Journal of Speech Sciences (9): 07-30. 2020.
Available at: http://revistas.iel.unicamp.br/joss

THE ANNOTATION OF GESTURE AND GESTURE / PROSODY SYNCHRONIZATION IN MULTIMODAL SPEECH CORPORA

CANTALINI, Giorgina[1]
MONEGLIA, Massimo[2]*

[1]Civica Scuola di Teatro “Paolo Grassi”, Milan
[2]University of Florence

Abstract: *This paper was written with the aim of highlighting the functional and structural correlations between gesticulation and prosody, focusing on gesture / prosody synchronization in spontaneous spoken Italian. The gesture annotation follows the LASG model (Bressem et al., 2013), while the prosodic annotation relies on the identification of terminal and non-terminal prosodic breaks which, according to L-AcT (Cresti, 2000; Moneglia and Raso, 2014), determine speech act boundaries and the information structure, respectively. Gesticulation co-occurs with speech in about 90% of the speech flow examined and gestural arcs are synchronous with prosodic boundaries. Gesture Phrases, which contain the expressive phase (Stroke) never cross terminal prosodic boundaries, finding in the Utterance the maximum unit for gesture / speech correlation. Strokes may correlate with all information unit types, however is infrequent with Dialogic Units (i.e. those functional to the management of the communication). The identification of linguistic units via the marking of prosodic boundaries allows us to understand the linguistic scope of the gesture, supporting its interpretation. Gestures may be linked to information belonging to different linguistic levels, namely: a) the word level; b) the information unit phrase; c) the information unit function; d) the illocutionary value.*

Keywords: Multi-Modal Corpora; Co-speech Gestures; Prosody; Synchronization; Pragmatics; Information Structure.

*Corresponding author: moneglia@unifi.it

1 Introduction

The connection between gesture and prosody is considered fundamental to the study of gesture (Kendon, 1972; 1980; McClave, 1991) since "their tight connection is a facet of the strong underlying linkage between gesture and speech in general, which in turn is felt to exist because speech is a fundamentally embodied phenomenon" (Loehr, 2014: p. 1388). This paper proposes an annotation model for multimodal speech corpora aimed at highlighting the functional and structural correlations between the level of co-verbal gestures and prosody. The model was created for specific reasons (Cantalini, 2018): for the comparison of gestures in acting and in spontaneous speech (Nencioni, 1976). The present article is aimed in particular at illustrating the annotation schema for gesture and prosody and focuses on the results obtained from the synchronization of gesture / prosody in spontaneous speech.

The annotated dataset consists of two collections of video recordings - spontaneous speech and recited speech - from which comparable samples were taken and annotated:

- Spontaneous: three samplings of at least two minutes each of continuous monological speech derived from three structured interviews with theatre actors on the work of the actor.
- Recited: the same monological joke, taken from the comedy "Il giuoco delle parti" (The Rules of the Game), recited by the actors interviewed above, plus a quarter recovered from the RAI archives.[1]

In section 2 we will briefly outline the theoretical reference models used for the analysis of gesture and prosody and implemented in the annotation schema, which is described in section 3. In section 4 we will briefly report the results of the validation work, illustrating how the model for gesture annotation was implemented following agreement between annotators on how to apply the specifications. In section 5 we will illustrate the level of synchronicity observed between the gestural and prosodic elements in the spontaneous speech examined. Section 6 discusses the possible advantages brought by gesture prosody synchronization for the interpretation of co-speech gestures, focusing on the different linguistic levels identified by prosody that may constitute an anchor for gesture interpretation.

Table 1: The Dataset [2]

Spontaneous	*Duration in minutes*	*N° Words*	*Recited*	*Duration in minutes*	*N° Words*
Actor- D	04:48.896	685	Actor- D	02:33.100	236
Actor -M	01:58.725	314	Actor -M	01:33.888	239
Actor -O	03:18.181	565	Actor -O	01:27.020	168
			Actor -V	02:15.337	251
Total	**10: 05.802**	**1564**	Total	**07: 49.345**	

2 Theoretical frameworks

[1] L. Pirandello, *Il giuoco delle parti* (Act 1, Scene 3rd – first edition 1918). Milano: Mondadori. The jokes in question are known in the theatrical world as "Il pezzo del guscio d'uovo" [the piece of the eggshell] in the monologue of the character Leone Gala.

[2] The words to be recited are the same in each instance, save for variations owing to interpretation. This is significant in the case of O who omits certain parts.

Gestures have a linear structure that can be segmented into hierarchically ordered units aligned with the flow of speech (Kendon, 1972; 1980; McNeill, 1992; 2005). For the analysis of gesture flow, the models proposed in Kita *et al.* (1998) were used, along with additions from Ladewig, Bressem (2013) and Bressem *et al.* (2013). The model foresees a hierarchy of gestural elements that develop around a mandatory nucleus in which a peak of energy occurs (the *Expressive phase*), constituting the semantic part of the gesture. The Expressive phase may be simple or compound, dynamic or stationary.

The typical expressive part (the *Stroke*) is recognizable from a dynamic point of view as the apex of the force and the vertex of the movement. The *Stroke* is nuclear and is sufficient to constitute a gesture or *Gestural phrase* (GPhrase), and may be preceded and / or followed by, respectively, a preparation phase (*Preparation*) and a phase of reabsorption of energy (*Retraction*). The expressive nuclear part may also occur when the hand takes a specific form while remaining firm (the *Hold*). Each of the elements that make up a GPhrase identifies a *Gestural phase* (GPhase), which is at the lower level of the hierarchy.

The framework considers that the gesture movement develops by starting with a surface from which the hand detaches (an armrest, a tabletop, a part of the speaker's body: the arm, the lap, etc) and finishes on a surface on which the hand rests. The entire 'sequence of movements' between the start position and the end position (*Home position* in Kendon, 2004: p. 111) constitutes the hierarchically higher level of gestures, referred to as *Gesture units* (GU). In the GUs, more expressive phases can be inserted, and thus more gestural phrases, which are included hierarchically.

The gestural structure may therefore consist of a single *Stroke* (coinciding with a GPhrase and GU) as well as a much more extensive segment in which a GU comes to contain several GPhrases, in their turn composed of multiple GPhases around an expressive nucleus. Figure 1 schematizes the hierarchical model.

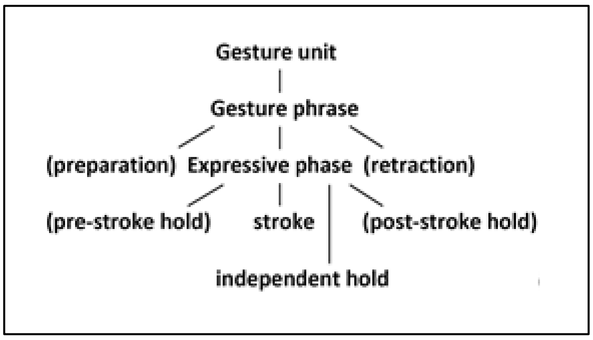


Figure 1: The hierarchical structure of the gesture (from Kita *et al.* 1998: p. 27, with adaptations).

On this configurational model we have overlaid a theoretical framework for the identification of prosodic units which is also configurational and based on perception: The Language into Act Theory (L-AcT) (Cresti, 2000; Moneglia and Raso, 2014). L-AcT focuses on the identification of the minimal units that can be pragmatically interpreted in the flow of speech (Quirk *et al.,* 1985), corresponding to *Speech acts* (Austin, 1962). These units are considered the *Units of Reference* for linguistic analysis (RU) (I'zreel *et al.,* 2020). Crucially, RUs are identified on a prosodic basis, and are considered necessarily separated from each other in the speech flow by perceptually relevant *Prosodic boundaries with a terminal value* (TB) (Karcevsky, 1931; Crystal, 1975).

The syntactic constituents of the RUs are realized as *Information Units* (IU) which also correspond to *Prosodic Units* (PU), following a principle introduced by Halliday (1967) and by Chafe (1994), according to which a PU corresponds to a unit of *flow of thought*. The Prosodic units within the RUs are, in turn, separated by perceptually significant *Prosodic boundaries with a non-terminal value* (NTB). From a perceptive point of view, a PU is a sequence of syllables bearing a recognizable prosodic movement and the PUs of a RU are organized into a *Prosodic Pattern* ('t Hart *et al.,* 1990) which is concluded, perceptually speaking.

In summary, the RUs are the hierarchically superior entities and are composed of IUs. Both types of elements correspond to *Prosodic units* (PU) and are marked by *Prosodic breaks*.

In the L-AcT methodology, as well as for the annotation of various large oral corpora (Du Bois *et al*., 1993; Spoken Dutch Corpus; Cheng *et al.,* 2005; Amir *et al.,* 2004; Cresti and Moneglia, 2005; Raso and Mello, 2012; Izre'el and Mettouchi, 2015) the key operating element is the identification of prosodic breaks: Terminal Breaks (TB), indicated by a double slash"//" in transcriptions, and Non Terminal Breaks (NTB), indicated by a single slash "/". The TBs mark the boundaries of the RUs, which is to say the minimal pragmatically interpretable linguistic entities, while the NTBs mark the boundaries of the information units (IU) that make up the RUs, allowing the identification in the *flow of speech* of the two hierarchical levels with functional values into which speech is structured, according to this theory.

At the upper level, pragmatic functionality expresses itself in terms of three types of RU:[3]

1) The *Utterance*, which corresponds to a linguistic act, and which can be *Simple* (consisting only of one IU) or *Complex*, composed of several IUs in a prosodic pattern;

2) The *Illocutionary pattern*, which structures several linguistic acts (usually two), performed within a prosodic pattern, functioning towards the expression of rhetorical relations (*reinforcement, list, comparison,* etc.);

3) The *Stanza*, which is made up of a sequence of weak illocutionary acts that follow the flow of thought. The stanzas develop through an additional process and do not correspond to the execution of a prosodic pattern.

