ISSN 2499-4553

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 Accademia university press

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

Rivista Semestrale dell'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

IJCoL

Volume 7, Number 1-2
june-december 2021
Special Issue

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
Francesco Cutugno, Hendrik Buschmeier 7

Knowledge Modelling for Establishment of Common Ground in Dialogue Systems
Lina Varonina, Stefan Kopp 9

Pragmatic approach to construct a multimodal corpus: an Italian pilot corpus
Luca Lo Re 33

How are gestures used by politicians? A multimodal co-gesture analysis
Daniela Trotta, Raffaele Guarasci 45

Toward Data-Driven Collaborative Dialogue Systems: The JILDA Dataset
Irene Sucameli, Alessandro Lenci, Bernardo Magnini, Manuela Speranza e Maria Simi 67

Analysis of Empathic Dialogue in Actual Doctor-Patient Calls and Implications for Design of Embodied Conversational Agents
Sana Salman, Deborah Richards 91

The Role of Moral Values in the Twitter Debate: a Corpus of Conversations
Marco Stranisci, Michele De Leonardis, Cristina Bosco, Viviana Patti 113

Computational Grounding: An Overview of Common Ground Applications in Conversational Agents
Maria Di Maro 133

Cutting melted butter? Common Ground inconsistencies management in dialogue systems using graph databases
Maria Di Maro, Antonio Origlia, Francesco Cutugno 157

Towards a linguistically grounded dialog model for chatbot design
Anna Dell'Acqua, Fabio Tamburini 191

Improving transfer-learning for Data-to-Text Generation via Preserving High-Frequency Phrases and Fact-Checking
Ethan Joseph, Mei Si, Julian Liaonag 223

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 Bressem 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). Bressem and Müller spotted a list of recurrent gestures in German that carry out pragmatics function and illocutionary values (Bressem 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). Bressem 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 (Bressem, 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

Symbol	*Value*
//	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

SPEECH UNITS		*GESTURAL UNITS*
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,

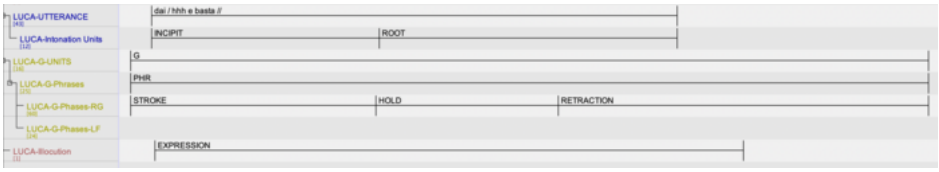

Figure 3
Stretch of transcription

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, out 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

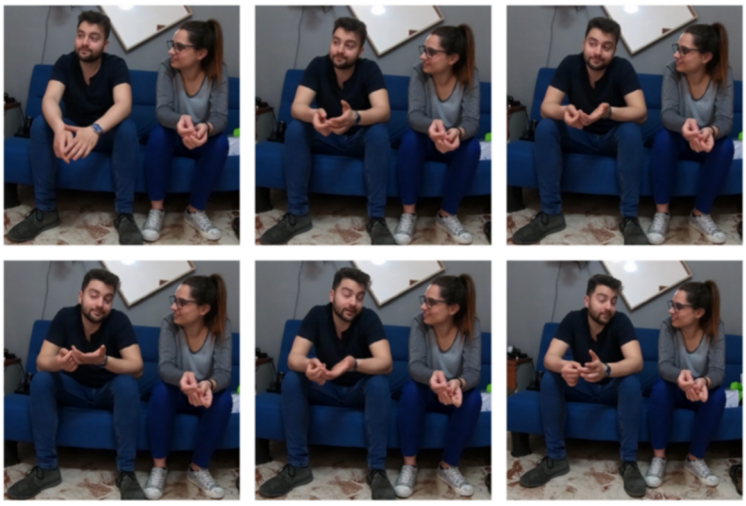

Figure 4
Example of gesture

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 occuring 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

INTERACTION GENRE	*INTERACTION TYPE*	*PLACE*
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 filed 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'illocuzione 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 (l-act). *Spoken corpora and linguistic studies. Amsterdam: John Benjamins*, pages 468–495.
Müller, Cornelia, Silva H. Ladewig, and Jana Bressem. 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 Tessendorf, 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.