IS_A] -> (:FRAME_INSTANCE) - [:INSTANCE_OF] -> (f:FRAME) RETURN
  Eval,
COLLECT(ID(a2)) [0] AS ConflictingAction,
NLEExplanation,
COLLECT(f.name) [0] AS ConflictingFrame,
apoc.text.join(COLLECT(DISTINCT pe3.name), ", ") AS OriginalEntity

//Condition 2: Verify that the involved PERCEIVED_ELEMENT is not a
  POWDER
UNION
// Get the last ACTION, the ENTITY nodes it refers to, the
  PERCEIVED_ELEMENTS they REFER_TO and having the POWDER label
// and the FRAME_ELEMENTS of type "Patient" ENTITY node are ASSIGNED_TO
.
MATCH (a1:ACTION) - [:REFERS_TO] -> (e:ENTITY) - [:ASSIGNED_TO] -> (fe:
  FRAME_ELEMENT {name: 'Patient'}),
(e) - [:REFERS_TO] -> (pe1:PERCEIVED_ENTITY)
WHERE NOT (a1) - [:IS_FOLLOWED_BY] -> () AND 'POWDER' IN labels(pe1)

// If at least one PERCEIVED_ELEMENT with the POWDER label is found,
  the pre-condition is not verified
WITH NOT COUNT(*) > 0 AS Eval

```

```

// If available, find a preceding version of the POWDER
  PERCEIVED_ELEMENT that did not have the POWDER label
MATCH (a1:ACTION)-[:REFERS_TO]->(e:ENTITY)-[:ASSIGNED_TO]->(fe:
  FRAME_ELEMENT {name: 'Patient'}),
(e)-[:REFERS_TO]->(pe1:PERCEIVED_ENTITY)
WHERE NOT (a1)-[:IS_FOLLOWED_BY]->()
WITH Eval, pe1, a1
OPTIONAL MATCH (pe1)-[:CREATED_FROM*]->(pe2:PERCEIVED_ENTITY)<-[:
  REFERS_TO]-(:ENTITY)<-[:REFERS_TO]->(a2:ACTION)-[:IS_A]->(:
  FRAME_INSTANCE)-[:INSTANCE_OF]->(f:FRAME)
WHERE a1 <> a2 AND NOT 'POWDER' IN labels(pe2)

// Return the necessary information to build the HNPQ if a previous
  ACTION caused the PERCEIVED_ENTITY to acquire the POWDER label
RETURN Eval, COLLECT(ID(a2))[0] AS ConflictingAction,
pe1.name + ' Ã in polvere.' AS NLEExplanation,
COLLECT(f.name)[0] AS ConflictingFrame,
COLLECT(pe2.name)[0] AS OriginalEntity

//Condition 3: Verify that the involved PERCEIVED_ELEMENT is not a
  LIQUID
UNION
// Get the last ACTION, the ENTITY nodes it refers to, the
  PERCEIVED_ELEMENTS they REFER_TO and having the LIQUID label
// and the FRAME_ELEMENTS of type "Patient" ENTITY node are ASSIGNED_TO
.
MATCH (a1:ACTION)-[:REFERS_TO]->(e:ENTITY)-[:ASSIGNED_TO]->(fe:
  FRAME_ELEMENT {name: 'Patient'}),
(e)-[:REFERS_TO]->(pe1:PERCEIVED_ENTITY)
WHERE NOT (a1)-[:IS_FOLLOWED_BY]->() AND 'LIQUID' IN labels(pe1)

// If at least one PERCEIVED_ELEMENT with the LIQUID label is found,
  the pre-condition is not verified
WITH NOT COUNT(*) > 0 AS Eval

// If available, find a preceding version of the POWDER
  PERCEIVED_ELEMENT that did not have the LIQUID label
MATCH (a1:ACTION)-[:REFERS_TO]->(e:ENTITY)-[:ASSIGNED_TO]->(fe:
  FRAME_ELEMENT {name: 'Patient'}),
(e)-[:REFERS_TO]->(pe1:PERCEIVED_ENTITY)
WHERE NOT (a1)-[:IS_FOLLOWED_BY]->()
WITH Eval, pe1, a1
OPTIONAL MATCH (pe1)-[:CREATED_FROM*]->(pe2:PERCEIVED_ENTITY)-[:
  CREATED_BY]->(a2:ACTION)-[:IS_A]->(:FRAME_INSTANCE)-[:INSTANCE_OF
  ]->(f:FRAME)
WHERE a1 <> a2 AND NOT 'LIQUID' IN labels(pe2)