At the lower level, the IUs convey a defined set of information functions belonging to two types: the *Textual*, which implements the semantic information of the Utterance, and *Dialogic,* which does not implement the semantics of the Utterance but is dedicated to the management of the interaction.

Similar to gestures, the L-AcT model postulates a mandatory textual element: the *Comment* (COM) IU, which is the information unit *necessary and sufficient* for the formation of a linguistic entity that can be interpreted pragmatically, given that its function is to specify how linguistic elements are to be interpreted in the world. In summary, the COM expresses the Illocutionary force (Austin, 1962) and corresponds to a PU of type *root* in the terminology of Hart *et al.* (1990).

Beyond the COM, there also exist the textual IUs of the *Topic* (TOP), *Parenthesis* (PAR), *Appendix of Topic* or *Comment* (APT / APC) and Locutive Introducer (INT). Among the Dialogical units, which may roughly be referred to as discursive signals, various functions are recognized including the *Discourse connector* (DCT), however their specification is beyond the scope of this paper. The theory ultimately predicts a *Scansion Unit* (SCA), which has no independent information function but scans one IU into different prosodic groups (see Moneglia and Raso, 2014, for definitions of the *tagset* for the IUs).

[3] See Moneglia and Raso (2014) and Saccone *et al.* (2018) for details on the reference unit types in L-AcT. For brevity and terminological transparency, we will refer to all types of RUs with the generic term *Utterance.*

Using the L-AcT framework, we distanced ourselves from the more traditional approach used for the study of the relationship between gesture and intonation, based on the autosegmental ToBI model (Beckman *et al.*, 2005), which observes in particular the alignment between the apices of the gestural movement and the *Pitch accent* (Loehr, 2004; 2014). L-AcT, constituting a model for the prosodic analysis of speech comparable to that outlined for gesture analysis, lends itself extensively to observing the relationship of the gesture with the prosodic units. Furthermore, the annotation methodology is, in both models, tested on the analysis of spontaneous speech corpora and is based on the perceptual recognition of hierarchically ordered units, from which the categorization process can be derived. In particular, by identifying the functional values of the prosodic units, it allows one to clarify the linguistic levels to which the co-verbal gestures refer, beyond the scope of words and syllables. In 6. we will test this concept considering actual cases of gesture interpretation.

3 The Annotation Schema

3.1 Gesture and Prosody

The gesture annotation system implemented is based on the *Linguistic Annotation System for Gestures* (LASG) by Bressem *et al.* (2013), in which hand movements are taken as the principal element in the annotation process. The software used is ELAN, which allows the annotation and transcription of the content of the audio-video recordings by multiple annotators, with the ability to distribute the annotations into several levels (*tiers*), connecting them hierarchically if necessary.

Before being reconciled in ELAN, the prosodic and gestural levels were annotated separately by different annotators using specific processes. For prosody, the procedure involved restricting access to just the acoustic information and omitting the images (and thus the ability to see the gestures). The transcription and annotation of the linguistic-prosodic units was carried out using WinPitch, which allowed for the alignment of the transcription and its audio wavelength, while simultaneously allowing access to all parameters of acoustic, segmental and suprasegmental analysis (F0, Intensity, Duration, spectrograms etc.), as well as multilevel annotation.[4]

In the practice of L-AcT, the Terminal and Non-terminal prosodic breaks (the principal objects of the annotation) are identified on a perceptual basis during the process of transcription, and then validated through instrumental verification of the acoustic correlates that breaks record in speech, which correspond roughly to the following list (Cruttenden, 1997; Amir *et al*., 2004; Sorianello, 2006):

a) F0 reset
b) final lengthening
c) drop in intensity
d) break
e) speed increase

The annotation of the functional value (according to the L-AcT tagset) is performed after the wave alignments for the Prosodic units, by consensus agreement between two annotators. WinPitch allows us to verify the perceptive judgments considering F0, intensity and duration in real time. In its second tier (DNTB) Figure 2 shows the scanning (SCA) into three Prosodic

[4] Winpitch is our preferred system, however the same procedure may be followed using Praat.

units of a single Parenthetical information unit (PAR), preceded by a Discourse Connector (DCT).

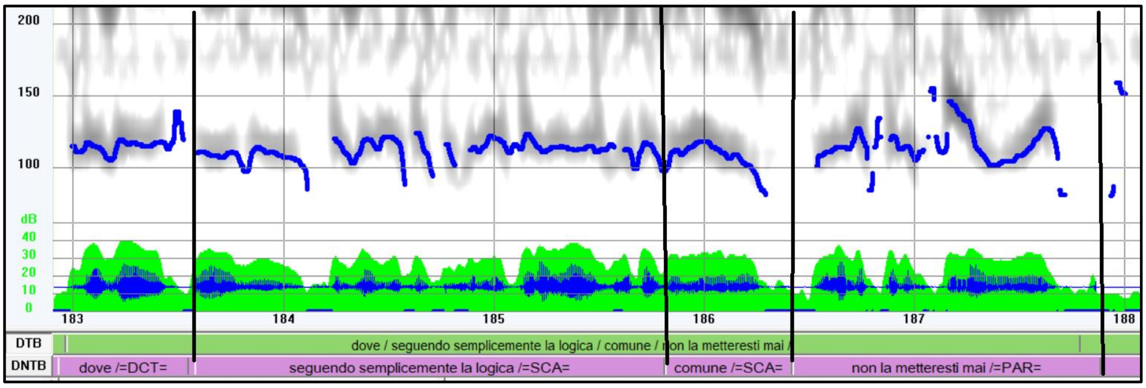

200
150
100
dB
40
30
20
10
0
183
184
185
186
187
188
DTB
dove / seguendo semplicemente la logica / comune / non la metteresti mai /
DNTB
dove /=DCT=
seguendo semplicemente la logica /=SCA=
comune /=SCA=
non la metteresti mai /=PAR=

Figure 2: Transcription, alignment and annotation of the prosodic units in WinPitch.

For gestures, annotations are made directly in ELAN and draw information from both the acoustic and visual modalities (in accordance with McNeill, 2005), although without access to transcriptions and prosodic annotations.[5] In our opinion, if the goal is to annotate the *co-speech gesture* then the removal of the information relating to speech, with respect to which the gesture finds relevance, does not seem justified as it eliminates perceptually relevant information for its identification.

The annotation model defines an ELAN tier for each hierarchical level of the theoretical frameworks and, thus, two distinct annotation sections: SP (*Speech*) and GE (*Gesture*). The SP section is composed of two tiers marked SP-1, which identifies the RUs, or the upper hierarchical level for the pragmatic analysis of speech and SP-2, which identifies the IUs, or the information level into which the RUs are organized.

The first tier in each of these sections is temporally subdivided into segments corresponding to the lengths of the units. The other tiers are hierarchically dependent and will necessarily be the same or included in this one. Thus, we have:

- *SP-1-TB transcript*: segmentation of the sound continuum into Reference Units with pragmatic values corresponding with the presence of Terminal Prosodic Breaks (TB); transcription of the text of the sequence between the two TBs;
- *SP-1-RU type:* specifies the typology for each Reference Unit between two TBs: Utterance; Stanza, Illocutionary-pattern;
- *SP-1-TB label*: gives the ID of each RU.

The first tier therefore has the function of segmentation and the others of classification and labeling. Analogously, the second section consists of the tiers:

- *SP-2-NTB transcript:* segmentation of the RUs identified above into IUs, in correspondence with the non-terminal Prosodic Breaks (NTB) present within; transcription of the text;

[5] This choice diverges from the German tradition of gestural studies that belongs to Cornelia Müller, in which one chooses to annotate the gesture without listening to the audio (see Bressem, 2013; and Müller, personal communication). In our view the adequacy of the annotation of the *co-speech gesture* without access to the audio must always be verified with respect to the acoustic source, in particular to avoid excessive granularity in the annotation.

- *SP-2-IU type*: indicates for each Information Unit the tag corresponding to its information function, according to the L-AcT types (e.g. COM, TOP, PAR etc.);[6]
- *SP-2-NTB label*: progressive numbering of the units.

You can choose to make the SP-2 section hierarchically dependent on SP-1, however, from the point of view of the annotation technique it is advantageous to leave the two sections independent during the annotation phase, so as to be able to make all necessary changes and link them hierarchically only after the analysis is completed. In ELAN it is in fact impossible to modify hierarchically higher tiers without also canceling the annotation of the lower ones included within them. This approach is therefore highly recommended, in particular with regard to the sections dealing with gesture analysis, which are often subject to refinements.

The gestural part, GE (*Gesture*), is structured into three levels, for each of which the start time and the end time of the unit considered are identified:

- *GE-1-GU*: in which the *Gesture Units* are identified;
- *GE-2-GPHR*: in which the *Gesture Phrases* are identified;
- *GE-3-GPHA*: where the *Gesture Phases* of the *Gesture Phrases* are identified.

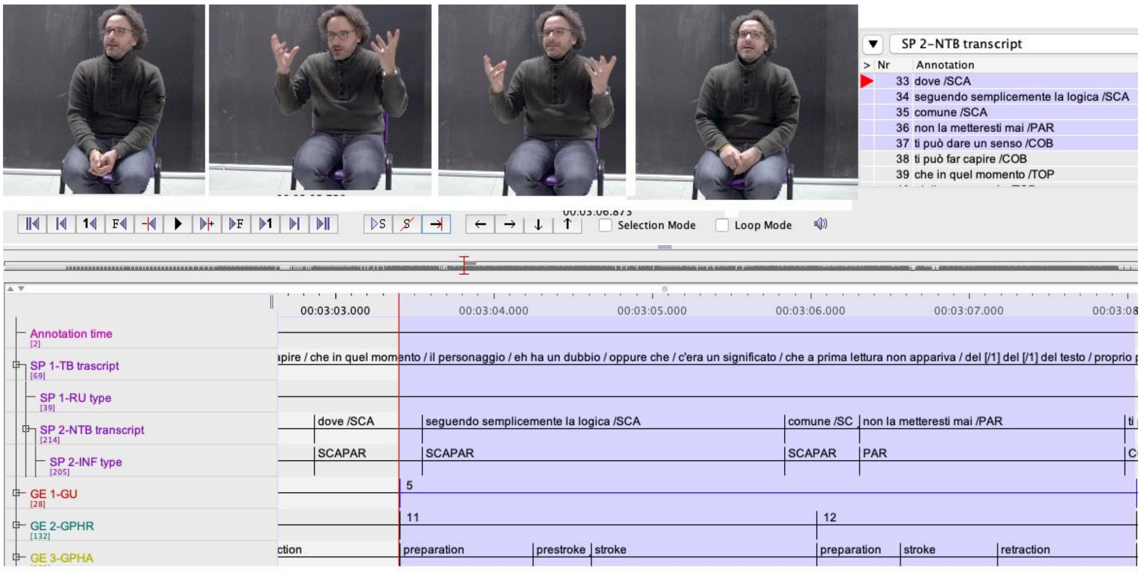
In this annotation structure, aimed at exploring the levels of synchronization between gestures and prosodic units, the gestures were not described, and the tier of each segmented unit would have remained empty. It was therefore decided to insert labels and classifications directly into each tier, which therefore takes on both the functions of segmentation and classification. Specifically we will have: a numeric label with no leading zeros for GE-1-GU; a numeric label with no leading zeros for GE-2-GPHR; and a tag for each unit within GE-3-GPHA that is chosen from a closed set of possibilities, namely: *Preparation*, *Stroke*, *Retraction*, *Hold*, and furthermore, as we will see, *Rest position.*

In the annotation of this dataset, which is monological, each section necessarily refers to a single speaker. In the case that we examine dialogues, and therefore more speakers, each section will be duplicated for each speaker and distinguished by its identifier (e.g. SP-1-TB-P1; SP-1-TB-P2; etc.).

Once the annotation of the gesture and the prosody are performed, the two annotations are reconciled in ELAN. In particular, the transcription and alignment of the prosodic units created with WinPitch are exported to PRAAT TextGrid files and then from PRAAT exported to ELAN, which has a specific function for linking with this environment, resulting in the prosodic annotations being temporally aligned with the gestural ones.

Figure 3 shows the standard annotation of a GU composed of two GPhrases and performed within the scanned parenthesis information unit shown in Figure 2. In the screenshots the participant highlights two *Strokes*, preceded and followed by the *Rest positions* at the start and end of the GU. The essential part of the annotation schema - below on the left side - outlines the annotation levels of the prosody and gesture, while center onward the vertical lines on each tier indicate the beginning and end of each unit (prosodic and gestural), whose synchronicity can thus be evaluated.

[6] If an extraction from ELAN is required of the CHAT format in the version that implements the prosodic breaks (Moneglia and Cresti, 1997) for the analysis of the text, it is worthwhile inserting some redundancy and transcribing the text both at the SP-1-TB transcript level and the SP-2-NTB level.



SP 2-NTB transcript

Nr	Annotation
33	dove /SCA
34	seguendo semplicemente la logica /SCA
35	comune /SCA
36	non la metteresti mai /PAR
37	ti può dare un senso /COB
38	ti può far capire /COB
39	che in quel momento /TOP

Annotation time [2]

SP 1-TB trascript [69]: apire / che in quel momento / il personaggio / eh ha un dubbio / oppure che / c'era un significato / che a prima lettura non appariva / del [/1] del [/1] del testo / proprio

SP 1-RU type [39]

SP 2-NTB transcript [214]: dove /SCA | seguendo semplicemente la logica /SCA | comune /SC | non la metteresti mai /PAR | ti

SP 2-INF type [205]: SCAPAR | SCAPAR | SCAPAR | PAR | C

GE 1-GU [28]: 5

GE 2-GPHR [132]: 11 | 12

GE 3-GPHA: ction | preparation | prestroke | stroke | preparation | stroke | retraction

Figure 3: Annotation tiers of a GU (id. 5) and screenshots from *Rest position* to *Rest position* and the *Strokes* from the GPhrase (id. 11 and id. 12).

3.2 Annotation in the presence of flow interruptions and other special cases

The described system fits well with the annotation of continuous speech flow in the presence of continuous co-verbal gestures. However, we must consider that both flows may be interrupted synchronously or asynchronously, producing interesting phenomena. The primary one of these is the interruption of gestures in the presence of continuous verbal flow (speech without co-verbal gestures), whose identification criteria - which are non-trivial - are described in the next paragraph.

The *Gesture interruption* is annotated in this schema in an independent manner, in a tier (GE-INTERRUPTION) that identifies blocks of speech flow in which the gesture ceases in a significant or noticeable manner. If greater than or equal to one second in length, the halting of a gesture - which from the point of view of gesture analysis would form a *Rest position* - is considered an *interruption of the gesture flow*. In this way, the actual co-verbal gesticulation time and the time of speech not accompanied by gesture are highlighted, which constitute the highest-level relationship between speech and gesture (see the results presented in 5).

However, we encountered not just speech without gesture, but also gesture without speech. For the study of this phenomenon, in which gestures can no longer be strictly considered as co-verbal, some ad hoc solutions were prepared, which marginally modify the L-AcT annotation practice. In particular, the pause in speech, indicated with '#' in CHAT - is not aligned as an independent unit in the annotated oral corpora according to L-AcT (Cresti and Moneglia, 2005), but is incorporated into the verbal unit. In order to consider the case of non-co-verbal gesticulation, the decision was made to modify the model, isolating the pauses greater than or equal to one second in length, which can occur after both terminated and non-terminated break units. This allows us to verify the behavior of the two modes.

For the occurrences of non-co-verbal gesticulation two special tiers were created: GE-1- # and GE-2- #; in this way the gestural units corresponding to silences may simply be subtracted from the general computation of those synchronous to linguistic units and measured temporally.

We have also found, particularly in recited speech, occurrences of gestures which correspond with non-linguistic communication units, such as screams, instances of loud coughing, and onomatopoeias. Also, in this case, to allow the analysis of *non-co-speech* gestures, we distanced ourselves from CHAT practice, which reports these events in transcriptions as ‘hhh’, without isolating them as an alignment unit. These units have been highlighted in the transcription tiers as independent units, indicated with breaks of a specific type ['@'] and then classified in a specific SP-COMM RU (*Communicative Reference Unit*) tier.

Clearly, in the case of pauses for speech, interruptions of gesture, and communication units, we are faced with cases that, according to our observations, are infrequent in ordinary speech, but nonetheless interesting, for which the annotation methodology deserves specific attention.

4 Validation of gesture annotation and the implementation of the annotation schema

The objective that the annotations resulting from the subjective judgments of researchers be "consistent" and reproducible to the level of current practice in the literature is ensured through so-called indices of *Inter-Annotator Agreement* (Carletta, 1996; Gagliardi, 2019).

As for the evaluation of the gesture annotation, the segmentation of the gesture flow and the classification of the gestures in the dataset were replicated by two researchers (independently of one another) on about 1:58 min of spontaneous speech and about 1:45 of acted speech, corresponding to roughly 18% of the data set. This validation was presented extensively in Cantalini *et al.* (2020) and its general results are reported here since the validation work gave rise to an implementation of annotation schema just presented.

The validation concerns two distinct types of tasks involved in the annotation of the different units of analysis (GUnit, GPhrase, GPhase). The GU and GPhase annotation is a task of identifying the unit’s boundaries in a continuum, or of *unitizing* (Krippendorff, 1980); the annotation of the GPhase, on the other hand, is a hybrid task, insofar as the identification of the boundaries of each phase of the gesture follows the attribution of a quality (*tag*) from a closed set of possibilities (*Rest position, Stroke, Hold, Preparation, Retraction, Pre-Stroke Hold, Post-Stroke Hold, Partial Retraction*). The measures for evaluating agreement for the two types of tasks are different:

- GUnit, GPhrase: "overlap/extent value", calculated by dividing the temporal extent of the overlap between the annotations by their total length;
- GPhase: “modified Cohen's Kappa” (Holle and Rein, 2013) which takes into account both the value attributed to the unit and the temporal overlap between the annotations.

The results obtained are comparable to the domain *benchmarks* reported in Lausberg (2013) and Helmich and Lausberg (2014). In particular, the segmentation of gestural units and phrases reached satisfactory levels of agreement (GUnit Av, overlap / extent ratio: 0.8358; GPhrase Av, overlap / extent ratio: 0.6106), as did the attribution of value to the phases of the gesture in all cases where segmentation is compatible (*raw agreement* K: 0.8804). The agreement is instead markedly lower when considering the phases for which there was no agreement on the segmentation (*raw agreement* 0.4402).

Focusing on the qualitative divergences encountered in the annotation of gestures is significant for establishing a reliable annotation schema. In this case the validation made it possible to reach consensus on three sensitive aspects of the gesture flow segmentation procedure, with regard to the GU segmentation criteria, the interpretation of the *Hold*, and the criteria for distinguishing the *Stroke* from the other phases.