// Return the necessary information to build the HNPQ if a previous
  ACTION caused the PERCEIVED_ENTITY to acquire the LIQUID label
RETURN Eval, COLLECT(ID(a2))[0] AS ConflictingAction, pe1.name + ' Ã
  un liquido.' AS NLEExplanation,
COLLECT(f.name)[0] AS ConflictingFrame, COLLECT(pe2.name)[0] AS
  OriginalEntity

```

Listing 4

Cypher query checking the pre-conditions of the Grinding frame.

Appendix B





















Table 4
Frame structures with features and sub-entities.

	Frame elements	Sub-Entities	Features		Frame elements	Sub-Entities	Features
Apply heat	Temperature setting	-	Temperature	Cause to amalgamate	Parts	-	Ingredient
	Heating instrument	-	Cooking appliance		Whole	-	Ingredient
	Food	-	Container		Means	-	Tool
			Ingredient		-	Container	
	Container	-	Cooking instrument		Place	-	Cooking instrument
Duration	-	Duration					
Cause to be included	Existing member	-	Ingredient	Cutting	Item	-	Ingredient
	Place	-	Container		Pieces	Quantity	Number
	New member	-	Ingredient			Size	Dimension
						Shape	-
Group	-	Ingredient	Instrument	Size	Dimension		
Dunking	Substance	-	Container	Grinding	Instrument	-	Tool
		-	Ingredient			Patient	-
	Theme	-	Ingredient				
Placing	Theme	-	Container	Removing	Source	-	Container
		-	Cooking instrument			-	Cooking appliance
		-	Ingredient			-	Cooking instrument
	Area	-	Container		-	Container	
	Source	-	Cooking appliance		Theme	-	Cooking instrument
			Cooking instrument				
	Means	-	Tool				
Duration	-	Duration					
Reshaping	Instrument	-	Tool	Separating	Whole	-	Ingredient
	Patient	-	Ingredient		Parts	-	Ingredient
	Result	-	-		Instrument	-	Tool
			Place		-	Container	

Table 5
Conflict Search Graph Results and Outcomes

Recipe	Result	Expected Result	Outcome
Besciamella	3	3	OK
Carbonara	10	10	OK
Cestini ripi- eni	7	7	OK
Crocchette	5	5	OK
Pancakes	5	1	KO
Patate al forno	4	4	OK
Piadina romagnola	2	1	KO
Pizzette rosse	3	3	OK
Polpettine di tonno	5	4	KO
Tiramisú	6	6	OK
Gnocchi	6	6	OK
Guacamole	6	6	OK
Hamburger di ceci	5	5	OK
Mousse al cioccolato	9	9	OK
Plumcake	1	1	OK
Polpette di zucchine	5	5	OK
Sformato di verdure	7	7	OK
Torta Tener- ina	6	6	OK
Zucchine alla scapece	4	4	OK
Zuppa	7	7	OK

Appendix C

<p>Action</p> 	<p>Parameters</p>  <p>On a low heat</p>	<p>Action</p> 	<p>Parameters</p> 
<p>Action</p> 	<p>Parameters</p> 	<p>Action</p> 	<p>Parameters</p> 
<p>Action</p> 	<p>Parameters</p> 	<p>Action</p> 	<p>Parameters</p> 
<p>Action</p> 	<p>Parameters</p> 	<p>Action</p> 	<p>Parameters</p> 
<p>Action</p> 	<p>Parameters</p> 	<p>Action</p> 	<p>Parameters</p> 

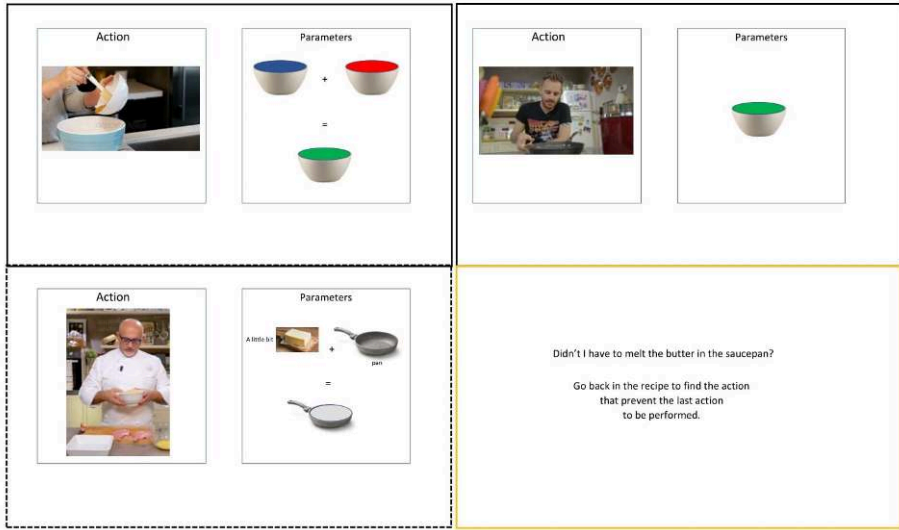


Figure 6

Slides representing the sequence of actions for the Pancake recipe. The conflicting quantity-related action (*Melt the butter in the saucepan*) is in red; the action that cannot be performed because of the conflicting one (*Put a little bit of butter in the pan*) is surrounded by dashed lines; the conflict is signalled with a High Negation Polar Question in the last slide (*Didn't I have to melt the butter in the saucepan?*)

Toward a linguistically grounded dialog model for chatbot design

Anna Dall'Acqua*
Università di Bologna & Injenia S.r.l.

Fabio Tamburini**
Università di Bologna

The increasing interest in various types of conversational interfaces has been supported by a progressive standardization of the technological frameworks used to build them. However, the landscape of available methodological frameworks for designing conversations is much more fragmented. We propose a highly generalizable methodology for designing conversational flows rooted in a functionalist-pragmatics perspective, with an explicit adherence to a conversationalist approach. In parallel, we elaborate a practical-procedural workflow for undertaking chatbots projects in which we situate the theoretical starting point. At last, we elaborate a general case-study on which we transpose the identified approach in Italian language and using one of the most authoritative NLU platforms.

1. Introduction

One of the most exciting innovations that we are experiencing in the last decade is the massive widespread of conversational interfaces, such as chatbots or virtual assistants (Tsvetkova et al. 2017; Chaves et al. 2019; Dale 2016). The various attempts that have been made to classify these technologies (Radziwill and Benton 2017; Følstad, Skjuve, and Brandtzaeg 2019; Hussain, Sianaki, and Ababneh 2019; Mathur and Singh 2018) and the absence of an unequivocal taxonomy (Braun and Matthes 2019) surely contribute to the lack of a methodological approach for designing conversational agents. They are perceived as something in between humans and web search engines characterised by a conversational way of expression and the capability of managing input and output in natural language (Dale 2016; Braun and Matthes 2019).

We are witnessing a flourishing literature about technologies, techniques and applications for building conversational interfaces (Ahmad et al. 2018; Adamopoulou and Moussiades 2020). Unfortunately, we cannot say the same for the elaboration of methodological guidelines that can be pursued for the designing of conversational interfaces, especially from a linguistic point of view.

This is the context in which our research comes to light. We think that a solid anchor in linguistics and therefore a scientific knowledge of what human conversation is may be the key for identifying a generalizable methodological approach for designing conversational agents.

* Dept. of Classical Philology and Italian Studies, University of Bologna, Italy and Injenia S.r.l., Bologna, Italy. E-mail: anna.dallacqua2@unibo.it

** Dept. of Classical Philology and Italian Studies, University of Bologna, Italy.
E-mail: fabio.tamburini@unibo.it. Corresponding author.

1.1 A methodological lack in conversation design

A relevant set of studies have been produced on this topic. The first agents were implemented with simple *pattern-matching* techniques and *template-based* responses (Weizenbaum 1966; Colby, Weber, and Hilf 1971). They could support a continuative concatenation of utterances but they were still far away from today's state of the art. The same pattern recognition model shaped the architecture of A.L.I.C.E., a chatbot annotated with AIML, a mark-up language derived from the metalanguage XML (Wallace 2003; Shawar and Atwell 2007). After 2016 there has been a growing interest for chatbots in various areas and applications, mostly because they were considered new productive and entertaining objects not reducible as mere assistants, but capable of a way of interacting that brings them closer to users (Dale 2016).

Today, most of them rely on machine learning algorithms and Natural Language Understanding modules, but still even the more conversational of the agents can just vaguely simulate conversational exchanges enriched by mutual understanding as we know it as humans.

Adamopoulou and Moussiades (2020) distinguish between two ways of developing chatbots: using any programming language like Java, Clojure, Python, C++, PHP, Ruby and Lisp or using state-of-the-art platforms. At this time, the leading NLU cloud platforms supported by machine learning are: Google's Dialogflow in both versions ES and CX, Facebook's wit.ai, Microsoft LUIS, IBM Watson Conversation and Amazon Lex. These platforms share a common *information-retrieval* approach based on what Moore and Arar (2019) identify as *Intent-Entity-Context-Response (IECR)* paradigm. An *intent* "[...] represents a mapping between what the user says and what action should be taken by the chatbot" (Adamopoulou and Moussiades 2020, 377). Intents recognize the conversational action a user is performing, while *entities* are tools used to extract particular details and parameters values from natural language inputs. They can be either system-defined or customized by the developer. *Contexts* are "[...] strings that store the context of the object the user is referring to" (Adamopoulou and Moussiades 2020, 378) used to capture the context of the current topic. *Responses* consist of what the chatbot actually answers in chat. This approach aims to extract contextual and detailed information from users' inputs and respond accordingly to the users' intention, extracting domain-specific entities and associating the corresponding intent, which means that machine learning algorithms are used for intent identification and entities extraction tasks, but responses are typically pre-authored by a designer (Moore and Arar 2019).

Generative models capable of automatically generating answers considering current and previous user messages are also in production, but there are still difficulties in building and training them and they are not available in the major commercial platforms (Adamopoulou and Moussiades 2020). Studies on the evaluation of these platforms show similar performances in terms of combined *f-score* (Liu et al. 2019; Braun et al. 2017), with slight differences in intent identification task (Canonica and De Russis 2018), especially with longer utterances (Zubania et al. 2020).

Although the technological progresses and the wide technical landscape here outlined, we are facing today a lack in designing domain specific conversational interfaces. The current state of conversational interfaces is limited in terms of established user interface design patterns: it is still unclear when chatbots should be text-based or button-based, or which are the best practices in designing a chatbot conversations. The question about how to structure the interaction with this new medium for creating efficient conversational experiences is still opened. Schiavo and Fadhil (2020) investigate the available scientific literature about interaction patterns and design principles in health-

care and identify four common theoretical themes in which the specific features are categorized: bot-user interaction, bot-response, bot-development, user experience. Since our work has a predominant linguistic focus, we mainly concentrate our attention on linguistic features, such as tone of voice, flexibility of responses, conversation length and user engagement in general. Schiavo and Fadhil (2020) treat each feature separately, offering relevant suggestions, but no univocal applicable design methodology in that sense.

Some studies embrace sociolinguistic theories: Chaves et al. (2019) apply to the design of a specific use case the *register theory*, Bennett (2018) and Dippold et al. (2020) identify Interactional Sociolinguistics as key to express chatbot's personality through language in responses and prompts design.

There are multiple works concerning the users' perceptions while texting with chatbots and what they would expect from a satisfying conversation with them: Hill et al. (2015) demonstrate that users hold long conversations also with conversational agents, adapting to chatbots language without overlooking that they were actually chatting with robots. Svenningsson and Faraon (2019) identify the factors of perceived humanness in chatbots' responses and underline their possible applications in terms of design guidelines. Jain et al. (2018) focus on new chatbots' users identifying guidelines more related to flow buildings. Kvale et al. (2019) draw practical and theoretical implications from a manual analysis of chatbots conversations, such as the value of cross-disciplinary teams and the need of diligence in chatbots training. Although these studies end with practical advice, they are far too generic for laying the foundations for a methodology.

There is consistent number of systematic guidelines on how to design conversational interfaces with a practical-computational procedure on how to approach chatbots' projects, reported also in McTear (2020). Some of them do have a commercial vocation (Hall 2018), other focus on technical issues (Shevat 2017; Dasgupta 2018) but even the more linguistically or cognitive oriented ones do not display a clear affiliation to a complete framework of analysis rooted in linguistics (Pearl 2016; Cohen, Giangola, and Balogh 2004).

We think that it is fundamental for creating effective conversational agents that should actually converse with humans and whose aim is to simulate the mechanism of human interaction to refer to a solid linguistic framework. Since pragmatics is the area of linguistics that primarily focuses on language in use also in interactional contexts, we agree with Bianchini et al. (2017) on the importance of pragmatics in developing new chatbots examples. Furthermore, we also agree with Bennett (2018) with the identification of Conversation Analysis as a methodological key to design better conversational flows.

2. Theoretical analysis of *dialogue* and *conversation*

2.1 A pragmatic perspective

The term *pragmatics* is conventionally credited to Charles W. Morris (1938) who first introduced a "pragmatic dimension" in the context of relations between signs, interpreters and objects (Bazzanella 2008). Influenced by Charles P. Peirce (1932) and in agreement with Carnap (1938), he distinguished three "dimensions of semiosis" (Morris 1938, 21), in which pragmatics addressed the relations between signs and who use and interpret them (Horn and Ward 2006). Since pragmatics is an interdisciplinary "hardly a well-integrated field of research" (van Dijk 2009, 13), it is preferable to speak about

a pragmatic *perspective* towards language instead of a pragmatic theory (Bazzanella 2008; Bublitz and Norrick 2011). The adoption of a pragmatic perspective also allows to embrace the contributions from philosophy, psychology, sociology, linguistics and the multiple definitions it has (Levinson 1983; Leech 1983; Katz and Fodor 1963; Ariel 2010; Turner 1999).

The philosophers of language Austin (1962), Grice (1975) and Searle (1969) influenced a notion of pragmatics in contrast with the chomskyan analysis of language as an abstract instrument independent by the context of use. Reflections in this direction are a consistent part of the most common handbooks of pragmatics, such as Levinson (1983), Leech (1983), Mey (1993), Yule (1996), and Verschueren (1999). It thus seems reasonable "[...] to claim that the ensuing pragmatic turn was most notably induced by J.L. Austin, J.R. Searle and H.P. Grice, who were interested in utterance meaning rather than sentence or word meaning, i.e. in studying unique historical events created by actual speakers to perform linguistic acts in actual situational contexts in order to accomplish specific goals" (Bublitz and Norrick 2011, 2).

The approach adopted in this work is *functionalist*, that is, "[...] that it attempts to explain facets of linguistic structure by reference to non-linguistic pressures and causes" (Levinson 1983, 7). This perspective aims to explain linguistic phenomena relying on pragmatics principles (Givon 1979; Hymes 1962) and opens to different developable possibilities such as an ethnomethodological method rooted in sociology (Garfinkel 1996; Goffman 1983; Sacks, Schegloff, and Jefferson 1992) and a psycholinguistic approach such as the alignment model (Pickering and Garrod 2004; Branigan, Pickering, and Cleland 2000; Szmrecsanyi 2005).

In both cases, the study of pragmatics is connected to the use of language in communication. Since communication inevitably involves at least two parties, the primary focus of pragmatics are "language use and language users in interaction" (Bublitz and Norrick 2011, 4). Towards this intersubjective dimension, Kerbrat-Orecchioni (2001) speaks of *pragmatique interactionniste*, whose main objects of research are manifestations of verbal interactions, such as dialogue and conversation.

Dialogue could be taken as "[...] the elementary and universal form of human communication" (Luckmann 1990, 58), whose basic principles are most salient in conversations and authentic discourses (Linell 2001). Levinson (1983, 284) defines conversation as "[...] the predominant kind of talk in which two or more participants freely alternate in speaking, which generally occurs outside specific institutional settings". In the broadest sense, it includes both face-to-face social communications and technology-mediated forms of interactions: all these different manifestations can be classified according to different criteria. A common distinction bases on the final scope of the exchange: social interactions' aim is building and maintaining rapports, while transactional interactions mainly fulfill practical goals (Brown and Yule 1983; Clark et al. 2019). Hakulinen's classification (2009) takes into account the degree of institutionality, the activity type or genre, the channel and participation framework; Linell and Luckmann (1991) consider the degree of asymmetry between the interlocutors. According to Schegloff, the *ordinary conversation* is the most general and flexible type of conversation from which the other types are adapted for particular purposes. It is defined as "[...] the basic medium of 'interactional exchange' [...] in whatever practices it is embodied in those settings" (Schegloff 1999, 413). Moore and Arar (2019) identify *service*, *teaching* and *counseling conversations* as derived typologies from the ordinary conversations. They are all characterized by roles' fixedness and influenced by their settings. Since Moore and Arar (2019) embrace a strictly conversationalist point of view, their focus is on identi-

fyng the underlying structure of conversation, which is suitable for slight adaptations according to the settings and contexts real conversations may occur in.

2.2 Pragmatic frameworks of analysis

There are several useful frameworks to analyze dialogical and conversational interactions. In the field of pragmatics, Haugh (2012) discerns two key trends to place conversational interactions: at the level of meaning and abstract principles referring to the works of language philosophers such as Grice (1989) and Searle (1969); at the level of the performance the analysis of authentic data referring to Conversation Analysis (Sacks, Schegloff, and Jefferson 1974) and Interactional Sociolinguistics (Gumperz 1982). In this section, we summarise the key points of each and the possible adaptation to a human-computer interaction.

The first trend is situated on a cognitive level and understands conversation as a “*joint activity*” whose progression is determined by the concatenation of “joint actions” (Clark 1996, 30). They are the result of the coordination of individual actions on two levels: “There is coordination of both content, what the participants intend to do, and processes, the physical and mental systems they recruit in carrying out those intentions.” (Clark 1996, 59). Regarding language use, “[...] a central problem is coordinating what speakers mean” (1996, 73). The idea of conversation as action determined by an undercurrent of communicative intention between the participants was formulated by Austin (1962) and Searle (1969, 1983). These studies have been enormously influential in the pragmatic approach of conversation: they allow to “[...] formalise rules and principles by which speakers mean (and to a lesser extent do) things” in conversation, abstracting from the conversation itself (Haugh 2012, 251). For example, the austinian notion of performativity frames new perspectives in human-computer-conversations, such as the collaborative action of “*We*” *Human-and-Technology* (Cho and Yoon 2013; Cho 2015) and the methodology of the *Performative Experience Design* (Spence 2016).

Intentionality and delivery of implicit meanings are Grice’s main objects of inquiry (1975). He can be considered one of the pioneers of *inferential pragmatics* (Ariel 2012). Most of all the cooperation principle and the conversational maxims proposed by Grice (1975) and later updated by Sperber and Wilson (1995) are a consistent part of the study of pragmatics today and involved in the implementation of dialogue systems from a methodological point of view: Jacquet et al. (2018; 2019; 2019b) evaluate the violation of the gricean maxims in textual online conversations; Saygin and Cicekli (2002) propose an empirical study of human-computer interactions within the context of the Loebner Prize Contest.

Lakoff’s theory of politeness (1973) is an attempt of expansion of Grice’s conversational maxims. This theory has been extensively criticized because it is hardly generalizable (Al-Duleimi, Rashid, and Abdullah 2016) and the key terms used in it are culturally determined and they therefore need to be clearly defined (Brown 1976; Tannen 1984).

Brown and Levinson (1987) propose instead an expansion of the studies on politeness made by Goffman (1967) introducing the concepts of *positive face* and *negative face*, which are respectively the need to be approved by the others and the need of autonomy. The importance of politeness in the realization of conversational interfaces is attested among the others by Følstad et al. (2018) and Nordheim, Følstad, and Bjørkli (2019), who list politeness as a factor perceived to affect trust in chatbots for customer service. As far as practical chatbots realizations are concerned, Hall (2018) includes it as an element to take into account during the conversation design, while De Jong, Theune, and Hofs

(2008) adapt the model of politeness strategy elaborated by Brown and Levinson (1987) and of the linguistic alignment (Pickering and Garrod 2004) to a virtual museum guide.

In relation to the second trend, the dominant perspectives are *Interactional Sociolinguistics* (IS) and *Conversation Analysis* (CA).

The core idea of Interactional Sociolinguistics is that what happens in a sequence of talk can be analyzed in its social contexts and that humans in talk accomplish social goals. A central concept of this approach are *contextualization cues*, “[...] by which speakers signal and listeners interpret what the activity is, how the semantic content is to be understood and how each sentence relates to what precedes or follows” (Gumperz 1982). Feine et al. (2020) offer an overview of the implementation of social cues in different kinds of conversational agents, while Bennet (2018) translates conversational cues in the realm of text-based chats, arguing that a strategic manipulation of orthography to convey conversation cues could help the design of chatbots personality and could situate them on different levels of *enthusiasm* or *considerateness*. Dippold et al. (2020) attest how a microlevel of design linguistic analysis based on Interactional Sociolinguistics can be useful for chatbot designers for creating engaging interactions and provide specific guidelines. Relevant studies in the field of Computer-Mediated Communication (CMC) describe the modification on the different levels of language that apply on digital mediated communication: Crystal (2001) and Herring (2012) are pioneers for the English language, while for Italian the work from Pistolesi (2018) is certainly a relevant reference point.

Another approach to dialogical interaction is Conversation Analysis. Levinson (1983) includes in his textbooks on pragmatics a chapter entitled *Conversational structure*, in which he compares *Conversation Analysis* and *Discourse Analysis* as two opposite methodological frameworks with a preference for the first one, characterized by an empirical and inductive vocation.

CA is a subfield of sociology whose origin is influenced by ethnomethodological studies (Garfinkel 1996), but the publications of the firsts and more influential CA contributions in 1974 *A simplest systematics* in the flagship *Journal of the Linguistic Society* and of *The preference of self-correction* in 1977 on *Language* both by Schegloff, Sacks and Jefferson established a relationship between CA and Linguistics from its origins (Fox et al. 2018).

The aim of this discipline is providing a *systematic description* of oral language *practices* (Schegloff 1992, 120) between humans, in order to formalise it into key structural elements that occur in the variety of contexts in which conversations may take place. Unlike ethnomethodology, whose observations are based on memory and intuition (Pallotti 2007), the methodological approach of CA is fundamentally *empirical* and its focus was a description of language as a tool used by social actors in interactions. In this respect, the object of study is an *interactive activity contextually situated*, where sentences are “[...] produced by someone, for someone else, at a certain time, in a certain way” (Hoey and Kendrick 2017). In the words of Schegloff and Sacks (1992, 70), CA is a “[...] naturalistic observational discipline that could deal with the detail of social action(s) rigorously, empirically and formally”.

The systematic descriptions of such practices leads to discovering the *machinery* (Sacks 1984, 84) underlying conversations, a *mechanics* of how people naturally talk in a variety of settings (Sacks, Schegloff, and Jefferson 1974; Schegloff 2007) made of some key concepts that regulate interactions. This approach supports the existence of some structural patterns that occur in conversational exchanges without consideration of the delivered content and with slight adaptations according to the settings and contexts of

realization: it is therefore clear the enormous potential it may have in Human-Computer interaction studies.

We will describe four elements of the descriptive apparatus for analyzing interactional structures, adhering to the analysis made by Pallotti (2007), Moore and Arar (2019) and Hoey and Kendrick (2017), which are today “[...] common stock for everyone doing CA” (Pallotti 2007, 7).

2.3 Key concepts of the conversation *machinery*

The strategy people use to manage the conversational traffic in interaction and the distribution of talk among the parties is *turn-taking*, known to be the feature that makes conversations orderly without significant clashes, overlappings or long pauses. Sacks et al. (1974, 702) describe it as a *simplest systematics* composed of two components and a coordination of the ending of the turn with the start of the next. The *turn-constructive units* (TCU) consist of linguistic unit-types such as sentential, clausal, phrasal or lexical constructions that form a “[...] recognizably complete utterance in a given context” (Hoey and Kendrick 2017). These bound units are defined in functional terms, being understood that usually “[...] a complete (linguistic) action corresponds to a complete syntactic unit, so that the TCU boundary turns out to coincide with the clause boundary” (Pallotti 2007, 8). Once a turn is perceived as completed, occur a turn-transfer using turn-allocation techniques such as self-selection and other selection in specific *transition-relevance place* (TRP). A hierarchically organized set of rules governs the turn construction and coordinates the transition so as to minimize the gaps and overlappings (Sacks, Schegloff, and Jefferson 1974).

Turns do not occur haphazardly, but are *sequentially organized* into coherent courses of actions (Schegloff 2007). The minimal unit of sequential organization is a two-move sequence, the *adjacency pairs*, in which the connection between the parts depends on the *conditional relevance*: the occurrence of a first pair sets up the relevance of the second part to follow (Schegloff 1968). Schegloff and Sacks (1973) give some examples of adjacency pairs, such as question-answer, greeting-greeting, offer-acceptance/refusal. Moore and Arar (2019, 65) also include farewell-farewell, assessment-assessment, inquiry-answer, request-grant/deny, invitation-acceptance/decline, accusing-admitting/denying. Since the first part sets up an *expectation*, the absence of the accepted second part is noticeable and the participants may require explanations or justifications for not having answered or for not having chosen the *preferred* option (Schegloff, Jefferson, and Sacks 1977). Even if the first part is usually directly followed by the second part, it could happen that for satisfying various requirements the completion of the first part has to be suspended for one or more turns. Sequences are therefore inherently expandable through additional turns over and above the two basic units of sequences. Expansions are allowed before the first pair part (*pre-expansion*), between the first and the second pair part (*inserted expansion*) and after the second pair part (*post-expansion*) (Schegloff 2007, 26). Another kind of sequence are the *storytelling sequences* (Jefferson 1978), used to express stories, anecdotes, or instructions whose content have to be distributed on multiple turns. They are often introduced by pre-announcement, namely the story preface (Schegloff 2007, 41). Sequences are an instrument for organizing utterances produced by the participants, while sequences themselves are organized into *activities* that define the “overall structural conversation” (Schegloff and Sacks 1973, 71), such as conversation openings and closings, instruction giving or troubleshooting.

Another relevant element in CA is *turn design*, which deals with how speakers build their turns to achieve some goals and to deliver contents for a specific audience (Drew

2020). Turn design principles concern the speakers' orientation to contiguity and their tendency to display connections between what they are saying and what the other said in prior turns; the specific lexico-syntactic adopted by the speakers to pursue the desired action and *the recipient design principle*, which is "[...] a multitude of respects in which the talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the [...] co-participants [...] with regard to word selection, topic selection, admissibility and ordering of sequences, options and obligations for starting and terminating conversations [...]" (Sacks, Schegloff, and Jefferson 1974, 727). It implies the consideration of the relationship between the interlocutors, their mutual knowledge and *common ground* (Clark 1996). The other principle to take into account in turn design is *minimization*, known as the speakers' tendency to deliver a message or to complete an action without using more words than necessary, but still being recognisable from the recipient the conversation is tailored to (Sacks and Schegloff 2007). In other words, the recipient's design is prioritarian over minimization: the speaker has to be efficient, using as few words as possible without preventing the interlocutor to understand.

The last element of conversation machinery are the *repair practices* spontaneously accomplished by the speakers in case of troubles in speaking, hearing or understanding (Schegloff, Jefferson, and Sacks 1977). Hoey and Kendrick (2017) describe the three basic components of a repair procedure: a trouble source, a repair initiation (i. e. a signal that begins the repair procedure) and the repair solution (i. e. the actual repair, for example paraphrasing or repeating a word part of the prior turns). Both the speaker (*self*) and the recipient (*other*) can initiate a repair procedure and/or accomplish a repair solution. Repair mechanisms are thus distinguished relying on who initiates the repair (*self-initiated* or *other-initiated*) and on who effectively accomplishes the repair (*self-repaired* or *other-repaired*).

Some attempts have been made adopting CA as a theoretical framework for implementing conversational agents: Luff et al. (1990) early imagined the potentialities of CA both as analytical tool in HCI applications and as inspiration for design methodology. Wooffit et al. (1997) adopt CA as a sociological perspective for studying human-computer dialogues. More recently, Lotze (2016) includes CA in the theoretical approaches used to analyze human-chatbot corpora. Hirst (2001) reviews Luff, Gilbert and Frohlich (1990) focusing on the different conceptual perspectives from which CA and Discourse Analysis are shaped, defending the necessity of considering CA studies in the field of Natural Language Understanding (NLU) technologies and applications. At that time technologies were not mature enough to adhere to such a complex theoretical framework, but at present time more and more studies go in this direction, such as the works from Gervits et al. (2020) and Michael and Möller (2020).

2.4 The Natural Conversation Framework (NCF) as a promising starting point for a design methodology

The work by Moore and Arar (2019) on NCF represents an innovative proposal in the scientific-industrial landscape. Moore and Arar worked together at IBM-Research in designing prototypes for novel forms of interactions for conversational interfaces and Moore is currently developing a conversational methodology founded on the qualitative models from the field of CA. This is not the first work in this direction: Moore (2013) and Moore et al. (2018) collect interesting contributions on specific design issues generalizable to various use-cases, such as Bickmore et al. (2018) and Candello and Pinhanez (2018) and lay the groundwork for the practical guide published in 2019.

Moore and Arar (2018) especially introduce the lack of a methodology for designing conversations and invite us to embrace the complexities of human dialogues in order to create machines we can interact with in a natural way.

Moore and Arar (2019) can be considered a potential starting point for drawing a complete conversation methodology for different reasons. First of all, they argue that among the possible natural-language interaction styles the *conversation-centric style* is the future of AI interfaces¹, since it aims to reproduce a real conversation-first way of interaction. Therefore, they situate themselves on a higher level than simply offering a versatile practical procedure: they embrace a specific linguistic approach and translate it in the realization of human-machine dialogues. The sub-field of research of the *User Experience Design* they steer is rooted in CA, whose methodological principles and key elements are described in previous section.

The NCF traces the basics mechanics of conversational patterns documented in CA and consists of four parts: an underlying interaction model based on expandable sequences, a distinctive content format based on the interaction model, a reusable pattern language for common conversational activities, a general method for navigating conversational applications.

The interaction model is based on the sequential structure of conversations, in which sequences are “[...] general patterns that [...] can be used and reused in all kinds of different situations and settings, for all kinds of different purposes” (Moore and Arar 2019, 65). Like adjacency pair sequences and the storytelling sequences in CA, this model should support sequence expansions, as “[...] natural indicators of the participants’ state of understanding of a turn-by-turn basis” (Moore and Arar 2019). This is a more natural and interactive pattern than the simple two-turn sequence model of the majority of chatbots and virtual assistants currently available.

In order to apply the interaction model accurately, a particular format is required for the content of the conversational application. (Moore and Arar 2019, 70) express it translating the principle of minimization (Sacks and Schegloff 2007) into three guiding principles: “limit agent utterances to a single sentence or less”, “break paragraphs down into their parts” and “let users control the level of detail”. These criteria enable designers to break-up document-formatted content into bite-sized intents, which can be requested by the users through simple queries.

Nevertheless NCF does not provide a library of industry-specific content in the form of intents and entities, but a systematic set of dialogue patterns that constitute various aspects of conversational competence, enable a variety of social activities and can be configured to a wide range of use cases (Moore and Arar 2019). The catalogue proposed by Moore and Arar (2019) is made of 15 types of patterns and 100 subpatterns² and is directly inspired from the examination of naturally occurring observations.

The patterns can be divided into three categories: *conversational activities*, *sequence level management* and *conversation-level management* (see Figure 1). Conversational activities include patterns for managing content inside the boundaries of the conversation, such as inquiries, requests or extended tellings. The remaining two categories help users and agents to manage the interaction itself and occur on two levels: management

1 Other interaction-styles that imply the recognition of natural language inputs are the *system-centric style*, the *content-centric style* and the *visual-centric style*. They differ from the conversation-centric style because they do recognize and produce strings in natural language, but they do not exhibit the conversational actions distinctive of the human way of communicating (Moore and Arar 2019).

2 The list is not exhaustive. Here <https://ibm.biz/BdzwQU> are some new patterns IBM researchers are experimenting with.

patterns on a sequence level help managing successful sequential interactions through repair procedures, management patterns on a conversation level help managing conversations' openings, closings and disengaging. Finally, the navigation method consists of six simple conversational actions the users can take to get around the conversational space.



Figure 1

Graphical representation of the *Natural Conversation Framework* patterns.

The so defined set of patterns is much more similar to a “*language*” than a “*library*”: these patterns are designed to be used together and are particularly useful in service encounters conversational interactions, which are fundamentally practical dialogues characterized by a goal-seeking nature (Allen et al. 2001).

To show how the patterns are presented, we provide an example of sequence closer patterns, from the category of the sequence management patterns (Moore and Arar 2019, 142). Sequence closers enable users to communicate that the previous action is completed and that they are ready to move on:

```

U: INQUIRY/REQUEST
A: RESPONSE
U: ACKNOWLEDGMENT/ASSESSMENT/LAUGHTER
<OPTIONAL PAUSE>
A: LAST TOPIC CHECK
  
```

```

U: what's your favourite movie?
A: I guess I like movies with a strong AI lead.
U: ok
(2.0)
A: Anything else?
  
```

Each pattern is expandable with other subpatterns or may present different options according to its conversational structure. Patterns display the same format: an abstract model of the pattern with generic social actions or intents in capital letters and an adaptation of the pattern to a particular use-case. Here the user does the first part of an adjacency pair (an inquiry or a request) successfully granted or answered by the agent. If the agent did not succeed, the third position would not have been filled with an acknowledgement but it would have been expanded with a repair initiator, like “what did you say?” or “define AI”. In this case, the agent does not need to say anything more after the acknowledgement and, after 2-3 seconds pause in a vocal interaction or 20-30 seconds pause in a textual interface, the agent checks if the user has other requests or topics.

The adherence to an Intent-Entity-Context-Response paradigm allow to apply the language pattern of the NCF to various platforms and use-cases. Specifically, the examples shown in the textbook are all realized on Watson, but we think that the same patterns could be realized also on other platforms with the same paradigm, like Google's Dialogflow. Furthermore Moore and Arar (2019) methodology is primarily focused on the realization of vocal conversational interaction in English, but we argue that the same patterns with slight adaptation could be implemented also in textual conversational interfaces and in other languages, like in Italian.

3. A roadmap towards the implementation

3.1 A possible proposal: a summary roadmap

In this section, we provide a practical-procedural workflow for approaching chatbot projects based on our work experience in the field and on the most influential textbook guidelines on this topic (Cohen, Giangola, and Balogh 2004; Pearl 2016; Shevat 2017; Hall 2018; Dasgupta 2018; McTear 2020). This workflow is made of macro and micro-levels integrated together and should involve stakeholders from different departments, considering at least the joint effort of marketing, linguistic design and the technical development orchestrated by a flexible methodology that opens with the project's requirements definition up to the implementation.

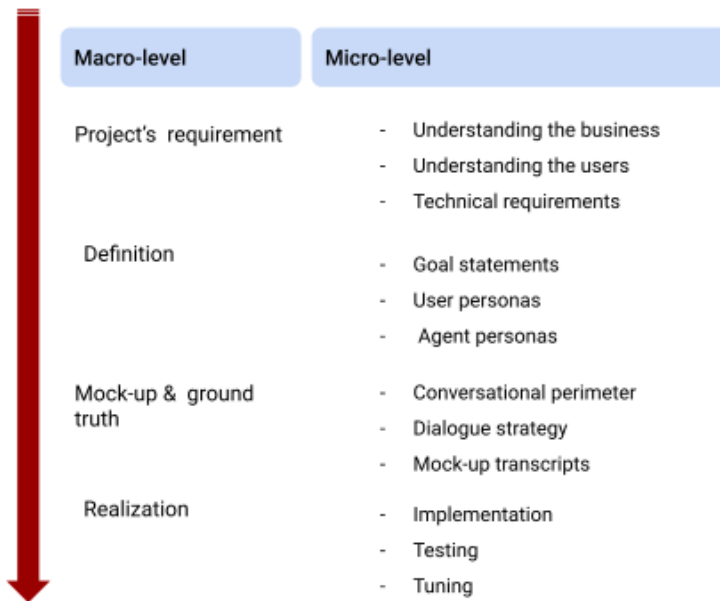


Figure 2
Schema for the proposed roadmap.

The stage of **project's requirements** definition produces as output the acknowledgement of the feasibility of the conversational agent's project, the identification of the potential target users and of the technical requirements the application should possess. We substantially agree with the procedure proposed in Cohen et al. (2004, 46) that covers three micro-levels: *understanding the business, understanding the user,*

technical requirements. There are multiple business issues that need to be understood and questions to be answered. Answering these questions can provide significant guidance to also write down metrics for the overall evaluation of the project.

The second element of the requirement phase is the understanding of the population of expected target users, which needs to be understood both in terms of *characteristics/needs* and of *usage modality*. Cohen et al. (2004, 49) list various elements for understanding and taking into account the final user's needs and characteristics.

The second perspective regards how, when, where and why the final users will use the application (Cohen, Giangola, and Balogh 2004, 50). They suggest also two practical ways to get these information: a preliminary overview of the various touchpoints of the company focusing on offered functionalities and transmitted feels, and organizing meeting with the company. The key attitude to develop to gain this information coincides with the first stage of the *Design Thinking* process assumed in Moore and Arar (2019), *empathize*. Empathy towards the interlocutors to get an understanding of the business and the users can be obtained through observations of how the final users currently interact with the industry or resolve the task that will be supplied by the chatbots, or through the engagement of the people directly involved. From a linguistic point of view, examples of useful materials to collect and analyze may be emails, call transcripts, or messages from the final users to the company.

The third requirement is getting an understanding of the application, focusing on the *technical requirements* of it. This stage is a prerogative of the technical department and its main point is getting an understanding of the application from a technical point of view, evaluating the feasibility itself and the strategies that need to be elaborated for solving tasks and subtasks.

The second macro-level point is **definition**, whose goal is to draw conclusions from the preliminary analysis of the project's requirements in order to define user needs more formally. We identify three elements that have to be defined: goal statements, user personas and agent personas. With *goal statements* we mean the definition of key design criteria learned from the analysis of requirements. They involve the definition of user goals grouped into broader statements, defined also with respect to the technical, financial and organizational constraints of the specific project.

The *user personas* is a fictional representation of the target user: a systematization of the collected user's qualities to define a prototype of who will typically interact with the final product. The user personas should also reflect the users' pain points, in order to improve the user journey. From our personal experience in the field, an efficient way to systematize the users' pain points is to write them down synthetically and associate them with benefits that the chatbot could provide in relation to them. An example is shown in Table 1.

The last element of the second section is the definition of some characteristics of the *agent personas* (Hall 2018; Cohen, Giangola, and Balogh 2004; Pearl 2016). In our experience on the field, we do agree with the synthetical sketch proposed in Moore and Arar (2019) that splits the agent personas design into three components: *agent job description*, *agent personality* and *agent self-knowledge*. A starting point for describing the qualities and the language of an agent, is imagining a job it is supposed to do. What role is the agent supposed to substitute or replace? Trying to list down duties and activities the chatbot is intended to assolve and the expected experiences and qualifications, as if it was a real job candidature, can help to be consistent also in the design itself and throughout the project. This procedure helps to identify jargon and recurrent technical terms the chatbot is supposed to manage. The *agent personality* needs to be characterized

Table 1

Pain points in the user journey associated with possible chatbot's benefits.

Pain Points	Chatbot benefits
Long waiting time on the phone	Instant answer or escalation to an operator
Difficult information retrieval	Personalised user journey to the retrieved information starting with an initial disambiguation
Hard understanding of complex and long documents	Systematisation of information in small slots and simplified language

in terms of communicative style, its level of formality and generally the tone of voice the target users expect to find in the conversational agent they are talking to. Two other core issues of the agent personality are the assignment of a gender to a chatbot, with the cultural and social implications that this may bring with, and the opportunity of humor in it. From our experience on the field, strongly anthropomorphized chatbots usually do belong to a gender, which most of the time is female (West, Kraut, and Chew 2019). There are also cases of neutral chatbots that reproduce an animal, a vegetable or a fantasy character. It depends on what kind of character or conception of gender we aim to reproduce in a virtual reality, being aware of the risk of reproducing virtually biases or prejudices belonging to the real world (Strengers and Kennedy 2020).

Humor can be an efficient strategy to build trust, especially if used in secondary responses that do not cover the main topics of the chatbot. Since the users expect a chatbot to be productive and efficient (Brandtzæg and Følstad 2017, 2018; Piccolo, Mensio, and Alani 2018; Zamora 2017), humoristic responses rather than informative ones can be counterproductive and may indeed frustrate the user. On the other hand, receiving a humoristic answer in an unexpected context such as online conversation with a machine, can increase the surprise effect and it may induce the users to continue the conversation (Jain et al. 2018).

Another correlated aspect involved in sketching a chatbot personality is the definition of some conversational paths that do not constitute the core topics of the chatbots but they are in some way related to it and can entertain the user. For example, a customer-service chatbot of an online motorbike clothing may provide an answer for a question like "What is your favourite motorbike brand?".³ Another way to reinforce the users' confidence towards the agent is working on the agent self-knowledge (Przegalin-ska et al. 2020; Følstad and Brandtzaeg 2020; Følstad, Nordheim, and Bjørkli 2018). A conversational agent can not have real perception of itself, but providing conversational paths that may help the users to navigate the conversational space created by the chatbot and the chatbot itself can be an efficient way for helping the user understanding what the chatbot can actually do and say and asking him more pertinent questions. Questions like this may regard the chatbot itself ("What are you?" "Are you a human?" "What is a

³ Business and commercial constraints have to be considered as well. In this case, the chatbot may not be able to provide an answer citing a specific brand, and therefore indirectly supporting a brand, but it can answer with a generic: "In my lonely virtual world I can only ride the wings of fantasy". Furthermore, some popular NLU platforms like Dialogflow contain pre-built conversational agents enriched with small-talks conversational paths covering generic topics such as weather, hour or day of the week.

chatbot?") or the competences of the chatbot ("What can you do?" "What can I ask you?" "What do you know?").

The third step in the proposed methodology is called **mock-up & ground truth**. Once the target users and the goal statements have been sketched out, we can move on to the drafting of the *conversational perimeter*. This term identifies a sort of table that holds together the groups of topics managed by the agent, structured in a way that reflects the Intent-Entity-Context-Response paradigm of the most common NLU platforms. In this phase, it is important to define the intents coherently and functionally to the goals of the agent. If conversational data collected during the preliminary phase such as call phone transcripts or emails are available, we suggest to group them following a *bottom-up labelling approach*. This approach consists of first grouping collected utterances into wide general categories, like "questions", "problems", "getting information" and then proceeds refactoring and splitting the so funded categories into more specific ones, considering for example the topic of the question, the action required to satisfy it and so on. Through this procedure we can both identify intents and train them with authentic linguistic material that constitute the so-called training phrases. A schema of how it works is shown in Figure 3 representing a schematisation of the reasonings behind the identification of the topics that will be handled by the agent, but it is not proper conversational perimeter as we mean it.

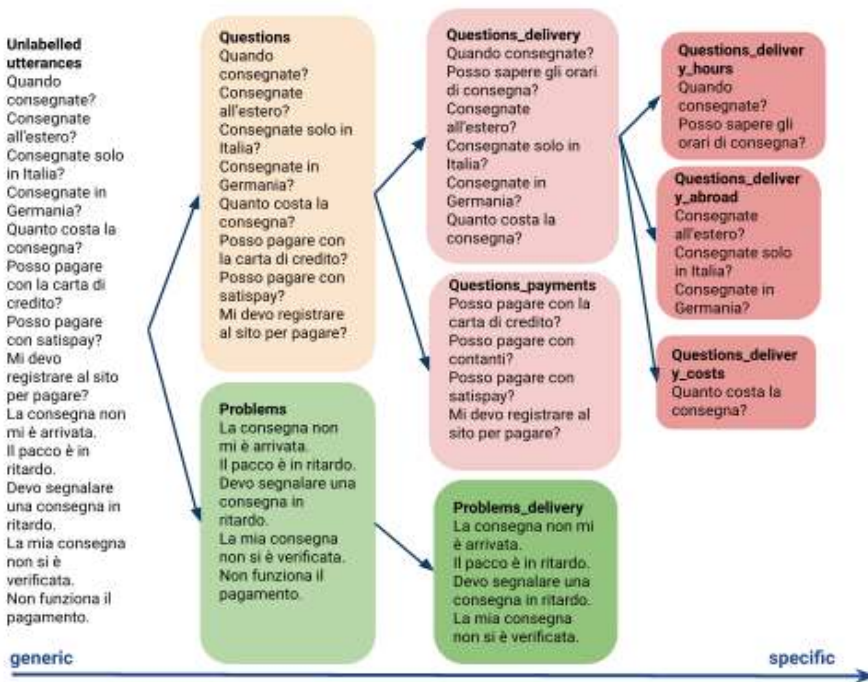


Figure 3
 Schema of the bottom-up labelling approach for the perimeter design.

The conversational perimeter is the final elaboration of this reasoning. It also includes a section dedicated to the responses, that should coincide with the actual responses that the agent gives in chat and has to be updated with new information

or modifications. Furthermore, it also contains some examples of the *training phrases*, utterances that final users can potentially say to formulate a request. Training phrases are categorized in *intents*, in turn grouped according to the topic they refer to. A conversational perimeter offers a systematization of the topics handled by the conversational agent, organized in intents like the platform it will be developed on; and it shows how simple question-answer exchanges will be managed directly in chat, with some examples of utterances the final users may produce and the answers they will be given. Table 2 shows the different section of a conversational perimeter, re-adapting the utterances from Figure 3 to the final elaboration of the perimeter.

Table 2
A section of a conversational perimeter.

Topic	Training phrases	Intent	Response
Consegna	Quando consegnate? Posso sapere gli orari di consegna?	Consegna_orari	Consegniamo tutti i giorni della settimana, compresi i festivi.
Consegna	Quando fate la consegna? Consegnate all'estero? Consegnate solo in Italia?	Consegna_estero	Consegniamo solo in Italia, non all'estero.
Consegna	Quanto costa la consegna?	Consegna_costi	Il costo della consegna è gratuito per gli utenti registrati e di 4 euro per gli ospiti.
Consegna	La consegna non mi è arrivata. Il pacco è in ritardo. Devo segnalare una consegna in ritardo. La mia consegna non si è verificata.	Consegna_ritardo	Se ritieni che ci sia un ritardo di consegna, posso inoltrare una segnalazione al servizio clienti, mi basta solo qualche tuo dato.
Pagamenti	Posso pagare con carta di credito? Posso pagare con satispay? Posso pagare con contanti?	Pagamenti_metodi	Puoi pagare con paypal, satispay e carta di credito registrata.
Pagamenti	Mi devo registrare al sito per pagare?	Pagamenti_account	Puoi comprare dal nostro sito sia come ospite sia come utente registrato, a te la scelta!

We need to consider that a complete conversational perimeter is usually much bigger than this: it depends on the specific project, but in our experience it can contain between fifty and one hundred intents. Furthermore, it contains only the simplest form of interaction: the responses are static, which means that they do not change dynamically taking information from external sources, but they are always the same, even though on the most common NLU platforms they can be randomized, that is to a specific intent can be associated with one or more responses randomly picked-up by the agent to create conversational variety. The form of the responses should respect the basic language-specific pragmatic norms, as shown in section 4.

The second micro-level step of the mock-up and ground truth mapping is dedicated to the *dialog strategy*, concerned with the effective building of the dialogue. It answers the question: how will the back and forth between the agent and the final users be? Will generally the agent start the conversation or the user? If the conversation flow is supposed to be more complicated than a simple question-answer and it requires the following of a specific path, we recommend the use of graphical tools for visualising

the steps of the path we have imagined and possible variations, such as Google Draw or XMind.

Mock-up transcripts is thought especially for more complex flows, that need more than one conversational turn to be developed and that may also be represented on one of the graphical tools mentioned above. In this phase there are two alternatives that can be pursued. From one side we can simply write down the dialogues we would like to reproduce with the agent (Cohen et al. (2004) and Pearl (2016) call them *sample dialogues*) and read it aloud to see if they sound human before implementing it, using either programming languages or commercial platforms that allow to concatenate conversational flows reproducing the characteristics identified by CA mentioned before. Alternatively, instead of creating the flows from scratch and implementing them, our proposal considers the Natural Conversation Framework and especially the sequential patterns proposed by Moore and Arar (2019). The latter approach appears more innovative and more structured, and it is the one we would like to adopt.

The final macro-level step of the roadmap is the concrete **realization** of the agent and consequently the effective implementation of the defined flows and selected sequential patterns on a NLU platform, on specific chatbot tools or using the most common programming languages.

After the implementation, a fundamental step before the roll out is *testing*. Like McTear (2020) claims, there is still no unified and univocal testing approach, but it depends on the implementation method behind the agent (Deriu et al. 2020) and on the project's requirements that need to be evaluated. Testing has a double function: to evaluate the efficiency of the developed application and, if possible, to improve the actual functionalities with a tuning activity. From our experience on the field, we suggest various test steps before the final roll-out. This allows developers and designers to evaluate the results and, if necessary, to tune some aspects of the application before the final version. A *dialog transversal test* (Cohen, Giangola, and Balogh 2004; Pearl 2016) for evaluating the behaviour of the system in every dialogue state and in every condition seems very important, trying out some out-of-perimeter utterances to verify the proper response of the system also in such cases. A fundamental aspect that needs to be considered in this context, is that a conversational agent is a constantly-in-progress creature: even though it is finished and especially if it supports a NLU tool, it needs to be updated with new linguistic materials in order to make the performances better and better. It has to be seen as an alive creature and project, that is nurtured by language and, in some way, produces language: therefore, training and maintenance are continuative activities. There are multiple elements that can be tuned or updated to make the performance of the chatbot better: the training phrases in natural language, the responses in case of changings in the information to deliver, new conversational paths. Table 3 shows a possible evaluation framework taken from our experience on the field and resulted from the combination of two variables: *in/out of perimeter*, which is referred to the coherence of the utterance produced by the user in relation to the conversational perimeter of the conversational agent; and *correct/wrong*, which is the effective evaluation of the response given by the conversational agent in that specific context. From every possible combination of these variables we provide possible improvements that can be undertaken.

4. Adaptation of the methodology and implementation

After the definition of a theoretical approach rooted in pragmatics and a practical-computational operating procedure towards the implementation, in this section we

Table 3
Evaluation grid.

	In perimeter	Out of perimeter
Correct	The user's utterance is in perimeter and the chatbot answers with the expected associated response: no improvements needed.	The user's utterance is out of perimeter and the chatbot answers properly activating the expected fallback intent: no improvements needed.
Wrong	The user's utterance is in perimeter and the chatbot answers wrongly with a response associated with another intent. Improvements may be: updating the training phrases of the missed intent, verifying the training phrases of other intents that may cause miss-match, or modify the response with more information.	The user's utterance is out of perimeter and the chatbot doesn't activate the fallback intent, but a response associated with another intent. Improvements may be: adding more training phrases to a fallback intent, adding training phrases to the missed intent, verifying training phrases that may have caused miss-match.

expand and enrich the procedural workflow with the implementation of a selection of patterns on one of the most authoritative NLU commercial platforms also for the Italian language, namely Google's Dialogflow (Zubania et al. 2020).

The selection of the patterns has been made according to two criteria: (a) the adaptation to a text-based modality of interaction; (b) the suitability to a customer-service context (Szymanski and Moore 2018).

We elaborate a general customer-service case study on which we transpose a selection of patterns in Italian. This practical section on the implementation aims to demonstrate the high generalizability of the approach, still considering the language-specific pragmatic implications in prompt design, and its suitability also to business-oriented contexts of use.

4.1 General case-study description

The structure of conversation belonging to a customer-service domain is similar to the more general category of service conversations, i.e. dialogic exchanges in which a person (in the role of a customer or a citizen) requests services or information and another person on behalf of an organization or an institution, provides services or information.

Since the focus of our work is not to build a comprehensive conversational agent but to demonstrate the effectiveness of our methodology for the design of sequential flows, we do not dwell on the details of the conversational perimeter of the case-study.

It suffices to say that it is a customer-service chatbot, whose goal is to provide repetitive information to support customers on the e-commerce of a chain of shops. It can provide information about typical online-shopping requests, such as deliveries, expeditions, payments and returns and it manages issues and problems related to the state of the orders. Customers may own a *fidelity card*, a card they collect points on to gain special discounts. Furthermore, customers can register on the website and activate an online profile with all the details on their customer's situation. Even though the main goal of the conversational agent is not to perform complex task, it should be able to support the final users through the registration process on the website. The chatbot