The first of these is certainly the most relevant for the purpose of comparing the flow of speech (which is substantially continuous) and the flow of co-verbal gestures (not necessarily continuous). The procedure indicated by Kita, *et. al*. (1998) for the segmentation of the GU may produce different annotation practices where it states: "The end of a movement unit is the moment at which the hand makes the first contact with the resting surface" (Kita *et. al*., 1998: p. 29). Based on this, in one the two annotation practices tested the relaxation phase after contact with the *Home position* was considered a GPhase of the type *Rest position*, and as such part of the GUnit regardless of its length; in the other, however, as soon as the hand touched a surface, the unit was considered terminated. Consequently, in the first encoding the sequence of GUs accompanying the speech showed a *continuous gestural flow*, while in the second the flow was configured as a *discrete* series of co-verbal gesture segments, with many short or long temporal interruptions.

Consider that the GUnit can continue even after contact with the *Home position*, like the final part of a *Retraction* phase, or as an adjustment of the hand in a *Rest position*[7], impedes the observation of relevant interruptions of the gesture flow. Therefore, only if the interruption is short should the *Rest position* constitute a phase of movement transition (ending of the previous or beginning of the next). Considering the limits of the average gesture duration (McNeill, 2016: p. 54), it was therefore decided to standardize the specification and consider *Rest positions* less than one second in length as movement transitions and those of a longer duration as actual interruptions in the GU flow.

With regard to the final relaxation of the hand, it was decided to consider such a phase after a contact with the *Home position* as a *Retraction*, and a placement of the hand less than or equal to 1s as a phase (the phase of *Rest position*), both to be included as options within the movement of the GPhrase following it (Duncan, 2008). On the basis of this standardization, it was therefore possible to highlight with precision to what degree co-verbal gestures insist on the oral component (see Table 2 below).

A similar problem occurs in the identification of the GPhase of *Hold*, in which the speaker's hands are envisioned as being in standby, both when the position is held *in the lap*, and when the position is in *mid-air.* The problem facing the annotator is in deciding whether to consider the standby position as actually a *Hold* (and thus an expressive phase), or on the contrary a *Rest position*, and therefore possibly an interruption in the flow (when equal to or greater than one second in length). In this choice any movement of the hands (which may indeed occur) is significant; in the presence of hand movements we tend to consider the position an expressive phase. Further, in the case of the *mid-air Hold* - at least in spontaneous speech where the position is intentionally held - the interpretation as a *Rest position* is discarded.

Finally, a significant implementation for the specification of the identification of gestures may occur in critical cases in which a movement is segmented into several apices, for which uncertainty may arise in the attribution of the expressive values. Based on Kita *et al*. (1998: p. 30) the specification was enriched through the addition of the "*Multi-segment"* GPhase. In parallel, the movement apices are considered *Strokes* in their own right only if they are very marked, which is to say when changes occur in both the speed and the direction of movement.

No specific validation was performed with regard the prosodic annotation. The prosodic and informational annotation of the dataset was achieved through the agreement of two expert annotators in the LABLITA laboratory of the University of Florence (*Consensus agreement*) and the reliability of the overall methodology has been the subject of many works in the last two decades and is beyond the scope of this paper. In particular, however, consensus for the perceptual identification of prosodic breaks in oral corpora has been evaluated for the Dutch Corpus (Buhmann *et al.,* 2002) and for C-ORAL-ROM and C-ORAL-BRASIL (Danieli *et al.,* 2004; Moneglia *et al.,* 2010). Overall, they show a high level of agreement regarding Terminal

[7] Kita (2018) - personal communication

breaks, both for expert and non-expert annotators, and an increase in the level of agreement on non-terminals for expert annotator groups (Raso *et al.*, 2020). Strong perceptual evidence for prosodic breaks, and especially for terminal ones, has also recently been demonstrated by comparing different approaches to the prosodic segmentation of speech. Panunzi *et al.* (2020) showed in particular that agreement on the detection of Terminal breaks in speech flow is strong, independently of the theoretical model being applied for their identification.

5 Synchronization: analysis of the results

The first rough piece of data that we obtain in considering the flow of spontaneous speech in correlation with gesture flow is the remarkable degree of co-occurrence of the two (see Table 2). In spontaneous speech, the gesture accompanies the speech almost continuously (about 90% of the time) and the occurrence of speech not accompanied by gesture (i.e. interruption of the gesture flow) is extremely low. In other words, the *co-speech gesture* constitutes a pervasive characteristic of spontaneous speech performance. Incidentally this is not true for recited speech, where the ratio between speech flow and gesture flow is about 1/2, that is to say, about half of the time taken up by the jokes is not accompanied by any gesture at all.

Beyond this point - which identifies a rough but important characteristic of spontaneous spoken Italian - in order to ascertain the manner and degree of synchronization between the levels of prosodic and gestural organization the Terminal and the Non-terminal breaks of the prosodic units have been considered in relation to the beginning and end of the gestural units.

Table 2: Speech without gesture: interruptions in the gestural flow

	Total Time	Gesticulation	Interruption of Gesticulation
Spontaneous	10:05.802	08:52.015	01:13.787
Acted	07:49.345	05:27.209	02:22.136

We have examined both the GUnit type and the GPhrase type, verifying their co-occurrence in the case of deviations of up to 200 ms between the unit boundaries (misalignments), and in the case in which the end of the gesture occurs within a prosodic unit with deviations of more than 200 ms between the unit boundaries. For example, the GU in Figure 3 starts about 100ms before an NTB and ends about 50ms after an NTB and the boundaries have been considered as aligned.

Following the above criteria, we examined in particular whether GUnits and GPhrases tended to align with prosodic units or to cross their boundaries (especially their terminal boundaries), as well as the exact nature of the relationships between the expressive phases and the typology of the information units.

As can be seen from the histograms in Figures 4 and 5, in the spontaneous speech examined the GUs are to a significant degree aligned with the prosodic breaks; the gestural *onset* coincides substantially with a prosodic reset. That is, when a new range of gestural movements begins, in about 92% of cases the beginning correlates with the presence of a prosodic break, and therefore in the L-AcT reference framework with the beginning of a new information unit as marked by prosody.

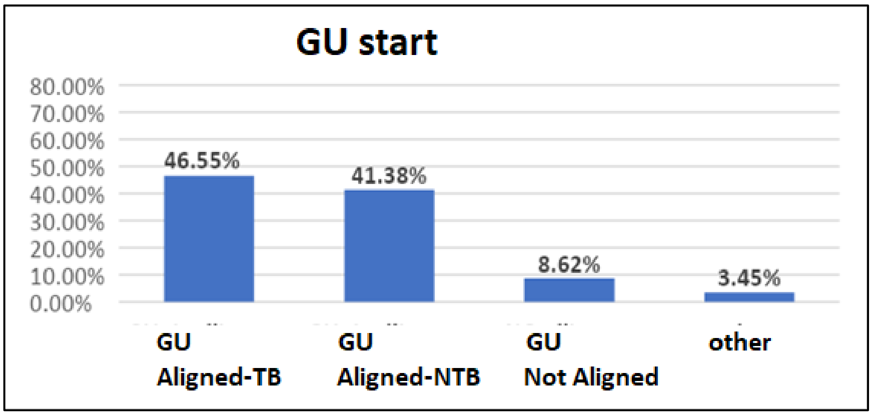


Figure 4: Alignment of the beginning of the Gesture Units with prosodic breaks.

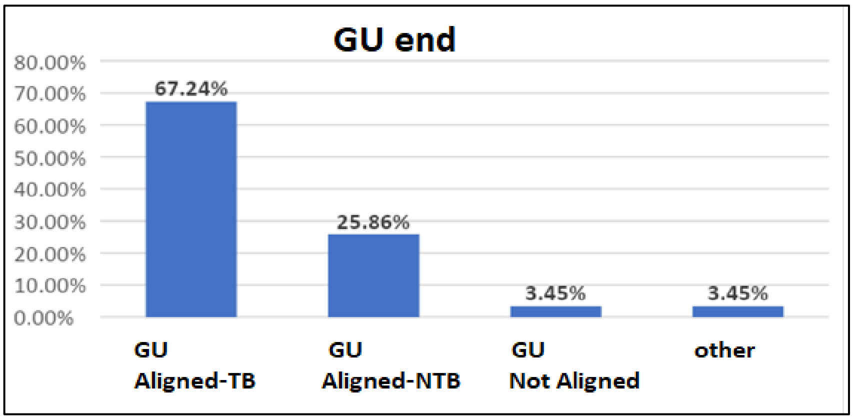


Figure 5: Alignment of the end of Gesture Units with prosodic breaks.

The slight preference for terminal *onsets*, i.e. when a new Utterance begins, may also be significant if we consider that Terminal breaks are much more infrequent than Non-terminal ones. Considering the histogram in Figure 6, the difference between GUs and GPhrases with respect to an *onset* that coincides with a TB is relevant: the first phrase of a GU is more likely to be correlated with the end of one Utterance and the beginning of another than any other phrase of the gestural arc (46% vs 21%).

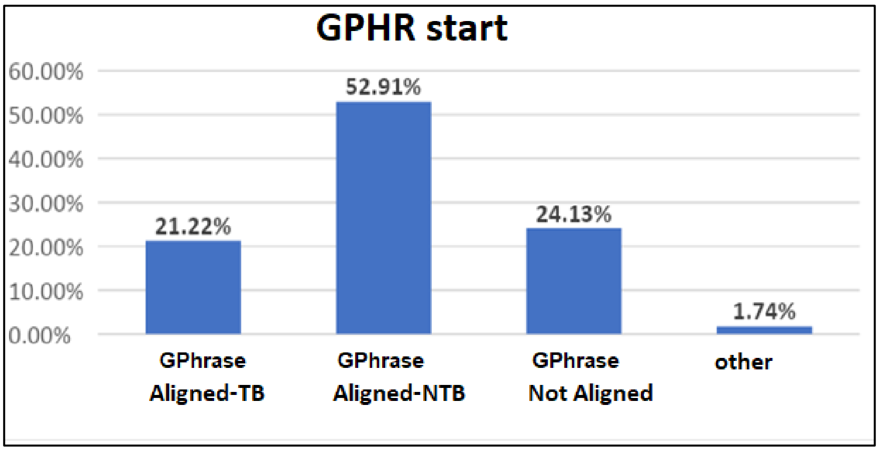


Figure 6: Alignment of the beginning of the Gesture Phrases with prosodic breaks.

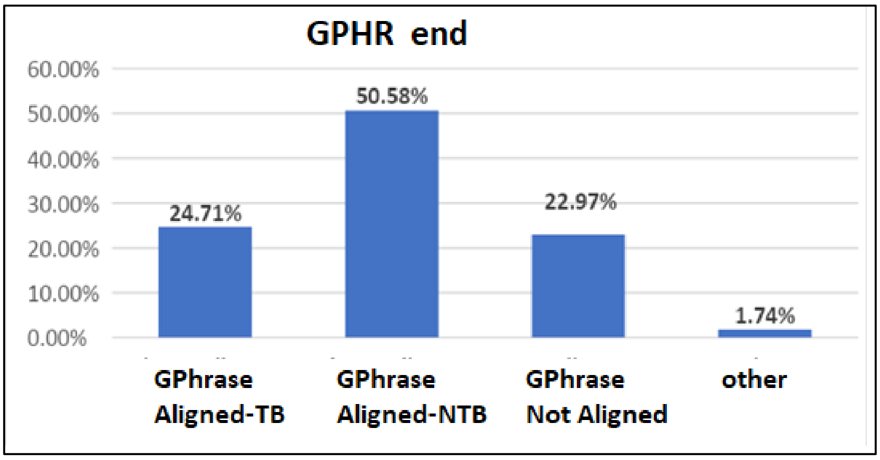


Figure 7: Alignment of the end of Gesture Phrases with prosodic breaks.

The correlation of the *onset* of a new phrase with a break, or with the beginning of a new information unit, is in any case high, but still lower than for GUs (about 75% compared to 92%). In other words, qualitatively interpreting the data the possibility that a GPhrase begins in correlation with a locution inside of the unit - above all at the conception of a new thought unit - even if less frequent still occurs with significant frequency.

Interpreting the data in terms of the linguistic scope of the gesture, we can hypothesize that roughly three times out of four the trigger for a GPhrase, being aligned with a break, might in principle have a scope that maps to the content of an entire information unit (usually consisting of a compositional syntagmatic constituent), rather than just a single word, or perhaps even an entire Utterance comprising its illocutionary value.

This hypothesis, although suggestive, can be validated by specifying the relationship between the semantics of the gesture and the linguistic units that constitute its scope in detail. In Section 6 we will consider to what extent this prevision may be true for gesture interpretation. The correlation between GUs and TBs is much higher at the conclusion of the gestural arc than at the beginning. As Figure 5 highlights, the GU tends to conclude very frequently in correspondence with the conclusion of an Utterance (in almost 70% of cases). More generally, the correspondence of the end of the GU with a *break* is even more compelling: in only 3.45% of cases does the gestural unity end without the occurrence of a prosodic break. Based on our data, we can therefore say that the release of the energy that bears the gesture almost necessarily correlates with the end of a prosodic program.

As for the GPhrases, their conclusion (Figure 7) aligns with a prosodic *break* in about 75% of cases, with the same preference for the non-terminal as with the *onset*.

The distinction between the level of the GU and the level of the GPhrase in relation to the linguistic Reference unit indicated by a TB is highlighted in the histograms in Figures 8 and 9. The distinction between the two levels is marked, in the sense that by its nature the GU can embrace multiple linguistic acts within it (in 43% of cases), while the gestural phrase lies within the linguistic act: a gesture practically never concerns elements belonging to two different Reference units (in only 1.5% of the GPhrases).

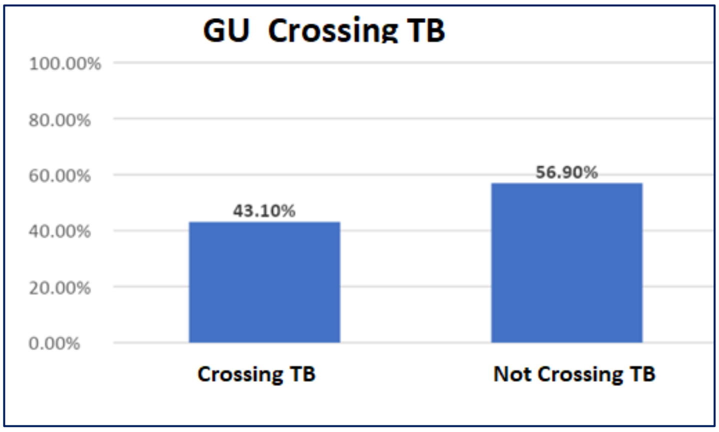


Figure 8: Gesture Units and multiple RUs.

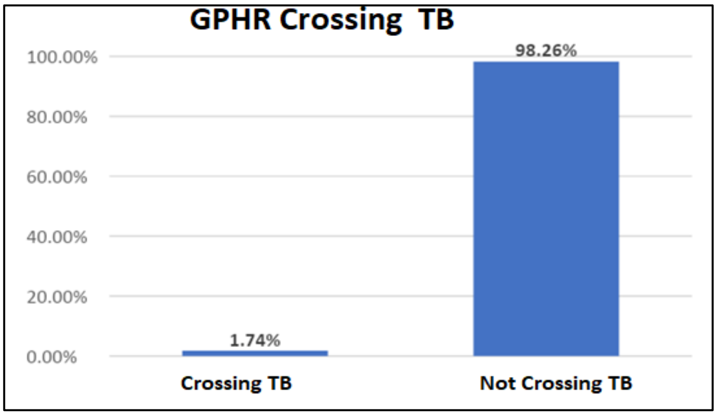


Figure 9: Gesture Phrases and multiple RUs.

This correlation is theoretically relevant as it corroborates with an independent argument the idea that underlies the L-AcT annotation system, that is the necessary relationship between the pragmatic concept of the Utterance (Reference unit) and the presence of a perceptually relevant terminal prosodic break: it is so true that the prosodic break signals the end of a linguistic act that a *co-speech gesture* never synchronizes with elements belonging to other linguistic activities; it forms a specific motor activity coordinated with the linguistic act itself, almost never (1.74%) different acts. In other words, the observable prosodic counterpart of the Utterance constitutes the maximum synchronization unit for the gestural phrase and is therefore the essential datum that the gestural annotation system must guarantee in order to correctly assign semantic relevance to a gesture.

An initial qualitative datum is also possible regarding the relationships between gesture and the level of the information units, which are annotated in the *dataset*. In particular, as we have reported, the prosodic units perform a specific information function of the Textual or Dialogical type. Each function defined in the L-AcT tag-set has been annotated independently of the gesture for each prosodic unit in the dataset, and it is therefore possible to check whether or not there are relationships between the gesture and the information level signaled by the intonation, considering whether there are restrictions in the gestural correlations, of the different types of information units that articulate the Utterance. We limit ourselves here to observing only basic data, which are nonetheless interesting in analyzing the relationship between the expressive phases of the gesture and the information functions.

Considering the expressive phases given in statements divided into several information units (i.e. consisting of units beyond the COM), it is observed that these may be found positioned, in principle, on all types of information units. Nonetheless, there exist preferences, which are highlighted in the histogram in Figure 10.

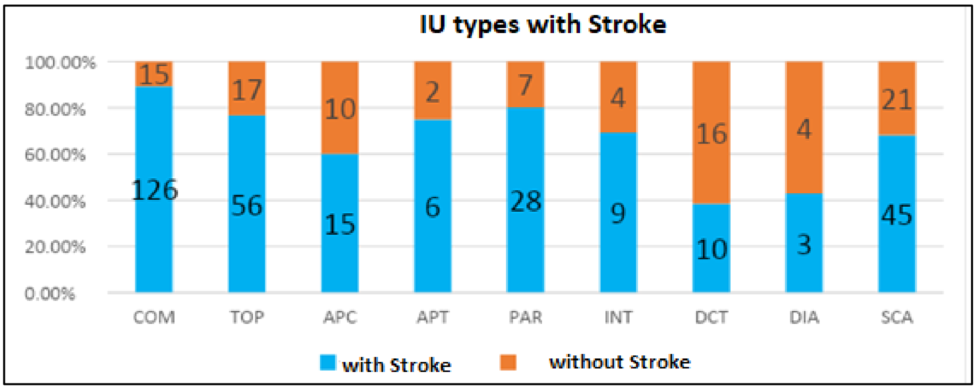


Figure 10: Types of information units and expressive phases within the Reference units.

The vast majority of units with textual functions contain a *Stroke*, while the probability that a Dialogic unit (the DCT and DIA bars) is marked by a *Stroke* is significantly lower (only two Dialogical information units out of 6 host an expressive phase, while the relationship is reversed for the textual units (5/6). This observation, if confirmed, could be linked to the semantic nature of the expressive phase of the gesture: the Dialogical units, by definition, are functional to the management of the relationship between the speaker and interlocutor and have little semantic content. The observation, however suggestive, is based on too scarce data and should be replicated on different types of corpora, since the spontaneous speech of the dataset in question (of the monological type) by definition hosts a very small number of Dialogical units.