handles it through some questions focused on the extraction of data. We can distinguish between two kinds of data to extract: (a) user's name and user's email address are necessary data to complete the registration process. Without the collection of these data, the registration can not be performed; (b) the fidelity card number is an optional data. If the user decides to not provide it or the user does not possess a fidelity card, the registration process it is not compromised.

The registration process is the interaction that we are going to transpose in Italian using the NCF pattern and implement on Google's Dialogflow, in order to demonstrate the validity and the high generalizability of the presented theoretical and methodological approach.

4.2 Pattern selection and transposition

The registration process on the website can be realised basically associating three patterns of the NCF described in Table 4, namely the **pattern A2.6 Open Request Summary** the **pattern A2.7 Warrant Request & Refusal** and the **pattern A2.11 Open Request Repairs**. They all belong to the first group of pattern, *conversational activities*, and therefore help to manage what happens *inside* the boundaries of the conversation itself.

As shown in the section 2.4, each pattern is made of an abstract model of social actions in capital letters and an adaptation of the pattern to an hypothetical dialogue.

The pattern *A2.6 Open Request Summary* manages requests that need a progressive level of detail to be satisfied, like a registration process, in which a set of data need to be collected in order to perform an action or to fill-in a module. The peculiarity of this pattern is the final summary at line (8), that should allow a user's confirmation or modification of the inserted information. The modification can be performed using the pattern *A2.11 Open Request Repairs*, which handles repair procedures. The pattern *A2.7 Warrant Request & Refusal* manages user's questions concerning a detail request and the possible refusal of giving a specific data. As described in 4.1, we have both compulsory and optional data that need to be handled in different way.

In the next section, we explain the pragmatic aspect we take into account for the transposition into Italian and we then implement them on Google's Dialogflow.

4.3 Pragmatic aspects for the transposition into Italian

The pragmatic-functionalist perspective adopted in this work emerges both in the turn-taking sequential structure of the patterns of NCF and in the transposition of the chosen sequential patterns into the Italian language. The observations made by Cohen et al. (2004) about prompts design are still actual and rich in ideas, but they are concerned exclusively with the English language.

In our work these aspects become relevant for the distribution of content in the Italian transposition of the selected patterns. We are going to transpose patterns from a language to another, namely from English to Italian, respecting the predominant tendency of positioning on the left side of the utterance the topic/given information and the comment/new information on the rest of the sentence. This norm can be violated for obtaining particular sense-effects and for enlightening a part of an utterance that in a normal order of the component would instead be positioned at the end of the sentence. In Italian the most used strategies for producing this effect on a written channel of communication are the inversion of the normal order of the syntactic components through *focalizations* such as *dislocations* and *left sentences*.

Table 4
Selected patterns for the transposition into Italian.

Pattern A2.6 Open Request Summary	
1 U: PARTIAL REQUEST	U: I need to register
2 A: DETAIL REQUEST	A: Okay. What's your name?
3 U: DETAIL	U: Dave Bowman.
4 A: DETAIL REQUEST	A: What's your job title?
5 U: DETAIL	U: astronaut
6 A: DETAIL REQUEST	A: What is your organizational affiliation?
7 U: DETAIL	U: NASA
8 A: REQUEST SUMMARY	A: So Dave Bowman, Astronaut at Nasa. Is that correct?
9 U: SUMMARY CONFIRM	U: yes
10 A: GRANT	A: Ok.
11 U: SEQUENCE CLOSER	U: thank you
12 A: RECEIPT	A: You're welcome!
Pattern A2.7 Warrant Request & Refusal	
1 A: DETAIL REQUEST	A: What's your job title?
2 U: WARRANT REQUEST	U: why do you need to know?
3 A: WARRANT	A: Job title enables us to know the range of professions of our participants.
4 U: REFUSAL	U: none of your business.
5 A: ACKNOWLEDGMENT	A: No problem. What's your organizational affiliation?
Pattern A2.11 Open Request Repairs	
1 U: FULL REQUEST	U: Can you recommend a nearby Mexican restaurant?
2 A: GRANT	A: Mario's located at Beach and Main.
3 U: REPAIR INITIATOR	U: What do you mean?
4 A: REPAIR	A: The Mexican Restaurant, Mario's, is located at the intersection of Beach Boulevard and Main Street.
5 U: SEQUENCE CLOSER	U: thanks

Another element on which Cohen et al. (2004) dwell on are *discourse markers* and their use in prompts's content design for conversational agents in English, organized into various classes of use and considered an efficient strategy for enhancing the natural flow of conversational technology. Without going deeper into the classification of discourse markers in Italian (among the others Bazzanella (1995) and Berretta (1984)), it suffice to say that we will take them into consideration for their pragmatic impact on sentences organizations and structures during the patterns' transposition. In Italian, they are indeed significantly important for their interactional function on both sides of the speakers, their meta-textual function as demarcations, focalizations and riformulations and finally for their cognitive functions (Bazzanella 2008). As Cohen et al. claim, it

can be that discourse markers are perceived as informal or slang, even though they are classified as a *functional* category with no explicit formality degree.

What instead has to do with the level of formality of an utterance and a piece of discourse in general (Clark 1996), are *register* and *consistency*. Even though the definition of register is controversial (Bazzanella 2008), we can say it involves the psychological and social rapports between the speakers, the circumstances in which the communication takes place and the adopted channel (Halliday 1994). This is one of the dimensions of variations of language. Other dimensions of variations are determined by the spatial origins and the geographical distribution of the speakers, by socio-cultural elements such as level of instruction, age and competences and channel of communication. Especially in technical or highly specialized work context, the use of jargon is an important issue to take into account. It is indeed acceptable if all the speakers do share similar background and analogous competencies in relations to the main content expressed by the conversational agent we are working on.

The channel of communication is also an issue with significant implications on the distribution of content and therefore on relation to the drafting of the responses. Without deepening into the characterization of the language variations adopted in digital contexts of communication, we have to consider that the variant of language adopted with a text-based conversational agent is an intermediate solution between the two opposite poles of written and oral language. This variation takes place in a written form but it shares important elements with speech (Pistolesi 2018), since it is a type of writing that considers more the acoustic effect than the visual one.

All these sociolinguistic and pragmatic aspects have to be considered as requirements for transposing the selected patterns into Italian. In the next section, we are going to outline a generic customer-service use-case and in followings there is the effective realization and transposition of the patterns on Google Dialogflow.

4.4 Implementation on Google's Dialogflow

We realize the selected NCF patterns on Google's Dialogflow ES.

The three patterns have been combined together in order to cover multiple scenarios we may face during a registration process. For doing that, we implemented fourteen intents on Dialogflow

<i>01_reg_00_registrazione_generico</i>	<i>01_reg_03_registrazione_riepilogo</i>
<i>01_reg_01_giustificazione_nome</i>	<i>01_reg_03_rifiuto_email</i>
<i>01_reg_01_registrazione_carta</i>	<i>01_reg_03_spiegazione_email</i>
<i>01_reg_01_rifiuto_nome</i>	<i>01_reg_04_modifica_dati</i>
<i>01_reg_02_registrazione_email</i>	<i>01_reg_04_riepilogo_corretto</i>
<i>01_reg_02_spiegazione_tessera</i>	<i>01_reg_04_riepilogo_negativo</i>
<i>01_reg_02_tessera_rifiuto</i>	<i>01_reg_05_ringraziamenti</i>

connected together by input and output contexts manually set and we use three system entities for the parameter extraction: (a) *sys.person* to extract and memorise the user's name, (b) *sys.number* to extract and memorise the fidelity card's number and (c) *sys.email* to extract and memorise the user's email.

The opening intent *01_reg_00_registrazione_generico* is activated by training phrases that express the intention of starting a registration process on the website. As Figure 4 displays, the contents' distribution of the agent's response follows the topic/comment order described in 4.3: the first part introduces the registration process already known

by the user, since he/she asked for it, and prepares the speaker to provide a set of data. In this case, the user provides the requested data and the user's name is thus memorised through the annotation of the training phrases with the *sys.person* entity and the extraction of the corresponding parameter.

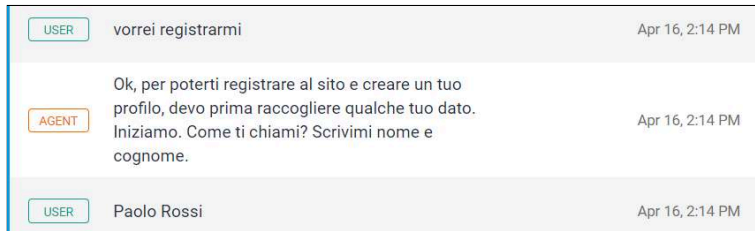


Figure 4

Realization of the pattern *Open Request Summary*. (TRANS. U:I would like to register /A:Ok, in order to register on the site and create your own profile, I must first collect some of your data. Let's begin. What's your name? Write your name and surname /U: Paolo Rossi).

As the pattern *A2.7 Warrant Request & Refusal* demonstrates, repetitive patterns in a process based on the data extraction, such as the outlined registration, are users' requests of warranty and explanation concerning a specific data. We provided two intents for managing questions about the reasons of requesting a name in this context and the refusal of providing it, respectively the intent *01_reg_01_giustificazione_nome* and the intent *01_reg_01_rifiuto_nome*, both connected to the hierarchical intent *01_reg_00_registrazione_generico* by contexts. The intent *01_reg_01_rifiuto_nome* is also subordinated to the *01_reg_01_giustificazione_nome*, which means that after an explanation's request by the user the chatbot is able to handle also a refusal of providing the requested data (see Figure 5).

The chatbot's responses are more extended than the examples provided by Moore and Arar (2019). This is because we are specifically working on a text-based interface, which allows us to insert also a bit longer prompt than with only vocal conversational assistant. As we claimed in Section 4.1, the user's name is a required information without which the registration process cannot be performed. The conversational agent offers an alternative to the user, (*"Se vuoi, puoi registrarti anche dal vivo, ti basta andare alla cassa [...]"*) because we are in a business-oriented context, where the customer satisfaction is the most important requirement (Gnewuch, Morana, and Maedche 2017). If after the activation of the intent *01_reg_01_rifiuto_nome* the user at the end provides the name, the registration process can proceed again.

The second data to extract is the fidelity card number, which is not compulsory and it can be skipped without compromising the registration. Also in this case we have a *good path* (Pearl 2016) where the number is immediately provided, and two expanding intents used to manage clarification's requests and refusal related to this topic, namely the intent *01_reg_02_spiegazione_tessera* and the intent *01_reg_02_tessera_rifiuto*. Since the fidelity card number is not a sensible data, instead of a warranty we provide further explanation about the number (where it is, what it is, why it is requested) and possibility of managing refusal, especially if determined by incomprehension, impossibility to find the card or absence of the card (see Figure 6).

In any case, if the fidelity's card has been provided or not, the registration conversational flows proceeds to the third and last data, the user's email, which is, like the name, a compulsory data to complete the registration, and it is therefore handled by the intents *01_reg_03_rifiuto_email* and *01_reg_03_spiegazione_email* with training phrases

AGENT	Ok, per poterti registrare al sito e creare un tuo profilo, devo prima raccogliere qualche tuo dato. Iniziamo. Come ti chiami? Scrivimi nome e cognome.	Apr 17, 11:17 AM
USER	perchè serve?	Apr 17, 11:17 AM
AGENT	Il nome è un requisito essenziale per iniziare la procedura di registrazione. Verrà inserito nei nostri database per associare tutte le informazioni del tuo profilo univocamente a te. Come ti chiami?	Apr 17, 11:17 AM
USER	non voglio fornirlo	Apr 17, 11:17 AM
AGENT	Ok, come preferisci. Senza il tuo nome, però, non posso aiutarti nella registrazione. Se vuoi, puoi registrarti anche dal vivo, ti basta andare alla cassa di uno dei nostri negozi durante gli orari di apertura e chiedere aiuto ad uno dei nostri addetti alla vendita.	Apr 17, 11:17 AM

Figure 5

Realization of the pattern *Warrant Request & Refusal*. (TRANS. A:Ok,...same as before...surname/U:Why is it needed?/A:The name is an essential requirement to start the registration process. It will be entered in our databases to associate all the information in your profile uniquely with you. What's your name?/U:I don't want to provide it/A:Ok, as you like. Without your name, however, I cannot help you register. If you want, you can also register live, just go to the cashier of one of our stores during opening hours and ask for help from one of our sales staff).

similar to ones dedicated to the name, with the difference that they are annotated with the *sys.email* entity.

As displayed in the pattern *A2.6 Open Request Summary*, after the collection of data we have a summary of all the provided information memorised through contexts and reported in the responses of the intent *01_reg_03_registrazione_riepilogo* through coding some specific variables through a specific Dialogflow syntax, as Figure 7 shows. These are the responses that the conversational agent displays once the *01_reg_03_registrazione_riepilogo* is activated by the insertion of an email address. They contain parameters associated to the data extracted with entities and stored in the conversational flow through contexts, according to the form: *#context-name.parameter-name*. The two responses are chosen randomly from Dialogflow: it means that there is no certainty that if the user did not provide the fidelity card's number in the flow the conversational agent will display the response with *"carta fedeltà non fornita"*. Our tests reveal a tendency to choose the appropriate response according to the collected data, but for more precise result we recommend to customize the response with some code. In Figure 8 we show how they appear if filled with the corresponding value.

The summary ends with a agent's request of confirmation before validating the inserted data. If the user confirms, the data are validated and the sequence can be closed with a sequence closer and its reception. If the users would like to modify one of the inserted data, a repair procedure can be started. The repair procedure can pass through the intermediate intent *01_reg_04_riepilogo_negativo*, where the user simply acknowledges that the proposed summary is not the correct, or it can directly activate the intent *01_reg_04_modifica_dati*, the core-intent of the repair. Figure 9 shows a

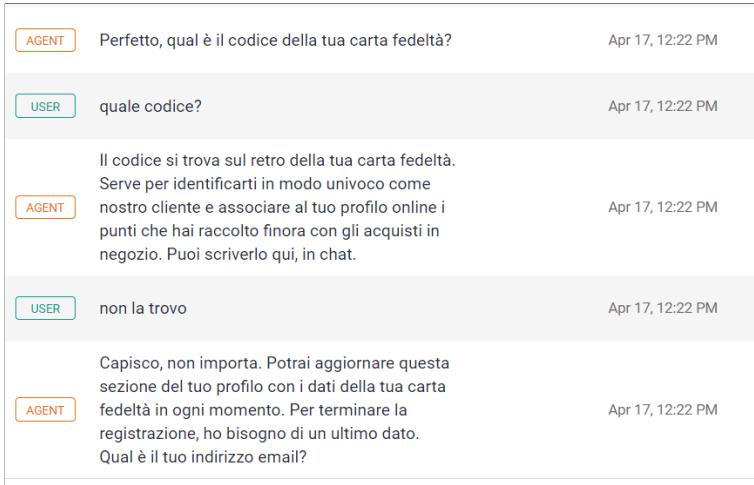


Figure 6
 Realization of a slightly modified version of the pattern *Warrant Request & Refusal* according to the type of data to extract. (TRANS. A:Perfect, what is your loyalty card code?/U:Which code?/A:The code can be found on the back of your loyalty card. It is used to uniquely identify you as our customer and associate the points you have collected so far with in-store purchases to your online profile. You can write it here in chat/U:I cannot find it/A:I see, it doesn't matter. You can update this section of your profile with your loyalty card details at any time. To end the registration, I need one more data. What is your e-mail address?).



Figure 7
 Dialogflow's syntax for the final recap in the pattern *Open Request Summary*.

repair procedure with the activation of both the intents *01_reg_04_riepilogo_negativo* and *01_reg_04_modifica_dati*.

The repair procedure is handled by the intent *01_reg_04_modifica_dati*, connected to the previous ones by contexts and whose training phrases are annotated with the system entities used before. The difference here is that the parameters associated to the entities and extracted from the user's sentences are saved with another value. The new extracted value is used in the response with the same `#context-name.parameter-name` syntax, in order to overwrite the previous value with the new one. The next figures show the parameter section in Dialogflow and the responses with the extracted parameters.

The prompts are enriched with discourse markers that functionally structure the progression of the conversational flow ("*Va bene*", "*Ok, quindi*", "*Ok*", "*Perfetto*", "*Ottimo, abbiamo finito*").

This demonstration does not expect to exhaust all the possibilities and expansions that a comprehensive registration process may present. The repair procedure handles for instance the modification of a single data and we did not focus on the insertion of

AGENT	Ok, per poterti registrare al sito e creare un tuo profilo, devo prima raccogliere qualche tuo dato. Iniziamo. Come ti chiami? Scrivimi nome e cognome.	Apr 17, 1:20 PM
USER	Anna Rossi	Apr 17, 1:20 PM
AGENT	Perfetto, qual è il codice della tua carta fedeltà?	Apr 17, 1:20 PM
USER	12345	Apr 17, 1:20 PM
AGENT	Ok, grazie. Ora, avrei bisogno di un indirizzo email su cui inoltrare comunicazioni e offerte.	Apr 17, 1:20 PM
USER	annarossi@gmail.com	Apr 17, 1:20 PM
AGENT	Ottimo, abbiamo finito. Riepilogando: Anna Rossi, 12345, indirizzo email annarossi@gmail.com. È tutto giusto?	Apr 17, 1:20 PM
AGENT	profilo, devo prima raccogliere qualche tuo dato. Iniziamo. Come ti chiami? Scrivimi nome e cognome.	Apr 17, 3:35 PM
USER	Giuseppe Verdi	Apr 17, 3:36 PM
AGENT	Perfetto, qual è il codice della tua carta fedeltà?	Apr 17, 3:36 PM
USER	non ce l'ho	Apr 17, 3:36 PM
AGENT	Capisco, non importa. Potrai aggiornare questa sezione del tuo profilo con i dati della tua carta fedeltà in ogni momento. Per terminare la registrazione, ho bisogno di un ultimo dato. Qual è il tuo indirizzo email?	Apr 17, 3:36 PM
USER	giuseppeverdi@gmail.com	Apr 17, 3:36 PM
AGENT	Ottimo, abbiamo finito. Riepilogando: Giuseppe Verdi, carta fedeltà non fornita, indirizzo email giuseppeverdi@gmail.com. È tutto giusto?	Apr 17, 3:36 PM

Figure 8

Realization of the pattern *Open Request Summary* with the two scenarios: in the first conversation the user provides all the requested data while in the second the users does not provide the fidelity card number. (TRANS. *First Part*: A:Ok,...as in Figure 5...surname/U:Anna Rossi/A:Perfect, what is your loyalty card code?/U:12345/A:OK thanks. Now I need an email address to forward communications and offers to/U:annarossi@gmail.com/A:Good, we have finished. Summarising: ... Is it correct? *Second Part*: ...same as before... U:/I do not have it/A:I see, it doesn't matter. You can update this section of your profile with your loyalty card details at any time. To end the registration, I need one more data. What is your e-mail address?/U:giuseppeverdi@gmail.com/A: ...same as before...).

unexpected or wrong-spelled data. The goal of our implementation is to demonstrate the effectiveness and the validity of an approach, which is supported by an high-level theoretical method and which is easily transposable into other languages and using various NLU commercial platforms.

AGENT	Ok, per poterti registrare al sito e creare un tuo profilo, devo prima raccogliere qualche tuo dato. Iniziamo. Come ti chiami? Scrivimi nome e cognome.	Apr 17, 1:54 PM
USER	Anna Bianchi	Apr 17, 1:54 PM
AGENT	Perfetto, qual è il codice della tua carta fedeltà?	Apr 17, 1:54 PM
USER	12345	Apr 17, 1:54 PM
AGENT	Ok, grazie. Ora, avrei bisogno di un indirizzo email su cui inoltrare comunicazioni e offerte.	Apr 17, 1:54 PM
USER	annabianchi@email.com	Apr 17, 1:54 PM
AGENT	Ottimo, abbiamo finito. Riepilogando: Anna Bianchi, 12345, indirizzo email annabianchi@email.com. È tutto giusto?	Apr 17, 1:54 PM
USER	no	Apr 17, 1:55 PM
AGENT	Va bene, quale informazione vorresti modificare?	Apr 17, 1:55 PM
USER	voglio modificare la mail in annabianchi@hotmail.com	Apr 17, 1:55 PM
AGENT	Ok, quindi: Anna Bianchi, carta fedeltà numero 12345, indirizzo email annabianchi@hotmail.com. Ora è tutto giusto?	Apr 17, 1:55 PM
USER	si	Apr 17, 1:55 PM
AGENT	Perfetto. A breve, riceverai una comunicazione all'indirizzo email che hai indicato per completare la registrazione.	Apr 17, 1:55 PM
USER	grazie	Apr 17, 1:55 PM
AGENT	Prego!	Apr 17, 1:55 PM

Figure 9
 Realization of the pattern *Open Request Summary* and *Open Request Repairs*. (TRANS. *The first 7 steps as in previous figures*/U:No/A:Ok, which information would you like to change?/U:I would like to change the email in annabianchi@hotmail.com/ A:Ok,...same as before.../U: Yes/A:Perfect. Soon, you will receive a communication to the email address you indicated to complete the registration/U:Thank you/A:You are welcome!).

5. Conclusions and future directions

We have therefore obtained a formal, highly generalisable procedure for building conversational agents. Our procedure is explicitly anchored to Conversation Analysis as authoritative methodological approach for realising sequential exchanges and reproducing typical features of human conversational interactions. We additionally transposed

REQUIRED	PARAMETER NAME	ENTITY	VALUE	IS LIST
<input type="checkbox"/>	emailmodificata	@sys.email	\$emailmodificata	<input type="checkbox"/>
<input type="checkbox"/>	personmodificata	@sys.person	\$personmodificata	<input type="checkbox"/>
<input type="checkbox"/>	numbermodificata	@sys.number	\$numbermodificata	<input type="checkbox"/>

Text Response	
1	Ok, quindi: #nome_corretto.person, carta fedeltà numero #modifica_dati.numbermodificata, indirizzo email #email_corretta.email. Ora è tutto giusto?
2	Ok, quindi: #nome_corretto.person, carta fedeltà numero #carta_corretta.number, indirizzo email #modifica_dati.emailmodificata. Ora è tutto giusto?
3	Ok, quindi: #modifica_dati.personmodificata, carta fedeltà numero #carta_corretta.number, indirizzo email #email_corretta.email. Ora è tutto giusto?