6 Gesture typology and the *scope* of gesticulation

Following McNeill's "Kendon's continuum" (1992), the co-speech hand movements can be arranged along a continuum and classified as: *Gesticulation*, *Emblems*, *Pantomime*, and *Signs*. At one end of the continuum gesticulation co-occurs with speech, at the other end signs occur instead of speech. In relation to speech, linguistic properties, degree of conventionalization, and co-occurring referential meaning, gesticulation is not conventionalized but highly motivated, though not in an immediately comprehensible way. Emblems are conventionalized and symbolic, pantomime is not conventionalized but highly iconic and representative, while signs are arbitrary and conventionalized.

McNeill gives a more fine-grained differentiation for the gesticulation, pointing out that it can have different *Semiotic dimensions*, such as: *Metaphoric* (representing abstract concepts through concrete images), *Iconic* (representing images of concrete entities or actions); *Deictic* (having pointing functions connected with deixis); and *Beats* (rhythmic). Metaphoric, Iconic, and Deictic gestures are *Propositional* i.e. linked to the onset of ideas. Beats are *non-propositional* i.e. not directly linked to the content of speech. The Metaphoric and Iconic dimensions are also identified as *"pictorial", or* alternatively as *"representational*" or "*referential*" (Mittelberg and Evola, 2014).

In our annotation, the Semiotic dimension of each GPhrase have been recorded in a specific tier according to this classification, however the annotation schema is still provisional and, considering gesture / speech relations, a more qualitative description of gestural meaning and typology is under development. In particular we considered that Kendon (2004: p. 158) notes that some gestural components are not part of the referential meaning of the Utterance (*Referential* gestures) but are linked to other components of the speech production (*Pragmatic* gestures). Pragmatic gestures relate to the Utterance but do not participate in its propositional content, performing *Modal*, *Performative*, *Parsing* functions or alternatively *Regulating the dialogue*. Accordingly, the tag *pragmatic* gesture has been used in our annotation to mark some metaphoric gestures whose function was not connected to the propositional content of the Utterance, but which can be interpreted by taking into account informational and illocutionary values.

The relation between the form and the content of gestures derives from a metaphoric process "metaphoricity recruits iconicity" (McNeill, 2016: p. 94). Gesture interpretation can be performed by guessing the underlying metaphor from its link to the context (linguistic or non-linguistic). In what follows we will consider how the annotation of information units in accordance with prosodic boundaries allows us to select the specific scope of the gesture meaning supporting its interpretation. We will consider in particular the relation of gestures to different linguistic levels: a) the word level; b) the Information unit phrase; c) the Information unit function; d) the Illocutionary value.

a) A gesture frequently has scope on a single lexical item, and this link aids gesture interpretation. For instance, although the gesture and the information unit are fully synchronized, the gesture concerns only one specific word (underlined) of the IU, as in the Comment unit below and in Figure 11:

*MAX: [32][8] di meno <u>usuale</u> per me //COM
‘less usual for me’

A circular movement with both hands is repeated in front of the chest. The gesture, interpreted specifically in relation to the word "usual", resembles an action of unrolling something and may refer to a "repetition of events".

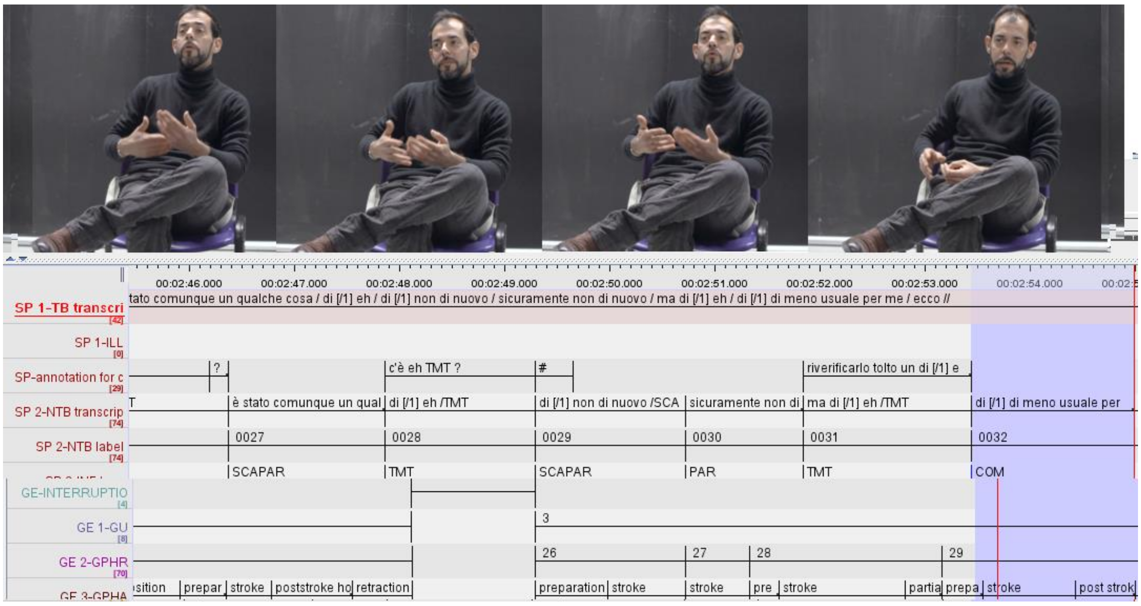


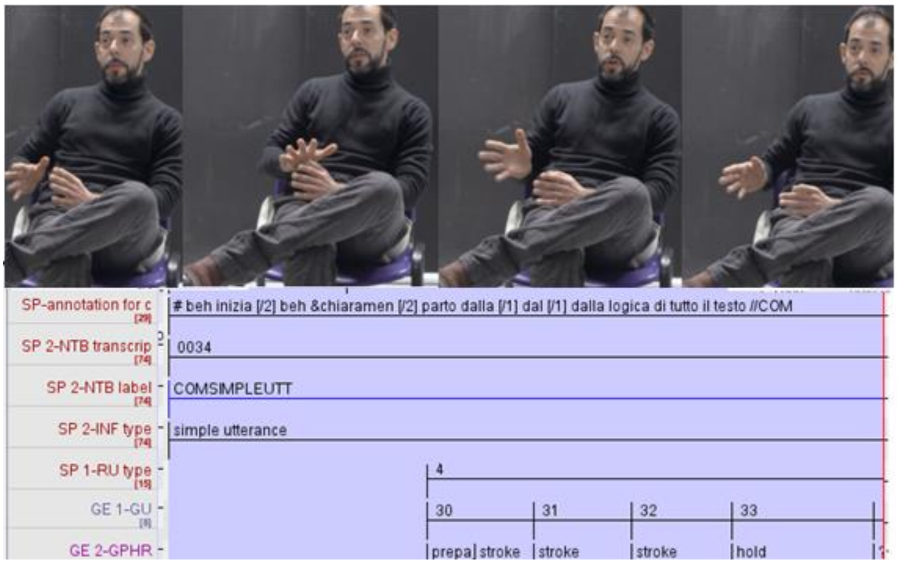
Figure 11: The gesture concerns only one specific word of the IU.

[8] Here and below the number refers to the ID of the IU

However, different words within the same information unit may be in relation with different expressive phases. For instance, the following four expressive phases occur within a Comment Unit, each one having the scope of a different concept expressed in the unit (Figure 12). Accordingly, we cannot claim a one to one correspondence of "gesture / information unit".

*MAX: [34] beh / inizia [/] beh chiarame [/] parto dalla [/] dal [/] dalla logica di tutto il testo //COM
'it starts [/] clearly [/] I start from [/] the logic of the whole text'

The interpretation of the metaphor underlying each gesture relies on its lexical anchor:



SP-annotation for c [29] # beh inizia [/2] beh &chiaramen [/2] parto dalla [/1] dal [/1] dalla logica di tutto il testo //COM
SP 2-NTB transcrip [74] 0034
SP 2-NTB label [74] COMSIMPLEUTT
SP 2-INF type [74] simple utterance
SP 1-RU type [15] 4
GE 1-GU [8] 30 | 31 | 32 | 33
GE 2-GPHR prepa | stroke | stroke | stroke | hold

Figure 12: Gestures concerning different words of the IU.

With the first *Stroke*, the speaker places both hands in front of him, like placing an object (or idea) in a stable position. The gesture, in relation to the words "inizia" [it starts], means "to fix a point of departure" for the interpretation of the text. The movement is coordinated with small, slight movements, that in relation to the word "chiarame<nte>" [clearly] might refer to the action of "weighing in his hands", as if considering or excluding alternatives.

In the second *Stroke* the actor moves his right-hand outwards slightly and rapidly repositions it to the central point (palms down), as if exploring a surface. The gesture, in relation to the word "parto" [I start] gives an iconic representation to the end of the mental process in which the actor explores alternatives for the interpretation of the text and reaches a valid conclusion to start with.

The third *Stroke* is a larger movement towards the outside of the wrist (open hand), like an action of "opening up", which in relation to the word "logica" [logic] represents the interpretation he found and is now pursuing, similar to a metaphor of "opening up a path".