Figure 10 Dialogflow’s syntax and parameters for the realization of the pattern *Open Request Repairs*.

the patterns in Italian considering the pragmatic implications of the chosen language and selected a flexible case-study that allows to easily recreate the patterns in multiple contexts and situations.

The generalisability of the selected patterns is thus subordinated to two main factors: the cultural implications that may occur in the transposition of the patterns into another language and the level of complexity of the conversational flows to implement. The problem of the transposition of the pattern deals with the pragmatics of the language and the cultural expectations related to the customer experience that may influence the dialogue design as well. As Brandtzaeg and Følstad (2017) claim the search of productivity in chatbot use is explicitly anchored to Western culture. This aspect conditions the entire dialogue design, for example in positioning the chatbot scope within the very first conversational turns, in order to not waste time. In our experience, the adaptation of the pattern have been made between two rather culturally similar language, but it might not have been the same with, for example, an oriental culture and its language. Beyond the strictly linguistic aspects, also the project requirements and the agent personas definitions are also culturally defined (Ruane, Birhane, and Ventresque 2019; Cardinal, Gonzales, and Rose 2020). The generalisability of the patterns is also determined by the scope of the specific conversational agent that need to be implemented and by the level of granularity of the conversational flows it should have: the patterns can be seen as a base model that can be simplified or complicated depending on the individual needs.

As future directions of our research, we are going to further investigating the applicability of this approach on the Dialogflow CX version and possibly on other commercial NLU platforms.

Acknowledgements

We would like to thank Injenia S.r.l. for supporting this research.

CRedit author statement; ADA: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Writing (Original Draft), Writing (Review & Editing); FT: Conceptualization, Supervision, Project Administration, Writing (Review & Editing), Funding Acquisition.

References

- Adamopoulou, Eleni and Lefteris Moussiades. 2020. An overview of chatbot technologies. *Artificial Intelligence Applications and Innovations*, 584:373–383.
- Ahmad, Nahdatul Akma, Mohamad Hafiz Che Hamid, Azaliza Zainal, Muhammad Fairuz Abd Rauf, and Zuraidy Adnan. 2018. Review of chatbots design techniques. *International Journal of Computer Applications*, 181(8):7–10.
- Al-Duleimi, Hutheifa Y., Sabariah Md. Rashid, and Ain Nadzimah Abdullah. 2016. A critical review of prominent theories of politeness. *Advances in Language and Literary Studies*, 7(6):262–270.
- Allen, James F., Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI Magazine*, 22(4):27–38.
- Ariel, Mira. 2010. *Defining Pragmatics*. Cambridge, Cambridge University Press.
- Ariel, Mira. 2012. Research paradigms in pragmatics. In K. Allan and K. M. Jaszczolt K. M. Jaszczolt, editors, *The Cambridge Handbook of Pragmatics*. Cambridge, Cambridge University Press, pages 23–46.
- Austin, John Langshaw. 1962. *How to Do Things with Words*. Oxford, Clarendon Press.
- Bazzanella, Carla. 1995. I segnali discorsivi. In L. Renzi, G. Salvi, and A. Cardinaletti, editors, *Grande grammatica italiana di consultazione*, vol. 3 (*Tipi di frase, deissi, formazione delle parole*). il Mulino, Bologna, pages 225–257.
- Bazzanella, Carla. 2008. *Linguistica e pragmatica del linguaggio. Un'introduzione*. Laterza.
- Bennett, Greg. 2018. Conversational style: Beyond the nuts and bolts of conversation. In R. J. Moore, R. Arar, H. Szymanski M., and G. Ren G. Ren, editors, *Studies in Conversational UX Design*. Springer, International Publishing, pages 161–180.
- Berretta, Monica. 1984. Connettivi testuali in italiano e pianificazione del discorso. In *Linguistica testuale. Atti del XV congresso internazionale della Società di Linguistica Italiana*. pages 237–254.
- Bianchini, Alessia, Francesco Tarasconi, Raffaella Ventaglio, and Mariafrancesca Guadalupi. 2017. “gimme the usual” - how handling of pragmatics improves chatbots. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, pages 30–35, Roma, Italy, December.
- Bickmore, Timothy W., Ha Trinh, Reza Asadi, and Stefán Ólafsson. 2018. Safety first: Conversational agents for health care. In Robert J. Moore, Raphael Arar, Margaret H. Szymanski, and Guang-Jie Ren, editors, *Studies in Conversational UX Design*. editors, Springer International Publishing, pages 161–180.
- Brandtzæg, Petter Bae and Asbjørn Følstad. 2017. Why people use chatbots. In I. Kompatsiaris, J. Cave, A. Satsiou, G. Carle, E. Kontopoulos, S. Diplaris, and D. McMillan and, editors, *Internet Science. INSCI 2017 (Lecture Notes in Computer Science)*. volume 10673, Springer, pages 377–392.
- Brandtzæg, Petter Bae and Asbjørn Følstad. 2018. Chatbots: changing user needs and motivations. *Interactions*, 25(5):38–43.
- Branigan, Holly P., Martin J. Pickering, and Alexandra A. Cleland. 2000. Syntactic coordination in dialogue. *Cognition*, 75(2):B13–25.
- Braun, Daniel, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany, August.
- Braun, Daniel and Florian Matthes. 2019. Towards a framework for classifying chatbots. In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 484–489, Heraklion, Greece, May.
- Brown, Gillian and George Yule. 1983. *Discourse Analysis*. Cambridge University Press.
- Brown, Penelope. 1976. Women and politeness: A new perspective on language and society. *Reviews in Anthropology*, 3(3):240–249.
- Brown, Penelope and Stephen C. Levinson. 1987. *Politeness - Some universals in language usage*. Cambridge University Press.

- Bublitz, Wolfram and Neal R. Norrick. 2011. Introduction: The burgeoning field of pragmatics. In W. Bublitz and N. R. Norrick, editors, *Foundations of Pragmatics*. vol. 1 of Handbooks of Pragmatics, Mouton de Gruyter, pages 1–20.
- Candello, Heloisa and Claudio Pinhanez. 2018. Recovering from dialogue failures using multiple agents in wealth management advice. In R. J. Moore, M. H. Szymanski, R. Arar, and G. Ren G. Ren, editors, *Studies in Conversational UX Design*. Springer, International Publishing, pages 139–160.
- Canonico, Massimo and Luigi De Russis. 2018. A comparison and critique of natural language understanding tools. In *CLOUD COMPUTING: The Ninth International Conference on Cloud Computing, GRIDs, and Virtualization*, pages 110–115, Barcelona, Spain, February.
- Cardinal, Alison, Laura Gonzales, and Emma J. Rose. 2020. Language as participation: Multilingual user experience design. In *Proceedings of the 38th ACM International Conference on Design of Communication (SIGDOC '20)*, Denton, TX, October.
- Carnap, Rudolf. 1938. Foundations of logic and mathematics. In O. Neurath, R. Carnap, and C. W. Morris and, editors, *International Encyclopedia of Unified Science, vol. I*. University of Chicago Press, pages 139–214.
- Chaves, A. P., E. Doerry, J. Egbert, and M. Gerosa. 2019. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 102–109, Kyoto, Japan, May.
- Cho, HyunKyoung. 2015. Toward a new design philosophy: Politics and the aesthetic of “we” human-and-technology in interaction design. In C. Stephanidis, editor, *HCI International 2015 - Posters' Extended Abstracts. HCI 2015. Communications in Computer and Information Science*. vol 528, Springer, Cham, pages 13–18.
- Cho, Hyunkyoung and Joonsung Yoon. 2013. Toward a new design philosophy of hci: Knowledge of collaborative action of “we” human-and-technology. In M. Kurosu, editor, *Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments. HCI 2013. Lecture Notes in Computer Science*. vol 8004, Springer, Berlin, pages 32–40.
- Clark, Herbert H. 1996. *Using Language*. Cambridge University Press.
- Clark, Leigh, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, pages 1–12, Glasgow, Scotland, May.
- Cohen, Michael, James P. Giangola, and Jennifer Balogh. 2004. *Voice User Interface Design*. Addison Wesley.
- Colby, Kenneth Mark, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.
- Crystal, David. 2001. *Language and the Internet*. Cambridge University Press.
- Dale, Robert. 2016. Industry watch the return of chatbots. *Natural Language Engineering*, 22(5):811–817.
- Dasgupta, Ritwik. 2018. *Voice User Interface Design. Moving from GUI to Mixed Modal Interaction*. Apress.
- de Jong, Markus, Mariët Theune, and Dennis Hof. 2008. Politeness and alignment in dialogues with a virtual guide. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, Estoril, Portugal, May.
- Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *CoRR*, abs/1905.04071.
- Dippold, Doris, Jenny Lynden, Rob Shrubsall, and Rich Ingram. 2020. A turn to language: How interactional sociolinguistics informs the redesign of prompt:response chatbot turns. *Discourse, Context & Media*, 37(10043):2.
- Drew, Paul. 2020. Turn design. In J. Sidnell and T. Stivers, editors, *The handbook of conversation analysis*. Wiley-Blackwell, Springer, Cham, pages 131–149.
- Feine, Jasper, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2020. Gender bias in chatbot design. In A. Følstad, T. Araujo, S. Papadopoulos, E. L. Law, O. Granmo, E. Luger, and P. B. Brandtzaeg, editors, *Chatbot Research and Design. CONVERSATIONS 2019. Lecture Notes in Computer Science*. vol 11970, Springer, Cham, pages 79–93.

- Fox, Barbara A., Sandra A. Thompson, Cecilia E. Ford, and Elizabeth Couper-Kuhlen. 2018. Conversation analysis and linguistics. In J. Sidnell and T. Stivers, editors, *Handbook of Conversation Analysis*. Oxford: Wiley-Blackwell, pages 726–740.
- Følstad, Asbjørn and Petter Bae Brandtzaeg. 2020. Users' experiences with chatbots: findings from a questionnaire study. *Qual User Exp*, 5:3.
- Følstad, Asbjørn, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? an exploratory interview study. In S. Bodrunova, editor, *Internet Science. INSCI 2018. Lecture Notes in Computer Science*. vol 11193, pages 194–208.
- Følstad, Asbjørn, Marita Skjuve, and Petter Brandtzaeg. 2019. Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design. In S. Bodrunova, O. Koltsova, A. Følstad, H. Halpin, P. Kolozaridi, L. Yuldashev, A. Smoliarova, and H. Niedermayer, editors, *Internet Science. INSCI 2018. Lecture Notes in Computer Science*. vol 11551, Springer, Cham, pages 145–156.
- Garfinkel, Harold. 1996. *Studies in ethnomethodology*. Polity Press.
- Gervits, Felix, Ravenna Thielstrom, Antonio Roque, and Matthias Scheutz. 2020. It's about time: Turn-entry timing for situated human-robot dialogue. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 86–96, virtual meeting, July.
- Givon, Talmy. 1979. *Syntax and Semantics 12: Discourse and Syntax*. Academic Press.
- Gnewuch, Ulrich, Stefan Morana, and Alexander Maedche. 2017. Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems (ICIS)*, Seoul, South Korea, December.
- Goffman, Erving. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Aldine Publishing Company, New York.
- Goffman, Erving. 1983. The interaction order. *American Sociological Review*, 48:1–17.
- Grice, H. Paul. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics*. Speech Acts. Academic Press, 3.
- Grice, H. Paul. 1989. *Studies in the Way of Words*. Harvard University Press.
- Gumperz, John J. 1982. *Discourse Strategies*. Cambridge University Press.
- Hakulinen, Auli. 2009. Conversation types. In S. D'hondt, J. Ostman, and J. Verschueren, editors, *The Pragmatics of Interaction*. John Benjamins, pages 55–65.
- Hall, Erika. 2018. *Conversational design*. A Book Apart.
- Halliday, Michael A. K. 1994. *An Introduction to Functional Grammar*. Edward Arnold.
- Haugh, Michael. 2012. Conversational interaction. In K. Allan and K. M. Jaszczolt, editors, *The Cambridge Handbook of Pragmatics*. Cambridge University Press, pages 251–274.
- Herring, Susan C. 2012. Grammar and electronic communication. In C. Chapelle, editor, *Encyclopedia of applied linguistics*. Wiley-Blackwell.
- Hill, Jennifer, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49:245–250.
- Hirst, Graeme. 2001. Does conversation analysis have a role in computational linguistics? *Computational Linguistics*, 17(2):211–272.
- Hoey, Elliott M. and Kobin H. Kendrick. 2017. Conversation analysis. In A. M. B. de Groot and P. Hagoort, editors, *Research Methods in Psycholinguistics: A Practical Guide*. Wiley, Blackwell, pages 151–173.
- Horn, Laurence R. and Gregory Ward. 2006. Introduction. In L. R. Horn and G. Ward, editors, *The Handbook of Pragmatics*. Blackwell, pages xi–xix.
- Hussain, Shafquat, Omid Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In L. Barolli, M. Takizawa, F. Xhafa, and T. Enokido, editors, *Web, Artificial Intelligence and Network Applications. WAINA 2019. Advances in Intelligent Systems and Computing*. vol 927. Springer, Cham.
- Hymes, Dell. 1962. The ethnography of speaking. In T. Gladwin and W. C. Sturtevant, editors, *Anthropology and Human Behavior*. Anthropological Society, pages 13–53.
- Jacquet, Baptiste, Jean Baratgin, and Frank Jamet. 2018. The gricean maxims of quantity and of relation in the turing test. In *Proceedings of the 11th International Conference on Human System Interaction - HSI 2018*, pages 332–338, Gdańsk, Poland, June.
- Jacquet, Baptiste, Jean Baratgin, and Frank Jamet. 2019b. Cooperation in online conversations: The response times as a window into the cognition of language processing. *Frontiers in Psychology*, 10(727):1–15.

- Jacquet, Baptiste, Alexandre Hullin, Jean Baratgin, and Frank Jamet. 2019. The Impact of the Gricean Maxims of Quality, Quantity and Manner in Chatbots. In *Proceedings of the International Conference on Information and Digital Technologies*, pages 180–189, Zilina, Slovakia, June.
- Jain, Mohit, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the Designing Interactive Systems Conference*, pages 895–906, Hong Kong, June.
- Jefferson, Gail D. 1978. Sequential aspects of storytelling in conversation. In J. Schenkein, editor, *Studies in the Organization of Conversational Interaction*. Academic Press, pages 219–48.
- Katz, Jerrold and Jerry Fodor. 1963. The structure of a semantic theory. *Language*, 39:67–80.
- Kerbrat-Orecchioni, Catherine. 2001. *Les actes de langage dans le discours. Théorie et fonctionnement*. Nathan.
- Kvale, Knut, Olav Sell, Stig Hodnebrog, and Asbjørn Følstad. 2019. Improving conversations: Lessons learnt from manual analysis of chatbot dialogues. In A. Følstad, T. Araujo, S. Papadopoulos, E. L. Law, O. Granmo, E. Luger, and P. B. Brandtzaeg, editors, *Chatbot Research and Design. CONVERSATIONS 2019. Lecture Notes in Computer Science*. vol 11970, Springer, pages 187–200.
- Leech, Geoffrey. 1983. *Principles of Pragmatics*. Longman.
- Levinson, Stephen C. 1983. *Pragmatics*. Cambridge University Press.
- Linell, Per. 2001. *Approaching Dialogue: Talk, interaction and contexts in dialogical perspectives*. John Benjamins Publishing Company.
- Linell, Per and Thomas Luckmann. 1991. Asymmetries in dialogue: some conceptual preliminaries. In I. Marková and K. Foppa, editors, *Asymmetries in Dialogue*. Harvester Wheatsheaf, pages 1–20.
- Liu, Xingkun, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *arXiv 1903.05566v3*.
- Lotze, Netaya. 2016. Chatbots. eine linguistische analyse. In J. Runkehl, P. Schlobinski, and T. Siever, editors, *Sprache – Medien – Innovationen*. vol. 9. Peter Lang.
- Luckmann, Thomas. 1990. Social communication, dialogue and conversation. In I. Marková and K. Foppa, editors, *The Dynamics of Dialogue*. Harvester Wheatsheaf, pages 45–61.
- Luff, Paul, David Frohlich, and Nigel Gilbert. 1990. *Computers and Conversation*. Academic Press.
- Mathur, Vinayak and Arpit Singh. 2018. The rapidly changing landscape of conversational agents. *arXiv 1803.08419v2*, pages 1–14.
- McTear, Michael. 2020. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool.
- Mey, Jacob L. 1993. *Pragmatics. An introduction*. Blackwell.
- Michael, Thilo and Sebastian Möller. 2020. Simulating turn-taking in conversations with delayed transmission. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 157–161, virtual meeting, July.
- Moore, Robert J. 2013. Ethnomethodology and conversation analysis: empirical approaches to the study of digital technology in action. In S. Price, C. Jewitt, and B. Brown, editors, *The SAGE handbook of digital technology research*. SAGE Publications Ltd, pages 217–235.
- Moore, Robert J. and Raphael Arar. 2018. Conversational ux design: An introduction. In R. J. Moore, M. H. Szymanski, R. Arar, and G. Ren, editors, *Studies in Conversational UX Design*. Springer, International Publishing, pages 1–16.
- Moore, Robert J. and Raphael Arar. 2019. *Conversational UX Design. A Practitioner's Guide to the Natural Conversation Framework*. Association for Computing Machinery.
- Moore, Robert J., Raphael Arar, and Guang-Jie Ren. 2018. *Studies in Conversational UX Design*. Springer, International Publishing.
- Morris, Charles W. 1938. *Foundations of the Theory of Signs*, Chicago University Press; reprinted in *Writings on the General Theory of Sign*. Mouton.
- Nordheim, Cecilie Bertinussen, Asbjørn Følstad, and Cato Alexander Bjørkli. 2019. An initial model of trust in chatbots for customer service—findings from a questionnaire study. *Interacting with Computers*, 31(3):317–335.
- Pallotti, Gabriele. 2007. Conversation analysis: Methodology, machinery and application to specific settings. In H. Bowles and P. Seedhouse, editors, *Conversation Analysis and Language for Specific Purposes*. Peter Lang, pages 37–68.
- Pearl, Cathy. 2016. *Designing Voice User Interfaces. Principles of Conversational Experiences*. O'Reilly.
- Peirce, Charles Sanders. 1932. *Collected Papers*. Cambridge University Press.

- Piccolo, Lara, Martino Mensio, and Harith Alani. 2018. Chasing the chatbots: Directions for interaction and design research. In S. Bodrunova, editor, *Internet Science, INSCI. Lecture Notes in Computer Science*, Springer, page 157–169.
- Pickering, Martin and Simon Garrod. 2004. Towards a mechanistic psychology of dialogue. *Behavioural and Brain sciences*, 27:169–225.
- Pistolesi, Elena. 2018. Storia, lingua e varietà della comunicazione mediata dal computer. In G. Patota and F. Rossi F. Rossi, editors, *L'italiano e la rete, le reti per l'italiano*. Accademia della Crusca – GoWare, pages 16–34.
- Przegalinska, Aleksandra, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2020. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, 62:785–797.
- Radziwill, Nicole M. and Morgan C. Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *arXiv 1704.04579*.
- Ruane, Elayne, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and ethical considerations. In *Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland, December.
- Sacks, Harvey. 1984. Notes on methodology. In J. M. Atkinson and J. C. Heritage, editors, *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, pages 21–27.
- Sacks, Harvey and Emanuel A. Schegloff. 2007. Two preferences in the organization of reference to persons in conversation and their interaction. In N. J. Enfield and T. Stivers, editors, *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives, Language Culture and Cognition*. Cambridge University Press, pages 23–28.
- Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for organization of turn-taking for conversation. *Language*, 50:696–735.
- Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1992. *Lectures on Conversation*, volume I, II. Blackwell.
- Saygin, Ayse Pinar and Ilyas Cicekli. 2002. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34:227–258.
- Schegloff, Emanuel A. 1968. Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095.
- Schegloff, Emanuel A. 1992. To searle on conversation: a note in return. In J. Searle, H. Parret, and J. Verschueren, editors, *(On) Searle on Conversation*. Benjamins, pages 113–128.
- Schegloff, Emanuel A. 1999. Discourse, pragmatics, conversation, analysis. *Discourse Studies*, 1:405–435.
- Schegloff, Emanuel A. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*, volume I. Cambridge University Press.
- Schegloff, Emanuel A., Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.
- Schegloff, Emanuel A. and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8:289–327.
- Schiavo, Gianluca and Ahmed Fadhil. 2020. Designing for health chatbots. *arXiv 1902.09022*.
- Searle, John R. 1969. *Speech Acts. An Essay in the Philosophy of Language*. Cambridge University Press.
- Searle, John R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.
- Shawar, Bayan Abu and Eric Atwell. 2007. Chatbots: Are they really useful? *LDV-Forum*, 22(1):29–49.
- Shevat, Amir. 2017. *Designing Bots: Creating Conversational Experiences*. O'Reilly.
- Spence, Jocelyn. 2016. *Performative Experience Design*. Springer, Cham.
- Sperber, Dan and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*. Blackwell.
- Strengers, Yolande and Jenny Kennedy. 2020. *Why Siri, Alexa, and Other Smart Home Devices Need a Feminist Reboot*. MIT Press.
- Svenningsson, Nina and Montathar Faraon. 2019. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2nd Artificial Intelligence and Cloud Computing Conference*, pages 151–161, Kobe, Japan, December.
- Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: a corpus linguistics analysis of persistence in spoken language. *Corpus Linguistics and Linguistic Theory*, 1(1):113–150.
- Szymanski, Margaret H. and Robert J. Moore. 2018. Adapting to customer initiative: Insights from human service encounters. In R. J. Moore, M. H. Szymanski, R. Arar, and G. Ren G. Ren,

- editors, *Studies in Conversational UX Design*. Springer, International Publishing, pages 19–32.
- Tannen, Deborah. 1984. *Conversational Style: analyzing talk among friends*. Ablex.
- Tsvetkova, Milena, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri. 2017. Even good bots fight: The case of wikipedia. *PLOS ONE*, 12(2):1–27.
- Turner, Ken, editor. 1999. *The Semantics/Pragmatics Interface from Different Points of View*. Elsevier.
- van Dijk, Teun A. 2009. *Society and Discourse: How Social Contexts Influence Text and Talk*. Cambridge University Press.
- Verschueren, Jef. 1999. *Understanding Pragmatics*. Arnold.
- Wallace, Richard S. 2003. *The Elements of AIML Style*. A.L.I.C.E. Artificial Intelligence Foundation Inc.
- Weizenbaum, Joseph. 1966. ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- West, Mark, Rebecca Kraut, and Han Ei Chew. 2019. *I'd Blush if I Could. Closing Gender Divides in Digital Skills through Education*. UNESCO.