In the last expressive phase, the speaker maintains the hand on the right side in a held position, and in conjunction with the phrase "tutto il testo" [all the text] the gesture indicates the support or continued relevance of the path just opened in the third *Stroke*, for the interpretation of the all text.

b) A gesture, as we saw, cannot cross Utterance boundaries, but may cross IU boundaries within the Utterance, as happens with the *Stroke* synchronized to the first two IUs of the following stanza (Figure 13):

*MAX: [61-62] e da lì /^{TOP} inizio a lavorare /^{COB} su [/] su tutta una serie di emozioni /^{SCA} di sentimenti /^{COB} di tutto ciò che provoca il personaggio / nell'arco di questo [/] del testo / del testo scritto //
'and from this point, I start working on the emotions, and sentiments, of all character stimulate, within the development of the text, of the written text'

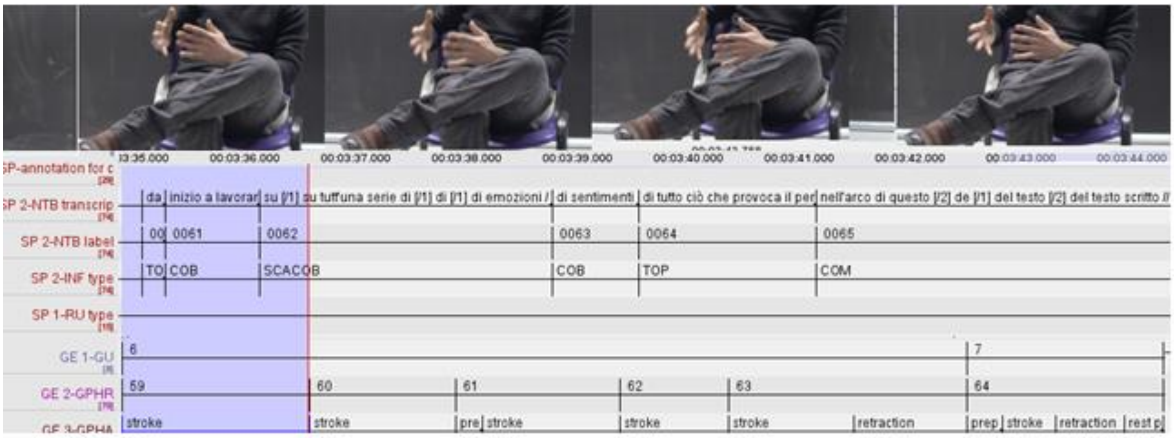
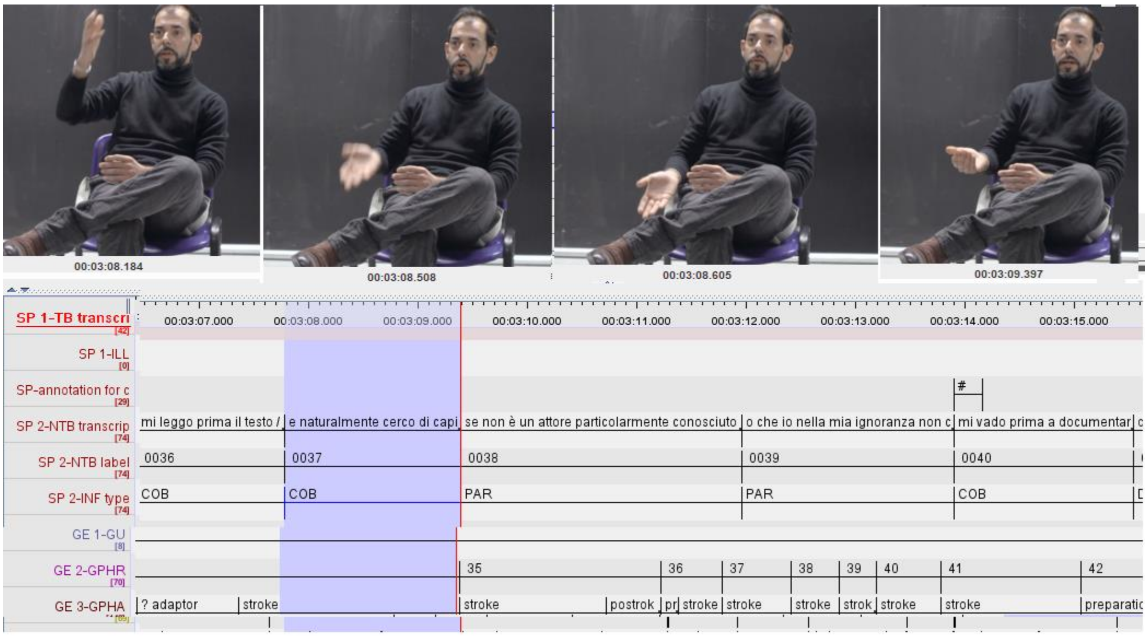


Figure 13: Gesture synchronized with two IUs and scope on the second Unit.

The *Stroke* consists of a down-up rotatory movement using both hands. In conjunction with the phrase "inizio a lavorare" [I start working], the gesture indicates the handling of an object with a significant volume. The repetition of this movement refers to its length over time. The whole gesture starts with the Topic unit, whose lexical content "da lì" [from there] is not the concern of the gesture, which synchronizes with both of the first two IUs but whose scope centers on the content of the second unit only.

c) A gesture does not necessarily have the scope of a single lexical unit but can concern a phrase in an IU, considered as a single unit of thought. The Bound Comment underlined in the following stanza provides an example (Figure 14):

*MAX: [36-37-38-39-40] mi leggo prima il testo /^{COB} naturalmente cerco di capire /^{COB} *se non è un attore particolarmente conosciuto* /^{PAR} o che io nella mia ignoranza non conosco/^{PAR} / mi vado prima a documentare //^{COM}
'I read the text before, of course I try to understand, if he is not a well-known actor, or I do not know him being an ignorant as I am, I try to be well documented'

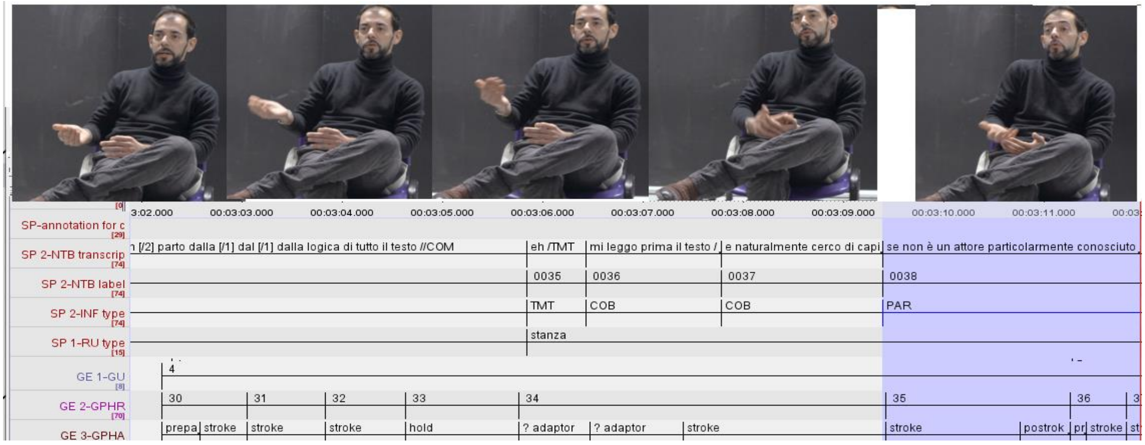


SP 1-TB transcri [42]
SP 1-ILL [0]
SP-annotation for c [29] | #
SP 2-NTB transcrip [74] | mi leggo prima il testo / | e naturalmente cerco di capi | se non è un attore particolarmente conosciuto | o che io nella mia ignoranza non c | mi vado prima a documentar | c
SP 2-NTB label [74] | 0036 | 0037 | 0038 | 0039 | 0040
SP 2-INF type [74] | COB | COB | PAR | PAR | COB | D
GE 1-GU [8]
GE 2-GPHR [70] | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42
GE 3-GPHA [69] | ? adaptor | stroke | stroke | postrok | pr | stroke | stroke | stroke | strok | stroke | stroke | preparatio

Figure 14: Gesture with scope on the locutive content of a IU.

The gesture consists of a slow movement lowering the hand from the head till the lap. The hand is opened toward the interlocutor and is raised a little at the end of the movement. The gesture refers to the process of understanding (“naturalmente cerco di capire” [of course I try to understand]). The underlying metaphor resembles “to bring the object under the eyes of everyone, keeping it visible at all times (as the raising movement in the last part of the gesture seems to suggest)”. The slow execution of the movement stresses the processual character of the understanding, while the focus is on the result reached (the understanding). Therefore, the scope of the single gesture concerns the whole concept expressed by the IU and is not relative to just one lexeme within it.

d) Beyond locutive relations, gestures can also regard functional aspects i.e *Pragmatic* gestures, according to Kendon. In this case the interpretation of the icon underlying the gesticulation is not anchored to any lexical expression (*propositional content*, in the traditional terminology), but is linked to the illocutionary value of the Utterance or to the information function of the IU marked by prosody. Our data set is too small to allow solid quantitative previsions, however we noticed that pragmatic gestures are much less frequent then propositional ones. In particular, we found only one gesture whose scope was clearly the information function expressed by the IU. This is the case with one co-speech gesture occurring in the Parenthetic unit of the previous Stanza (IU 38 *in italics*), which hosts two gestures (Figure 15). The first of the two (reported in the first four screenshots) can only refer to the Parenthetic function of the IU, according to our interpretation.