- Wooffitt, Robin, Norman Fraser, Nigel Gilbert, and Scott McGlashan. 1997. *Humans, Computers and Wizards. Analyzing human (simulated) computer interaction*. Routledge.
- Yule, George. 1996. *Pragmatics*. Oxford University Press.
- Zamora, Jennifer. 2017. I'm sorry, Dave, I'm afraid I can't do that: Chatbot perception and expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction, HAI '17*, page 253–260, New York, NY, USA. Association for Computing Machinery.
- Zubania, Matteo, Luca Sigalini, Ivan Serina, and Alfonso Emilio Gerevini. 2020. Evaluating different natural language understanding services in a real business case for the italian language. In *Proceedings of the 24th International Conference of Knowledge-Based and Intelligent Information & Engineering Systems*, pages 995–1004, Verona, Italy, September.

Improving Data-to-Text Generation via Preserving High-Frequency Phrases and Fact-Checking

Ethan Joseph*
Rensselaer Polytechnic Institute

Julian Lioanag**
Rensselaer Polytechnic Institute

Mei Si†
Rensselaer Polytechnic Institute

Transforming numerical data into natural language descriptions (data-to-text) requires presenting the data in the correct context, supplementing plausible details, and creating an overall coherent and non-conflicting narrative. In this work, we propose a generate-extract-correct pipeline for the task. We use transfer learning with an auxiliary task of keeping high-frequency word sequences from the training data for text generation. We then apply information extraction to the generated text to check its accuracy, followed by correction, and thus ensure the coherence of the generated narrative. We demonstrate the effectiveness of this approach with both objective and subjective evaluations. Using an empirical evaluation, we show that people rated our system's outputs similarly to human-written text regarding its coherence, conciseness, and grammar.

1. Introduction

In recent years, there has been an increasing interest in automatically generating text descriptions or dialogue from structured data (Puduppully, Dong, and Lapata 2019; Wiseman, Shieber, and Rush 2017; Rebuffel et al. 2020; Kale and Rastogi 2020). Data-to-text, broadly speaking, refers to tasks where a system is provided with data in a machine-readable format, e.g., RDF or tabular data, and needs to produce human-readable text based on the data. Because data-to-text techniques can enable machines to communicate with people in a natural, narrative way, they have enormous potential for real-world applications, especially with the fast development of semantic web, knowledge graph, and automated data analysis tools in recent years.

Given its primary function of communicating data with people, we infer three desiderata for data-to-text generation techniques. First of all, the data-to-text generation needs to ensure it conveys accurate information. This requires providing correct data and avoiding confusion in the generated text. Confusion can come from multiple sources, including redundancy and inconsistency in information, violating common sense, and incorrect grammar. Since people read sequentially, data-to-text generation can be viewed as an iterative grounding process, where the beginning part should ground the latter part of the generated text. Secondly, the generated narrative needs to be relatively concise while not hurting readability. Being concise can help deliver

* Dept. of Computer Science - Rensselaer Polytechnic Institute. E-mail: josepe2@rpi.edu

** Dept. of Computer Science - Rensselaer Polytechnic Institute. E-mail: lioanj@rpi.edu

† Dept. of Cognitive Science - Rensselaer Polytechnic Institute. E-mail: sim@rpi.edu

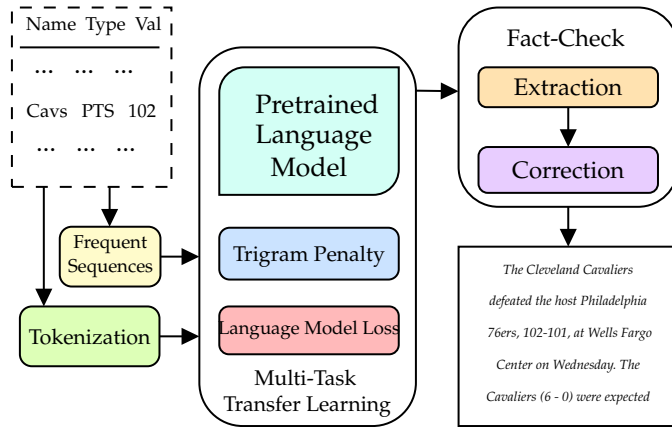


Figure 1
Overview of the approach

information more effectively and reduce the inclusion of data that does not exist in the input, and is hallucinated by the language model. Finally, the generated text should typically follow the same writing style and have the same topic and word choice preferences as the training examples to provide a familiar reading experience.

Multiple datasets have been used for exploring the data-to-text task, including RotoWire (Wiseman, Shieber, and Rush 2017), WebNLG (Gardent et al. 2017), and E2E (Novikova, Dušek, and Rieser 2017). We choose to work with the RotoWire dataset. The RotoWire dataset contains statistics of NBA basketball games with corresponding human-written narratives. This dataset presents a unique challenge by requiring models to form relatively long narrative descriptions with many numbers embedded in them (14.25 sentences and 25.49 numbers per description on average). Previous works on this dataset utilized explicit content planning, attention, and copy mechanisms, but can still suffer from insufficient fluency and accuracy in the generated text. Both hurt peoples’ reading experience and prevent them from understanding the data without confusion. Section 2 summarizes related work, and Section 3 discusses the imperfections in their generated text.

We propose a three-step generate–extract–correct pipeline for the data-to-text task as shown in Figure 1. This model helps to ground the generated text by emphasizing its accuracy and reducing potential confusion. It does so by having specific fact-checking and correction procedures after text generation. For generating the text, we investigate two techniques for enhancing transfer learning-based data-to-text techniques. First, to improve the language model’s capacity to learn the local structure and word choices from the training data, we mine high-frequency trigrams from the training data. During the transfer learning process, we add an auxiliary task of learning these trigram word combinations. This technique helps our model produce text written in the same style as the training samples, and hence, helps the readers comprehend them. Secondly, instead of directly outputting the generated text, we employ an extract-correct post-processing step to improve the generated text’s accuracy. First, information extraction is applied to the generated text for retrieving the information mentioned in it. The retrieved contents are then compared with the input data for checking their accuracy. If mistakes are found, they are fixed in the subsequent correction step. This can significantly reduce generation errors introduced by the pre-trained language models.

2. Related Work

Early approaches to data-to-task have relied on domain-specific knowledge and curation by experts. Such techniques can generate coherent narratives, but suffer from lacking flexibility and variations in the generated text. These approaches often involve developing complex rule-based templates in collaboration with experts in the field, as in (Reiter et al. 2005). More recently, deep learning techniques have been employed to encode data records into a semantic vector space, which can then be decoded and translated into output summaries. Early work in deep-learning-based data-to-text models often linearizes the input records, encoding them as a sequence of facts. (Wiseman, Shieber, and Rush 2017) shows the limitations of using recurrent architectures on such large structured data, which often fails to capture long-term relationships in the data. More recently and in contrast to the practice of linearly encoding records, (Puduppully, Dong, and Lapata 2019; Rebuffel et al. 2020) have used more complex schemes to encode input records, taking into account content planning and the structure of the input records. These models focus on end-to-end training and utilize planning or attention mechanisms, arguing that the previous linear encoding of input records has prevented models from extracting meaningful relationships hidden in the data.

Many recent advances in natural language processing have been attributed to the Transformer architecture (Vaswani et al. 2017), which not only have a strong language comprehension capacity but are also able to leverage language modeling skills to generate fluent text (Radford et al. 2019). Transfer learning, in which models are pre-trained on an unrelated, data-rich task, and later finetuned on a downstream task, has been shown to be very effective in many tasks (Raffel et al. 2020). In particular, (Kale and Rastogi 2020) demonstrates that finetuning the T5 model outperformed many other multi-stage pipelined approaches in three data-to-text benchmarks. The tasks in (Kale and Rastogi 2020) only require short-scale generations. In contrast, the RotoWire dataset contains longer narrative descriptions with many numbers (average of 24), posing a very different challenge.

The idea of rewriting part of the generated text for achieving a better quality has been explored in a few works. (He, Peng, and Liang 2019) used rewriting to increase the "surprise" factor of a generated sentence, and thus make the sentences more fun to read. (Song et al. 2020) rewrites the generated dialogue to make its tone consistent with the speaker's personality profile. In this work, we apply the rewriting idea to improve the accuracy of the generated text.

This work seeks to combine multiple aspects of recent advances in data-to-text and broader text generation by performing a multi-task (Luong et al. 2016) transfer learning on transformer architectures for the data-to-text task, and by introducing a post-processing module to improve the accuracy of generated descriptions. In contrast to previous work, we argue that the transformer model would be good at extracting latent relationships in input data due to their strong language and understanding skills, even if that data is encoded linearly. Our results show we can dependently improve transfer learning for data-to-text tasks based on multiple language models, including T5 (Raffel et al. 2020).

3. Case Studies of Generation Errors

The sentences in the generated text need to be grounded in their context, i.e., they need to be accurate and consistent with each other. Unfortunately, because of the complexity of the task, existing models often cannot ensure self-consistency, contain

Table 1

Duplicate percentage, average numbers of records, sentences, erroneous records, and duplication per generated description on the test set. Compared between human written descriptions (**Gold**), [Wiseman, Shieber, and Rush 2017]’s template-based model (**Template**) and neural model (**WS-2017**), [Rebuffel et al. 2020]’s best model (**Heir-k**), [Puduppully, Dong, and Lapata 2019]’s best model (**NCP+CC**), and our best model (**Bart_{Tri+Fact}**).

Model	Dup %	# Rec	# Sent	# Err	DupSent
Gold	0.14 %	25.49	14.25	1.49	0.05
Template	0.01 %	54.26	8.11	0.59	0.88
WS-2017	30.58 %	45.18	15.19	11.23	1.69
Heir-k	13.34 %	32.61	14.10	6.38	0.21
NCP+CC	15.77 %	45.96	12.11	5.52	0.89
Bart _{Tri+Fact}	1.27 %	55.38	13.03	5.10	0.07

inaccurate records, and suffer other readability issues. This section provides examples of these challenges in generations achieved with previous SOTA models on the RotoWire dataset. Complete examples can be found in Appendix A.

3.1 Duplicate Information

A common issue with the generated text is that it includes redundant or repeated information. Take, for example, the following excerpt generated from the model defined in (Puduppully, Dong, and Lapata 2019):

Tristan Thompson *chipped in with seven points and 13 rebounds, marking his first double-double of the year.* *Tristan Thompson* *chipped in seven points and 13 rebounds as the starting power forward.* **Ersan Ilyasova had a solid game off the bench with 21 points (8-13 FG, 4-6 3Pt) and four rebounds.** *Gerald Henderson* *scored 11 (5-9 FG, 1-4 3Pt) and Ersan Ilyasova had a team-high of 21 points (8-13 FG, 4-6 3Pt) and grabbing four rebounds.* It was a season-high for Ilyasova, who hadn’t reached double figures in points twice this season. Gerald Henderson had 11 points (5-9 FG, 1-4 3Pt) as well.

Sets of duplicate information are highlighted with italics, boldface, and underlines respectively. To get an estimate of the number of semantically similar sentences in the generated descriptions, we run a simple cosine similarity test. Two sentences are considered duplicate if the cosine similarity of their average word2vec embeddings (Rehurek and Sojka 2011) is greater than 0.9. Using this technique, we get an average of 0.89 pairs of redundant sentences per description on the test set for (Puduppully, Dong, and Lapata 2019) (See Table 1 for full statistics), implying that almost every generated description has some form of duplicate information. Further, by extracting records from generated descriptions using an information extraction system, we see 15.77% duplicate records for (Puduppully, Dong, and Lapata 2019) and 30.58% for (Wiseman, Shieber, and Rush 2017) as shown in Table 1.

The duplication can affect the overall readability of the generated descriptions, impacting their coherency and conciseness. We address this issue by finetuning large transformer language models, which have been shown to generate consistent text with minimal duplicates.

3.2 Erroneous Information

In many cases, SOTA models generate sentences with erroneous information, such as records that didn't exist in the data, or incorrect scores. Below are excerpts from (Puduppully, Dong, and Lapata 2019) and (Rebuffel et al. 2020) that show these inconsistencies, highlighted in bold:

1. **Greg Beasley** led the bench with 17 points, two rebounds, two assists and one steal.
2. Kobe Bryant led the Lakers with **26** points (**10 - 20** FG, **2 - 4** 3Pt, **4 - 4** FT), 12 rebounds, four assists, one steal and one block in 38 minutes.
3. The Memphis Grizzlies (5 - 2) defeated the Phoenix Suns (3 - 2) Monday **1 - 2** at the Talking Stick Resort Arena in Phoenix.

In the first example, "Greg Beasley" is not an actual player in the NBA, and in the second and third examples, incorrect scores were generated. Using an information extraction system (see Section 5.1 for details) and comparing extracted records to actual input records, we found an average of 11.23 incorrect records per generated description for (Wiseman, Shieber, and Rush 2017)'s neural model, and 5.52 incorrect records per generated description for the (Puduppully, Dong, and Lapata 2019) model. We address this issue by post-processing generated descriptions and correcting erroneous information in an ad hoc fashion.

A more significant issue is that the text descriptions used in training often contain sentences that refer to information not existing in the input data and, therefore, are not grounded by data. For example, the text in Table 2 mentions, "The Sixers will return to action on Wednesday, when they host the Sacramento Kings for their next game." Data-to-text models often learn the "need" of adding sentences like this due to their prevalence in the text used for training. However, the RotoWire dataset does not contain data on each team's next match, so the model ends up making up the information in the generated text. This issue where models generate text but cannot relate it with real-world data is a severe limitation of many data-to-text models. Since writers often utilize information outside of the paired data in their writings, it is hard for machine learning models to address this without external knowledge. In (Reiter et al. 2005) where human authored templates are used for text generation, this problem is particularly avoided by generating more concise descriptions and including more real data in the generated text. As shown in Table 1, the average number of records mentioned in the generated text is 54.26 in (Reiter et al. 2005), while the average number of sentences used is only 8.11. In Gold, i.e., the training data, the average number of records is only 25.49, and the average number of sentences is 14.25. Other models typically also generate text that includes more records than Gold. Our model generates text with a very similar number of records as the template model. We believe, as a result, our generated text contains less made-up information and is more grounded.

3.3 Grammar and Consistency of Text

Sometimes, generated text can be awkwardly phrased, affecting readability. Excerpts from descriptions generated using the model from (Puduppully, Dong, and Lapata 2019) display this:

Table 2

Sample data-record table (top) paired with a truncated human-written news summary (bottom). Corresponding records are bolded.

Team	WIN	LOSS	PTS	DREB	FG3_PCT	...	
Raptors	11	6	122	34	68	...	
76ers	4	14	95	26	41	...	
Player	H/V	PTS	AST	REB	FG	TO	...
Carroll	H	10	3	5	4	0	...
Siakam	H	8	0	3	4	1	...
Henderson	V	0	2	1	0	1	...

The host Toronto **Raptors** defeated the Philadelphia **76ers**, **122 - 95**, at Air Canada Center on Monday. The Raptors came into this game as a monster favorite and they didn't leave any doubt with this result. Toronto just continuously piled it on, as they won each quarter by at least four points. The Raptors were lights-out shooting, as they went **55 percent** from the field and **68 percent** from three-point range. They also held the Sixers to just **42 percent** from the field and dominated the defensive rebounding, **34 - 26**. ... The Sixers will return to action on Wednesday, when they host the Sacramento Kings for their next game. ...

1. However, a standout effort in the second half was the play of the dynamic duo of **D'Angelo Russell and D'Angelo Russell**, who combined for 51 points on the night.
2. **Derrick Favors (knee) sat this one out with a sore back**, while Gordon Hayward returned ...
3. **The Pacers are now 2 - 3 in the first three games of their nine - game homestand**. They are now 2 - 3 on the road this season.

The first example duplicates the same entity in the same sentence. The last two examples contain contradictions: Derrick Favors injured his knee but sat out with a sore back, and the Pacers' win-loss ratio is 2 - 3, when they are described to have only played three games so far. We address this issue by adding an auxiliary objective while finetuning the language model, which is designed to help the model keep high-frequency sequences of words together and better learn the writing styles of professional sports summaries.

3.4 Balance of Statistics vs. Descriptors

While it is often beneficial for the descriptions to include many statistics about a game, there has to be a balance between the number of records and descriptive sentences about the game or the players. If a description contains too many records, it can often feel like reading a wall of data, in which case the information would be better conveyed through a table. However, if there are too few records, readers may not be satisfied. (Wiseman, Shieber, and Rush 2017)'s Template model skews towards "wall of data", containing over 54 records in only 8.11 sentences on average (Table 1). We rely on our language model and the auxiliary training objective to learn the correct ratio of records to descriptors.

4. Proposed Approach

To address the challenges presented in data-to-text generation, we decompose the generation task into three steps: generation, information extraction, and correction. The generation pass involves four steps as shown in Figure 1:

1. Tokenize the input record table R , extract frequent sequences (trigrams) from R .
2. Use transfer learning to finetune a language model with an auxiliary task of learning high-frequency word sequences from training data via trigram penalty.
3. Feed tokenized input into the finetuned language model, generate text y .
4. Feed y into the Fact-Check module, receive final text y' .

We hypothesize that a pretrained language model will be able to overcome the duplication and fluency challenges identified in Sections 3.1 and 3.3. We test our approach on three state of the art language models: T5 (Raffel et al. 2020), Bart (Lewis et al. 2020), and Pegasus (Zhang et al. 2019). T5 was employed by (Kale and Rastogi 2020) for data-to-text tasks. BART and Pegasus are selected because we believe their BERT style bi-directional encoders can efficiently attend to our input records, and their GPT-2 style auto-regressive decoders are ideal for generating fluent text.

While our generation system is not trained end-to-end, it is automated and does not require human intervention during execution. We also argue for its simplicity. Retraining the system for working with another dataset will only require a quick finetuning pass (averaging about 1 hour on an Nvidia Titan RTX) rather than the full training process from scratch.

4.1 Tokenization and Notation

To pass a table of records R to a language model, we first tokenize the data by prefacing records with “special field” tokens. Our finetuning pass then optimizes a cross-entropy loss between the model’s output y and professionally written text \hat{y} , with an added auxiliary task for learning high-frequency word sequences in the training data.

Adopting the notation from (Puduppully, Dong, and Lapata 2019), the input to our model, R , is a table of records from match m (see the top of Table 2 for an example.) Each data record r_j has 5 features: the entity which it belongs to ($r_{j,1}$; e.g. Cavaliers, Stephen Curry), its value ($r_{j,2}$; e.g. 102, Golden State), its relation type ($r_{j,3}$; e.g. POINTS, REBOUNDS), whether the record belongs to the home or away team ($r_{j,4}$; HOME or AWAY), and whether the record belongs to a team or a player ($r_{j,5}$; TEAM or PLAYER), represented as $\{r_{j,k}\}_{k=1}^5$. The total number of records is given by $|R|$. The output y is a text description of R containing words $y_1 \cdots y_{|y|}$, where $|y|$ is the length of the text. The gold text description paired with each R in the dataset is then \hat{y} . See Table 2 for an example record table (top) and paired text description (bottom).

Records and descriptions are tokenized using byte-pair encoding (BPE). To model each record r_j , we introduce multiple special field tokens that each correspond to a specific record relation type and whether the record belongs to a team or a player ($r_{j,3}$ and $r_{j,5}$). This ensures that the representation for a record type is never split by the

tokenizer, and reduces the total size of our tokenized input (at the cost of a slightly increased vocabulary), allowing us to pass a longer context to the model.

For each match m 's table of records R , we start by tokenizing the team-level records such as team-wins and team-points, then we follow with the records for all the players. We also add special "HOME" and "AWAY" tokens that separate each new entity and gives the model information about which team each record belongs to ($r_{j,4}$). For each match, we first convert the table of records to an easily tokenizable string. For example, the table of records given by the top part of Table 2 would be converted to the following string:

```
<|HOME|> Raptors <|TM-PTS|> 122 <|TM-REB|> 42
<|TM-AST|> 22 <|TM-WINS|> 11 <|TM-LOSSES|> 6
<|AWAY|> 76ers <|TM-PTS|> 95 <|TM-REB|> 38
... ..
```

4.2 FineTuning and Trigram Penalty

Using the tokenized dataset, we finetune large transformer models to generate the text description y given the tokenized input R . Finetuning is the process of taking a model that was initially trained (pre-trained) on a large dataset, and further training (finetuning) it using the same objective on a different and typically smaller dataset. This is known as transfer learning (Raffel et al. 2020), and it allows the language model to learn vocabulary, grammar, structure, and linguistic features of language during the pre-training step on a vast amount of data, then further finetuning on the data-to-text dataset allows the model to generate fluent and consistent text with linguistic features of the new dataset (in our case, the descriptions y).

Algorithm 1: Trigram loss penalty

```
Data:  $x, x_{t-1}, x_{t-2}, TG$ 
Result: Penalty:  $p$ 
 $p \leftarrow 1$ ;
foreach  $R \in TG$  do
  if  $(x_{t-2}, x_{t-1}) \in R$  then
    if  $x \notin R$  then
       $p \leftarrow p + 1$ ;
    end
  end
end
```

To better learn common phrases in the paired gold text descriptions, we add an auxiliary task during the finetuning step to increase the likelihood of generating word sequences frequent in \hat{y} . We hypothesize that this objective can help the language model generate text that more closely follow the language patterns in \hat{y} , and improve the frequency of expressions commonly seen in the training data. After some experimentation, we chose to focus on trigram sequences. To generate a list of frequent trigrams, we comb through gold human written text in the training set, create a count of each sequence of 3 words, and choose the 100 most common sequences. For now, we ignore any word sequences that contain data records, e.g. "scored 2 points". Examples of enforced trigrams include: "double-digit favorite", "led the way", "triple - double", and "of the

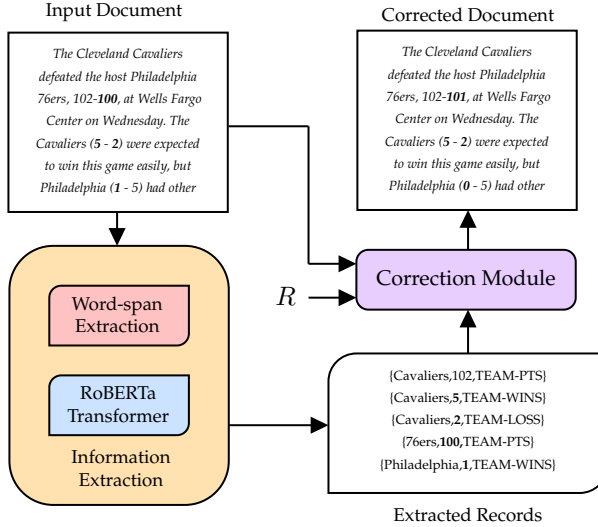


Figure 2
Extraction and Correction

season". Because of the inclusion of such word sequences, this task may indirectly help the model with topic selection as well.

During finetuning, we minimize a cross-entropy loss with label smoothing (Pereyra et al. 2017) combined with a penalty factor that scales the loss if frequent trigrams aren't being generated (or are only partially generated). Given target word y , output token x , the previous two tokens x_{t-1} and x_{t-2} and a list of frequent trigrams TG , we minimize

$$\mathcal{L}(x, x_{t-1}, x_{t-2}, y, TG) = \text{Cross Entropy}(x, y) + \alpha \log f(x, x_{t-1}, x_{t-2}, TG)$$

where α is a hyperparameter to scale the trigram penalty $f(x, x_{t-1}, x_{t-2}, TG)$, given by Algo. 1.

We also enforce a minimum and maximum length penalty. Output texts are generated using beam search with a beam size of 4, and we remove duplicate trigrams during the search to avoid repetition following (Paulus, Xiong, and Socher 2017).

To examine how robust this proposed auxiliary task is for improving the performance of transfer learning, we test our approach on three state-of-the-art language models: T5 (Raffel et al. 2020), Bart (Lewis et al. 2020), and Pegasus (Zhang et al. 2019) in Table 3.

4.3 Post-Processing Fact-Check

For post-processing, we employ a two-step, extract-and-correct process shown in Figure 2, relying on an information extraction system to extract records from the generated text y , then passing these extracted relations along with the accurate input R to a correction module that replaces those incorrect values in y .

4.3.1 Information Extraction

For the information extraction (IE) component of our fact-check system, we finetune a RoBERTa (Liu et al. 2019) transformer with a classifier head to predict $r_{j,3}$ (i.e. the relation type between an entity and value) given all pairs of word-spans in an input. Unrelated pairs are predicted as ε and ignored. Thus, the model learns to minimize

$$\mathcal{L}(\theta) = - \sum_j \log \sum p(r_{j,3} = r'_{j,3} \mid r_{j,1}, r_{j,2}; \theta)$$

for all text spans $\{r_{j,1}, v\}$. The training dataset for this task was developed in the same way as the IE dataset proposed in (Wiseman, Shieber, and Rush 2017). We programmatically extract text spans in the gold summaries by looping through each entity and number in each sentence, then search the records in R for a relation $r_{j,3}$ that corresponds to the extracted span. If an entity and number aren't found together in R , we give a ε label. A RoBERTa model trained on this dataset achieves an 83.6% accuracy when evaluated on the test set. This is sufficient to improve generations as part of the fact-check module, despite being lower than the 90% accuracy claimed by the CNN/LSTM ensemble IE system (Wiseman, Shieber, and Rush 2017) used to calculate the objective evaluation metrics.

4.3.2 Correction

Given a *sentence*, trained RoBERTa *model*, and corresponding table of records R , we begin the fact-check by looping through the sentence to extract text spans. For each word in tokenized *sentence*, we first check if the word corresponds to an entity $r_{j,1}$ in R . Next, we loop through each number v in the sentence and construct a span from the entity and that value. This span is passed to *model* along with *sentence*, which predicts the relation between the entity and value $r'_{j,3}$. Finally, we check the correct value of v (i.e. $r_{j,2}$) given $r_{j,1}$ and the predicted $r'_{j,3}$, and replace v in *sentence* with $r_{j,2}$ if the values diverge. This way, we can find sentences with incorrect values in y , replace the wrong values with the correct ones from corresponding input records, and finally rewrite to new output text y' . Pseudocode of the whole extraction/correction process is provided by Algo. 2.

5. Evaluation

We train and evaluate our model on the RotoWire data from the BoxScore dataset (Wiseman, Shieber, and Rush 2017). There are a total of 4853 distinct text descriptions covering basketball games played between 1/1/2014 and 3/28/2017. Each game is paired with an average of 628 records (with an average of 28 separate entities). The descriptions are relatively long, averaging 337 words in 14 sentences. We followed the same split introduced in the dataset, training on 3398 data/description pairs, using 727 for validation, and 728 for testing.

To show that the effectiveness of our approach is model-agnostic, we ran objective evaluations on each of the pretrained T5, Bart, and Pegasus models (See Table 3 for comparisons.) Note that these models follow the encoder-decoder transformer architecture. Tests with encoder or decoder only models such as BERT and GPT2 were unable to generate grammatical text for this task. We believe that having a separate encoder and decoder is ideal for the data-to-text task as it allows for the model to better learn an internal representation for the input records R , then separately focus on translating

Algorithm 2: Fact-Check

```

Data:  $R, model, sentence$ 
Result: Corrected:  $sentence$ 
 $p \leftarrow 1;$ 
 $S \leftarrow tokenize(sentence);$ 
foreach  $ent \in S$  do
  if  $ent \in [r_{j,1} \text{ for } j \in |R|]$  then
    foreach  $value \in S$  do
      if  $isNumber(value)$  then
         $span \leftarrow \{ent, value\};$ 
         $rel \leftarrow model.forward(S, span);$ 
         $r_{j,2} \leftarrow R[ent, rel];$ 
        if  $value \neq r_{j,2}$  then
           $sentence[value] \leftarrow r_{j,2}$ 
        end
      end
    end
  end
end
return  $sentence$ 

```

Table 3
Transformer Model Comparison on Test Set.

Model	RG		CS		CO
	#	P%	P%	R%	DLD%
T5 _{Base}	15.33	46.71	22.76	29.40	10.81
T5 _{Tri}	19.33	66.56	29.47	33.67	12.38
T5 _{Tri+Fact}	25.97	77.68	28.99	35.64	12.67
Psus _{Base}	21.06	54.88	24.85	37.61	14.52
Psus _{Tri}	31.31	72.92	28.53	48.43	16.21
Psus _{Tri+Fact}	33.31	87.17	31.67	47.33	17.06
Bart _{Base}	44.10	80.89	26.66	57.25	14.09
Bart _{Tri}	46.19	86.14	27.62	58.47	16.56
Bart _{Tri+Fact}	50.60	89.90	27.60	60.49	16.18

that representation into a text description y . The subscript **Base** models were trained without trigram loss and unprocessed. Subscript **Tri** models were trained with trigram loss but generated without fact-checking. Finally, subscript **Tri+Fact** models utilize our full pipeline, and were trained with trigram loss and processed with fact-checking.

In addition, we compared the performance of our best model with that of (Wiseman, Shieber, and Rush 2017)’s template-based (**Template**) model and neural model (**WS-2017**), (Puduppully, Dong, and Lapata 2019)’s best model (**NCP+CC**) and (Rebuffel et al. 2020)’s best model (**Heir-k**). Results on the test set can be found in Tables 4-5 (SOTA results bolded).

Table 4
Objective Evaluation on Test Set.

Model	RG		CS		CO
	#	P%	P%	R%	DLD%
Template	54.23	99.95	26.61	59.15	14.44
WS-2017	23.58	75.09	28.25	35.81	15.37
NCP+CC	34.12	88.12	34.49	51.13	18.66
Heir-k	22.83	79.22	34.12	37.88	17.10
Bart _{Tri+Fact}	<i>50.60</i>	<i>89.90</i>	27.60	60.49	16.18

5.1 Objective Evaluations

We evaluated model outputs on the validation and test sets using the metrics defined in (Wiseman, Shieber, and Rush 2017). These metrics use a neural ensemble IE system to extract records from gold description \hat{y} and our models' output y . This system ensembles the predictions from 3 CNN based architectures and 3 Bi-Directional LSTM based architectures trained to predict relations given all pairs of word-spans in an input. We then compared whether the extractions align or diverge from the gold summaries. The following metrics are used:

Relation Generation (RG): measures the “correctness” of the records extracted from y , as the proportion of extracted records that is also in R , given in terms of precision **P%** and number of unique generations **#**.

Content Selection (CS): measures how well y matches \hat{y} in terms of selecting which records to generate, as the proportion of records extracted from y that are also in \hat{y} , given in terms of precision **P%** and recall **R%**.

Content Ordering (CO): measures how well the order of records in y matches the order of records in \hat{y} , given as the normalized Damerau-Levenshtein Distance **DLD%** between records extracted from y and \hat{y} .

(Wiseman, Shieber, and Rush 2017) notes that CS primarily targets the “what to say” aspect of evaluation, CO focuses on the “when to say it”, and RG targets both.

In addition, we report BLEU, ROUGE-L, and METEOR, using paired human-written descriptions as a reference. A lot of work on this dataset only reports BLEU. Like BLEU, ROUGE-L and METEOR are commonly used metrics when evaluating automated text generation. ROUGE-L emphasizes recall, and METEOR has been shown to correlate better with human judgment and doesn't penalize using synonyms (Denkowski and Lavie 2014).

Table 3 shows that our proposed auxiliary task and post-processing procedures improved the performances of all three language models. Our pipeline improves evaluation results by 2.34% up to 26.4% on average comparing to finetuning the language models alone. Overall, the Bart_{Tri+Fact} model performed the best and is what we will use to compare to the previous state-of-the-art.

Table 4 shows that the Bart_{Tri+Fact} model performs better on RG# and RG P% than all other models except for the Template model. For CS, the Bart_{Tri+Fact} has higher recalls than all other models, including the Template model.

To further investigate these results, we computed the average length and the amount of duplication that exists in each model's output on the test set. The results are shown in Table 1. Bart_{Tri+Fact} only generated 0.07 pairs of duplicated sentences per

Table 5
BLEU, ROUGE-L, and METEOR Scores.

Model	Validation			Test		
	BLU	ROG	MET	BLU	ROG	MET
Gold	100.00	100.00	100.00	100.00	100.00	100.00
Template	8.97	18.54	21.67	8.93	18.59	21.38
WS-2017	14.57	23.00	31.44	14.19	22.86	31.39
NCP+CC	16.19	23.69	32.06	16.50	23.67	31.81
Heir-k	16.30	23.27	33.26	16.50	23.33	33.53
Bart _{Tri+Fact}	14.19	24.34	34.88	14.52	24.24	34.48

description and 1.27% of duplicate records. This is a notable improvement compared to other models. While the total numbers of sentences generated by these models are similar, with less duplication, the descriptions generated by Bart_{Tri+Fact} contain more unique records. This can explain why we have better results on RG metrics. Similarly, the more unique records can account for our higher CS recall. The fact that we have lower CS precision indicates our generated descriptions do not necessarily follow the same content plan that the gold descriptions use, and may generate more records that aren't mentioned in Gold. As shown in Table 1, generating more records than Gold is common; and having a higher number of records reduces the amount of made-up information in the generated text. In fact, our model generated a similar amount of records as the Template model. However, unlike Template, our model also generates sufficient descriptor text such that reading the generated descriptions doesn't feel like reading a wall of data, as shown by our conciseness and coherence scoring higher than Template in the subjective evaluations (Table 6). Therefore, we believe our generated text descriptions are better grounded for the readers.

As shown in Table 5, the Bart_{Tri+Fact} model has higher METEOR and ROUGE-L scores, but slightly lower BLEU when compared to other models. This suggests that the text generated from our models contains a lot of synonyms, which is expected when using a pretrained language model.

Interestingly, Bart_{Tri+Fact} improves the CS scores over the base Bart model in both test and validation sets, while in theory, the post-processing we perform should not affect content selection (CS). We suspect this may result from the IE model being able to extract more accurate information in the text generated by Bart_{Tri+Fact}. Further, the auxiliary task of learning high-frequency word sequences may have helped the model select more accurate records.

5.2 Subjective Evaluations

Using the same design as in (Puduppully, Dong, and Lapata 2019), we conducted a human evaluation study on Amazon Mechanical Turk (MTurk) to assess the subjectively perceived quality of the generated text. We randomly picked 30 basketball matches in the test set. We then asked crowd-workers to compare a human-written description (Gold), and descriptions generated by Template, NCP+CC, Heir-k, and our Bart_{Tri+Fact} with each other. For each game, we arranged the 5-tuple of generated description into pairs for comparison, resulting in 10 pairs. Each pair was shown to 3 different crowd-workers. They were asked to choose the *better* description according to:

Table 6
Results from Subjective Evaluations

Model	Grammar	Coherency	Conciseness
Gold	24.444	26.111	-3.889
Template	-48.889	-44.444	-1.667
NCP+CC	-10.000	-1.111	-1.111
Heir-k	10.556	1.667	1.667
Bart _{Tri+Fact}	20.000	13.889	0.000

Coherence: Is the summary easy to read? Does it follow a logical order?

Conciseness: Is the summary concise? Does it avoid redundancy and repetition?

Grammar: Does the summary read fluently? Does it use proper grammar?

All of these questions are important for people's subjective experiences of whether the generated text is well-grounded. We recruited 450 subjects. Each made two comparisons. This results in a total of 900 comparisons. We then calculated the score of a system for each criterion as the difference between the percentage of times it was chosen as the *better* one and the percentage of times it was chosen as the *worse* one. The scores range from -100 (absolute worst) to +100 (absolute best).

The results of this study are displayed in Table 6. The evaluations for Bart_{Tri+Fact} are similar to those for Gold with slightly lower Coherence and Grammar scores, but a better Conciseness score. Counting a score of 1 each time a description generated from an algorithm is selected, and 0 otherwise, we performed one-way ANOVA on the subjects' ratings of Grammar, Coherency, and Conciseness. The results show a significant difference ($p < .05$) among the subject's ratings for Grammar and Coherency, but not for Conciseness. We performed additional T-tests between the evaluations for Bart_{Tri+Fact} and other algorithms using two-tailed unpaired T-tests. At the .05 level, there is no significant difference between Bart_{Tri+Fact} and Gold or Heir-k. However, Bart_{Tri+Fact} did perform significantly better than NCP+CC and Template in regards to Grammar and Coherency. Template performs significantly worse in Coherence and Grammar, probably because of its restricted and rigid sentence templates. Overall, Bart_{Tri+Fact} was rated higher than the other generative models (Heir-k and NCP+CC). Our Conciseness is also slightly higher than every model except Heir-k. This may imply that our system strikes the right balance between data and descriptors.

6. Conclusion and Future Work

We aim at generating well-grounded text for the data-to-text task by emphasizing its accuracy, coherency, and conciseness. We propose a generate-extract-correct pipeline and incorporate an auxiliary task of learning high-frequency word sequences. Evaluations on the RotoWire dataset demonstrate the auxiliary task and the ad hoc extract-correction processes improved transfer learning performances using all three language models – BART, T5, and Pegasus. Subjective evaluation using mTurk show that the results generated by our model are comparable to Gold descriptions.

For future work, we want to look further into the consistency of the generated text. Minimally, the usage of transition phrases, e.g., "also" and "but" should be consistent with the conjunction or contradiction relationship between sub-sentences. Furthermore, the sentiment of a sentence, should be consistent with the comparison in it. This means

that the attitude towards subjects in a sentence should correlate with the generated text. For instance, if A defeats B, then A's score should be higher than B's. We are also interested in connecting this work with common sense reasoning. One limitation of work in this area is the generated text can only state factual information, but not offer any explanations while human written text often involves some form of explanations and inferences.

References

- Denkowski, Michael and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June. Association for Computational Linguistics.
- Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain, September. Association for Computational Linguistics.
- He, He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. 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 1734–1744, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Kale, Mihir and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland, December. Association for Computational Linguistics.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Luong, Minh-Thang, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Novikova, Jekaterina, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany, August. Association for Computational Linguistics.
- Paulus, Romain, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.
- Pereyra, Gabriel, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.
- Puduppully, Ratish, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6908–6915, Honolulu, Hawaii, USA, January 27 - February 1. AAAI Press.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Rebuffel, Clément, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. A hierarchical model for data-to-text generation. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances*

- in Information Retrieval*, pages 65–80, Cham. Springer International Publishing.
- Rehurek, Radim and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).
- Reiter, Ehud, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169, September.
- Song, Haoyu, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, Online, July. Association for Computational Linguistics.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, pages 6000–6010, Long Beach, California, USA, December. Curran Associates Inc.
- Wiseman, Sam, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Appendix A: Qualitative Comparisons

Below we provide two examples of outputs generated by different systems and manually mark issues identified in Section 3. Erroneous information (3.1) in **red**, duplicate information (3.2) in **blue**, inarticulate/illogical sentences (3.3) in **green**, and inconsistent sentences (3.3) in **orange**.

Note that there is a small amount of error, even in gold. For duplication, we only marked the places where the information appeared at its second or third times. For spotting inconsistencies, the context needs to be taken into consideration. For example, in the NCP+CC version of the first example, it first says the game’s date is Friday, then, later on, says it is Monday and then Tuesday. Similarly, in the Bart_{Tri+Fact} version, the Raptors are said to play with both the Knicks and the Nets. In the NCP+CC version of the second example, Middleton never surpassed the 20-points mark based on prior information in the paragraph.

Gold – Summary 1

DeMar DeRozan and Terrence Ross combined for a whopping 55 points on 17-of-28 shooting, helping lead the Raptors to outstanding **51** and 56 percent success rates from the field and three-point range, respectively. Kyle Lowry went for 18 points, while DeMarre Carroll supplied 13 points of his own. Jonas Valanciunas registered an 11-point, **13**-rebound double-double. The top-heavy Bucks got their own stellar 30-point performance from Giannis Antetokounmpo, while Jabari Parker followed with 27 points. However, only one other Bucks player managed to get into double digits in the scoring column, and that effort came from the bench in the person of Greg Monroe (11 points). Tony Snell (eight points), John Henson (seven points) and Matthew Dellavedova (two points) considerably underwhelmed, while Milwaukee particularly struggled from long range, shooting just 17 percent from beyond the arc.

Template – Summary 1

The Toronto Raptors (17-7) defeated the Milwaukee Bucks (11-12) 122-100. Giannis Antetokounmpo scored 30 points (13-23 FG, 0-3 3Pt, 4-6 FT) to go with 9 rebounds. DeMar DeRozan scored 30 points (7-11 FG, 1-1 3Pt, 15-15 FT) to go with 3 rebounds. Jabari Parker scored 27 points (11-17 FG, 2-3 3Pt, 3-5 FT) to go with 3 rebounds. Terrence Ross scored 25 points (10-17 FG, 4-6 3Pt, 1-3 FT) to go with 5 rebounds. Kyle Lowry scored 18 points (5-9 FG, 4-7 3Pt, 4-4 FT) to go with 3 rebounds. DeMarre Carroll scored 13 points (5-12 FG, 3-7 3Pt, 0-0 FT) to go with 6 rebounds. The Bucks’ next game will be at home against the Dallas Mavericks, while the Raptors will travel to play the Bulls.

WS-17 – Summary 1

The Toronto Raptors defeated the visiting Milwaukee Bucks, 122-100, at Air Canada Centre on Friday. The Raptors (17-7) checked in to **Saturday’s contest** with only two wins in their last 11 games, but the Bucks (11-12) checked in to **Friday’s contest** with a 33-point first quarter, **and the Bucks (11-12) checked in to Wednesday’s contest with a 33-point first quarter. However, the Bucks (11-12) battled back in the second half, outscoring the Raptors by a 33-18 margin over the final 12 minutes. However, the Raptors battled back to make it a one-point game heading into the fourth, but the Raptors battled back to make it a one-point game at the end of the third quarter.**

The Raptors were led by Giannis Antetokounmpo's 30 points, which he supplemented with nine rebounds, five assists, three steals and **three** blocks. Parker was next with 27 points, three rebounds, two assists and one steal. **Parker was next with 27 points, three rebounds, two assists, one steal and one block.** Greg Monroe was next with 11 points, two rebounds, **two** assists and a steal. **Greg Monroe supplied 11 points, two rebounds, two assists and a steal.** Greg Monroe led the second unit with 11 points, two rebounds, **two** assists and a steal. Malcolm Brogdon led the second unit with **23** points, two assists, two rebounds and three steals. Milwaukee was led by a pair of **30-point** efforts from Kyle Lowry and DeMar DeRozan, with the former adding three rebounds and **three** assists, and the latter supplying three rebounds, three assists, **three** steals and a block. Norman Powell was next with a bench-leading **25** points, which he supplemented with **five** assists, **five** rebounds and a steal. Jonas Valanciunas posted a 11-point, 12-rebound double-double that also included an assist, a steal and a block. Cory Joseph led the second unit with **27** points, seven assists, four rebounds and two steals. The Bucks remain in last place in the Eastern Conference's Atlantic Division. They head to Detroit to take on the Pistons on Tuesday.

NCP+CC – Summary 1

The Toronto Raptors defeated the Milwaukee Bucks, 122-100, at Air Canada Centre **on Friday**. The Raptors (**11-12**) **checked in to Monday's contest** having lost four of their last six games, but the Raptors (17-7) were able to pull away in the second half. The Raptors (**17-7**) **checked in to Tuesday's contest** with only **two road wins in 17 tries**, but the Raptors (17-7) were able to prevail with a solid road victory. The Raptors (17-7) were led by Jabari Parker's 27 points, which he supplemented with three rebounds, two assists, a steal and a block. Giannis Antetokounmpo was right behind him with 30 points, nine rebounds, five assists, three steals and a block. **Jabari Parker was next with 27 points, three rebounds, two assists and a steal.** Matthew Dellavedova was next with a bench-leading 27 points, which he supplemented with **three** rebounds, **two** assists and a steal. **Matthew Dellavedova supplied two points, 10 assists and a rebound.** The Raptors remain in last place in the Eastern Conference's Central Division. They head to Cleveland to take on the Cavaliers **on Sunday night**. DeRozan led the way for the Raptors with 30 points, three assists, three rebounds and a steal. Kyle Lowry was next with 18 points, seven assists, three rebounds and a steal. DeMarre Carroll registered 13 points, six rebounds, two assists and a steal. Jonas Valanciunas posted a 11-point, **12-rebound double-double** that also included an assist, a steal and a block. The Bucks head back home to take on the Atlanta Hawks on Monday night, **while the Raptors visit the Cleveland Cavaliers on Friday night as well.**

Heir-k – Summary 1

The Toronto Raptors defeated the Milwaukee Bucks, 122-100, at Air Canada Centre on Monday. The recent form of these teams would indicate an easy Milwaukee win, but Toronto had other ideas. **In fact, the Raptors led for nearly every minute of this game, with the Raptors having their only lead in the first quarter.** Both offenses were extremely efficient, as they each shot over 53 percent from the field. Toronto dominated in the interior, as they won the rebounding battle, 46-37 and outscored Milwaukee, 62-37 in the paint. The assist-to-turnover ratio was a huge factor as well, with the Raptors recording 26 assists to six turnovers, while the Bucks committed 11 turnovers to 27 assists. Toronto (**11-12**) has been regressing over recent weeks and this was a tough

task north of the border. **The bench was the only thing keeping them around**, as **Terrence Ross and Terrence Ross** both scored in double figures off the bench. **In fact, Toronto led for the entirety of the second half**, while holding a double-digit lead for the majority of that stretch. DeMar DeRozan once again led the team in scoring, as he tallied 30 points, three rebounds and three assists. Kyle Lowry was second on the team, finishing with 18 points, three rebounds and seven assists. Jonas Valanciunas provided 11 points and 12 rebounds. **Terrence Ross was a nice spark off the bench, scoring 25 points** on 10-of-17 shooting. Giannis Antetokounmpo led the Bucks in scoring, tallying 30 points, nine rebounds and **three** assists. **Kyle Lowry was second on the team, with 18 points, three rebounds, seven assists** and **two** steals.

Bart_{Tri+Fact} – Summary 1

The Toronto Raptors (17-7) defeated the Milwaukee Bucks (11-12) 122-100 on Wednesday at the Air Canada Centre in Toronto. The Raptors got off to a hot start in this one, out-scoring the Bucks 33-23 in the first quarter and 36-18 in the second quarter. They were able to coast to a comfortable lead for the rest of the game, as the Bucks weren't able to keep up with the Raptors' high-powered offense. The Bucks were led by Giannis Antetokounmpo, who tallied 30 points (13-23 FG, 0-3 3Pt, 4-6 FT), nine rebounds, five assists and three steals in 39 minutes. Jabari Parker followed up with 27 points of his own, going 11-for-17 from the field and 2-of-3 from the three-point line to score 27 points, while also adding three rebounds, two assists and one steal in 35 minutes. The only other player to score in double digits for the Bucks was Malcolm Brogdon, who finished with 4 points (2-6 FG, 2-1 FT) in 23 minutes off the bench. Meanwhile, DeMar DeRozan led the way for the Raptors, scoring a game-high of 30 points on 7-for- 11 shooting from the floor and 15-for 15 from the free throw line. He added three rebounds and three assists in 32 minutes. Kyle Lowry finished with 18 points, seven assists and three rebounds in 34 minutes. DeMarre Carroll and Jonas Valanciunas added **11 points each**, while Terrence Ross chipped in 25 points (10-17 FG) and five rebounds in 21 minutes as a reserve. Jonas Valancunas recorded a double-double of 11 points and 12 rebounds in 23 minutes, while Cory Joseph added seven assists, four rebounds and two steals in 27 minutes. For the Bucks, it was a forgettable night for the starting five, as John Henson, Matthew Dellavedova and Tony Snell combined for just nine points on 1-for 8 shooting in 28 minutes. Up next, the Bucks will head home to take on the Wizards on Friday, while the Raptors will head to New York **to play the Knicks** on Friday. The Raps will look to extend their winning streak to five games as they travel to Brooklyn on Friday **to face the Nets**.

Gold – Summary 2

The Milwaukee Bucks (18-17) defeated the New York Knicks (5-31) 95-82 on Sunday at Madison Square Garden in New York. The Bucks were able to have a great night defensively, giving themselves the scoring advantage in all four quarters. The Bucks showed superior shooting, going 46 percent from the field, while the Knicks went only 41 percent from the floor. The Bucks also out-rebounded the Knicks 48-36, giving them in an even further advantage which helped them secure the 13-point victory on the road. Brandon Knight led the Bucks again in this one. He went 6-for-14 from the field and 1-for-3 from beyond the arc to score 17 points, while also handing out five assists. He's

now averaging 21 points per game over his last three games, as he's consistently been the offensive leader for this team. Zaza Pachulia also had a strong showing, finishing with 16 points (6-12 FG, 4-4 FT) and a team-high of 14 rebounds. It marked his second double-double in a row and fourth on the season, as the inexperienced centers on the Knicks' roster weren't able to limit him. Notching a double-double of his own, Giannis Antetokounmpo recorded 16 points (6-9 FG, 1-1 3Pt, 3-6 FT) and 12 rebounds. The 12 rebounds matched a season-high, while it was his second double-double of the season. Coming off the bench for a big night was Kendall Marshall. He went 6-for-8 from the field and 3-for-3 from the free throw line to score 15 points in 20 minutes. The Knicks really struggled to score without Carmelo Anthony and Amare Stoudemire. Tim Hardaway Jr led the team as the starting shooting guard, going 6-for-13 from the field and 3-for-5 from the three-point line to score 17 points, while also adding four assists. He's now scored 17 or more points in three out of his last four games, as he has put it on himself to pick up the slack with other key players sitting out. J.R. Smith also put together a solid outing as a starter. He finished with 15 points and seven rebounds in 37 minutes. Like Haradaway Jr, he's also benefitted from other guys sitting out, and has now combined for 37 points over his last two games. While he didn't have his best night defensively, Cole Aldrich scored 12 points (6-10 FG) and grabbed seven rebounds in 19 minutes. The only other Knick to reach double figures in points was Jason Smith, who came off the bench for 10 points (3-11 FG, 4-4 FT). The Bucks' next game will be at home against the Phoenix Suns on Tuesday, while the Knicks will travel to Memphis to play the Grizzlies on Monday.

Template – Summary 2

The Milwaukee Bucks (18-17) defeated the New York Knicks (5-31) 95-82. Brandon Knight scored 17 points (6-14 FG, 1-3 3Pt, 4-5 FT) to go with 2 rebounds. Tim Hardaway Jr. scored 17 points (6-13 FG, 3-5 3Pt, 2-4 FT) to go with 3 rebounds. Giannis Antetokounmpo scored 16 points (6-9 FG, 1-1 3Pt, 3-6 FT) to go with 12 rebounds. Zaza Pachulia scored 16 points (6-12 FG, 0-0 3Pt, 4-4 FT) to go with 14 rebounds. Kendall Marshall scored 15 points (6-8 FG, 0-2 3Pt, 3-3 FT) to go with 2 rebounds. JR Smith scored 15 points (6-16 FG, 3-7 3Pt, 0-0 FT) to go with 7 rebounds. The Bucks' next game will be at home against the Dallas Mavericks, while the Knicks will travel to play the Bulls.

WS-17 – Summary 2

The Milwaukee Bucks (18-17) defeated the New York Knicks (5-31) 95-82 on Tuesday at Madison Square Garden in New York. The Bucks got off to a quick start in this one, out-scoring the Knicks 22-22 in the first quarter alone. **The Bucks were able to use a strong first half**, where they out-scored the Knicks 31-18 to seal the victory **in front of their home crowd**. The Bucks were the superior shooters in this game, going 46 percent from the field and 36 percent from the three-point line, while the Knicks went 41 percent from the floor and just 25 percent from deep. The Bucks were also able to force the Knicks into 16 turnovers, while committing **just 16 of their own**. The Bucks were led by the duo of Greg Monroe and Khris Middleton. Knight went 6-for-14 from the field and 1-for-3 from the three-point line to score a team-high of 17 points, while also adding five assists and **two** steals. He's now averaging 20 points and 8 rebounds on the year. Khris Middleton also had a solid showing, finishing with 8 points (2-6 FG, 1-2 3Pt, 3-3 FT) and five rebounds. He's now averaging 16 points and 6 rebounds on the year. The only other Knick to reach double figures in points was **Brandon Knight, who chipped in with**

17 points (6-14 FG, 1-3 3Pt, 4-5 FT) and five assists. The Knicks' next game will be on the road against the Cleveland Cavaliers on Friday, while the Knicks will be at home against the New York Knicks on Friday.

NCP+CC – Summary 2

The Milwaukee Bucks (18-17) defeated the New York Knicks (5-31) 95-82 on Wednesday at Madison Square Garden in New York. The Bucks were the superior shooters in this game, going 46 percent from the field and 25 percent from the three-point line, while the Knicks went just 41 percent from the floor and a meager 36 percent from beyond the arc. The Bucks were the superior shooters in this game, going 46 percent from the field and 25 percent from the three-point line, while the Knicks went just 41 percent from the floor and a meager 36 percent from deep. The Bucks also forced the Knicks into 18 turnovers, while committing just 11 of their own, which may have been the difference in this game, as the Bucks forced the Knicks into 21 turnovers, while committing just 11 of their own. The Bucks' frontcourt did most of the damage in this game. Giannis Antetokounmpo led the team with 16 points (6-9 FG, 1-1 3Pt, 3-6 FT), 12 rebounds and two blocked shots, while Middleton had eight points (2-6 FG, 1-2 3Pt, 3-6 FT), five rebounds and two steals in 22 minutes. It was the first time he's surpassed the 20-point mark this season, so it was good to see him get things turned back around. Coming off the bench, Khris Middleton had eight points (2-6 FG, 1-2 3Pt, 3-3 FT), five rebounds and two steals in 22 minutes. The Bucks' next game will be on the road against the Boston Celtics on Friday, while the Knicks will travel to Brooklyn to play the Nets on Friday.

Heir-k – Summary 2

The Milwaukee Bucks (18-17) defeated the New York Knicks (5-31) 95-82 on Monday. Milwaukee has won four straight games, and the deeper they get into the season, the more believable the Bucks' turnaround under coach Jason Kidd appears to be a sustainable change. The Bucks gave all five of the bench players they used at least 21 minutes. Giannis Antetokounmpo led the way with a game-high 16 points and 12 rebounds, while Giannis Antetokounmpo had a double-double of his own with 16 points and 12 rebounds. It was an off-night for New York, as the team shot just 41 percent from the field and 36 percent from beyond the arc. Jared Dudley (12), Kendall Marshall (10) and Johnny O'Bryant (10) round out the six New York players who scored in double figures. Up next, the Bucks will stay home Wednesday to take on the 76ers, while the Knicks will head to Los Angeles on Saturday to take on the Clippers. As has been the regular season for the Knicks, but they didn't have enough swag to win the Bucks. They will hope to continue their hot start as they take on the Bulls in Madison Square Garden on Monday. The Bucks will also have a few days off before traveling to Orlando to take on the Magic on Wednesday. For the Knicks, meanwhile, the Knicks play their seventh straight loss as they take on the Milwaukee Bucks on Tuesday, in a total of 15 games.

Bart_{Tri+Fact} – Summary 2

The Milwaukee Bucks (18-17) defeated the New York Knicks (5-31) 95-82 on Wednesday at Madison Square Garden in New York. The Bucks got off to a quick start in this one, out-scoring the Knicks 22-21 in the first quarter and never looking back after that. They were led by Giannis Antetokounmpo, who finished with 16 points (6-9 FG, 1-1

3Pt, 3-6 FT) and 12 rebounds in 30 minutes. He also added two assists, two steals and two blocks. Zaza Pachulia added 16 points and 14 rebounds of his own, while Brandon Knight chipped in 17 points and five assists in 32 minutes. It was a nice bounce back game for the Bucks, who had lost four of their last five games coming into Wednesday's contest. The Knicks, on the other hand, have now lost three straight games, as they continue to struggle on offense. They shot just 41 percent from the field and 36 percent from beyond the arc. Tim Hardaway Jr. was the high-point man for the Knicks, finishing with 17 points on 6-of-13 shooting, while J.R. Smith added 15 points, seven rebounds and four assists in 37 minutes. The only other Knick to reach double figures in points was **Willy Hernangomez**, who scored 10 points (3-11 FG, 4-4 FT) in 24 minutes off the bench. Up next, the Bucks will head home Friday **to take on the road** to play the Bulls, while the Knicks will travel to Boston on Friday to play against the Celtics.