SP-annotation for c [29]

SP 2-NTB transcrip [74]: n [/2] parto dalla [/1] dal [/1] dalla logica di tutto il testo //COM | eh /TMT | mi leggo prima il testo / | e naturalmente cerco di capi | se non è un attore particolarmente conosciuto

SP 2-NTB label [74]: 0035 | 0036 | 0037 | 0038

SP 2-INF type [74]: TMT | COB | COB | PAR

SP 1-RU type [15]: stanza

GE 1-GU [8]: 4

GE 2-GPHR [70]: 30 | 31 | 32 | 33 | 34 | 35 | 36 | 3

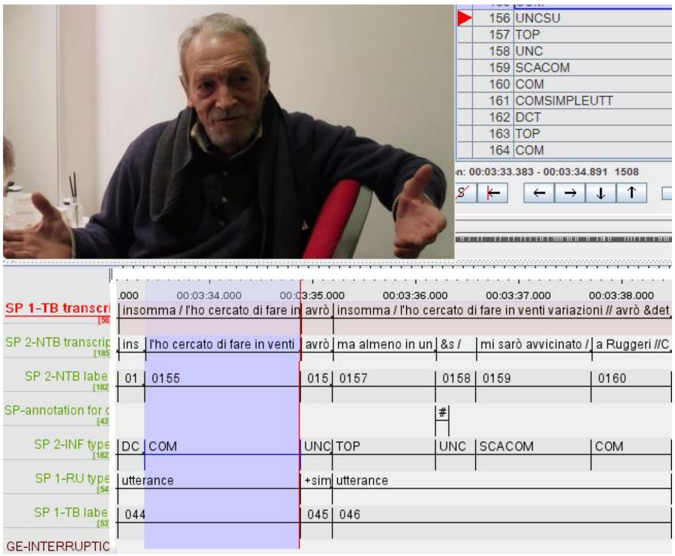
GE 3-GPHA: prepa | stroke | stroke | stroke | hold | ? adaptor | ? adaptor | stroke | stroke | postrok | pr | stroke | st

Figure 15: Gesture correlating with the Parenthetic function of the IU.

The gesture consists of a rapid hand movement from the speaker's right side to their lap. The hand is open and turned toward the speaker. The movement does not appear to have any link with the propositional content expressed in the parenthesis ("se non è un attore particolarmente conosciuto" [if he is not a well-known actor]). On the contrary, the gesture can be interpreted metaphorically as "from the side of the speaker" that explicitly, in coordination with prosodic change, introduces the epistemic point of view proper to the Parenthesis information unit. The second gesture (in the last screenshot) on the contrary refer to the content of one word of the IU; i.e. "conosciuto" [known].

Finally, gestures can be interpreted specifically in relation to the illocutionary value of the Utterance (Figure 16). The following Utterance, performing an Expressive act (*Expression of obviousness* according to the tag set of Cresti, 2020), is a clear example of this.

*ORS:[155] l'ho cercato di fare in venti variazioni … COM
'I tried to do it in twenty different variants …'
[the actor tells of the circumstance in which he was not able to imitate the performance given by a colleague]



156 UNCSU
157 TOP
158 UNC
159 SCACOM
160 COM
161 COMSIMPLEUTT
162 DCT
163 TOP
164 COM

n: 00:03:33.383 - 00:03:34.891 1508

SP 1-TB transcri [56]: insomma / l'ho cercato di fare in | avrò | insomma / l'ho cercato di fare in venti variazioni // avrò &det

SP 2-NTB transcrip [185]: ins | l'ho cercato di fare in venti | avrò | ma almeno in un | &s / | mi sarò avvicinato / | a Ruggeri //C

SP 2-NTB labe [182]: 01 | 0155 | 015 | 0157 | 0158 | 0159 | 0160

SP-annotation for c [43]: #

SP 2-INF type [182]: DC | COM | UNC | TOP | UNC | SCACOM | COM

SP 1-RU type [54]: utterance | +sim | utterance

SP 1-TB labe [53]: 044 | 045 | 046

GE-INTERRUPTIC

Figure 16: Gesture correlating with the Illocutionary value of the Utterance (*Expression of obviousness*).

The Utterance performs an illocutionary act of *Expression of obviousness* and is synchronous with a gesture in which arms are enlarged and hands are opened toward the interlocutor. Again, the gesture cannot be anchored and interpreted in relation to the propositional content ("l'ho cercato di fare in venti variazioni …" [I tried to do it in twenty different variants …]) but is immediately clear in correlation with the concept of "obviousness" ("open arms" is a possible visual metaphor for "as everybody can see").

6 Conclusions

The hierarchical annotation models adopted for the annotation of gesture and prosody allow their relationship to be examined on multiple quantitative and qualitative levels. Gesticulation goes hand in hand with spontaneous Italian speech for about 90% of the speech flow examined. The beginning and end of the gestural arcs are synchronous with the beginning and end of the prosodic units, independent of the positioning of these units across one or more Reference units, while the gestural phrase, although they have a tendency to start and end with the prosodic units, may also be located within them. Gesture Phrases always remain within a specific Reference unit and never cross the terminal prosodic boundaries. The Reference unit therefore appears to be the largest unit for the correlation of gestures. The expressive phases may be placed in all types of textual information units, but, as far as we can suppose on the basis of a small monological dataset, they tend not to mark the discursive signals, that is to say the dialogical information units, which are functional in managing the communication event.

From a qualitative point of view considering the synchronization of gestures with linguistic units (marked by terminal and non-terminal prosodic boundaries) allows us to identify the scope of the gesture and to support its interpretation. Gestures may have scopes at different linguistic levels: a) the word level; b) the Information unit phrase; c) the Information unit function; d) the Illocutionary value.

REFERENCES

1. Amir N, Vered SV, Izre'el S. *Characteristics of Intonation Unit Boundaries in Spontaneous Spoken Hebrew: Perception and Acoustic Correlates.* In Bel B, Marlien I. (eds.), Proceedings of Speech Prosody 2004, ISCA, 2004, 677–680.

2. Austin J. *How to Do Things with Words*. Oxford: Oxford University Press, 1962.

3. Bressem J, Ladewig SH., Müller C. *Linguistic Annotation System for Gestures (LASG).* In Müller C, Cienki A, Fricke E. Ladewig SH., McNeill D, Teßendorf S. (eds.), Body – Language – Communication: An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38) Vol. 1, Berlin: De Gruyter Mouton, 2013, 1098–1125.

4. Beckman M, Hirschberg J, Shattuck-Hufnagel S. *The original ToBI system and the Evolution of the ToBI Framework. In Sun-Ah J.* (ed.), Prosodic Typology. The Phonology of Intonation and Phrasing. Oxford: Oxford University Press, 2005, 9–54.

5. Boersma P, Weenink D. *Praat: Doing phonetics by computer*. Software. 2005. Retrieved from: http://www.praat.org/.

6. Buhmann J, Caspers J, van Heuven V, Hoekstra H, Martens JP, Swerts M. *Annotation of prominent words, prosodic boundaries and segmental lengthening by no-expert transcribers in the spoken Dutch corpus.* In Gonzales Rodriguez M, Suarez Araujo C. (eds.) Proceedings of the 2nd International Conference on Language Resources and Evaluation *(LREC 2002).* Paris: ELRA, 2002, 779–785.

7. Cantalini G. *La gestualità co-verbale nel parlato spontaneo e nel recitato. Annotazione del gesto e correlati prosodici in campioni comparabili di attori italiani,* PhD thesis - Università Roma Tre, 2018.

8. Cantalini G, Gagliardi G, Moneglia M, Proietti M. *La correlazione gesto/prosodia e la sua variabilità: il parlato spontaneo di contro alla performance attorale*. In: Atti del convegno: *GSCP* (Napoli, 12-14 Dicembre 2018), 2020.

9. Chafe W. *Discourse, consciousness, and time*: *The flow and displacement of conscious experience in speaking and writing*. Chicago: UCP, 1994.

10. Cheng W, Greaves Ch, Warren M. *A Corpus-driven Study of Discourse Intonation: The Hong Kong Corpus of Spoken English.* Amsterdam: Benjamins, 2005.

11. Cresti E. *Corpus di italiano parlato.* Firenze: Accademia della Crusca, 2000.

12. Cresti E, Moneglia M. (eds) *C-ORAL-ROM. Integrated reference corpora for spoken romance languages.* Amsterdam: Benjamins, 2005.

13. Cruttenden A. *Intonation*. Second edition. Cambridge: Cambridge University Press, 1997.

14. Crystal D. *The English Tone of Voice*. London: Edward Arnold, 1975.

15. Danieli M, Garrido JM, Moneglia M, Panizza A, Quazza S, Swerts M. *Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech C-ORAL-ROM*. In Lino MT, Xavier MF, Ferreira F, Costa R, Silva R. (eds), Proceedings of the 4th LREC Conference. Paris: ELRA, 2004, 1513–1516.

16. Du Bois JW, Schuetze-Coburn S, Cumming S, Paolino D. *Outline of discourse transcription.* In Edwards JA., Lampert MD. (eds): Talking data: Transcription and coding in discourse research. Hillsdale NJ: Lawrence Erlbaum, 1993, 45–89.

17. Duncan S. *Annotative practice (under perpetual revision)*. Retrieved from: http://mcneilllab.uchicago.edu/pdfs/susan_duncan/Annotative_practice_REV-08.pdf, 2008.

18. ELAN (Version 5.9) [Computer software]. (2020). *Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive*. Retrieved from: https://archive.mpi.nl/tla/elan

19. Gagliardi G. *Inter-Annotator Agreement in linguistica: una rassegna critica / Inter-Annotator Agreement in Linguistics: A Critical Review.* In: Cabrio E, Mazzei A, Tamburini F. (eds) Proceedings of the Fifth Italian Conference on Computational Linguistics - CLiC-it 2018. (Torino, 10-12 Dicembre 2018). Torino: Accademia University Press, 2019, 206–212.

20. Halliday M, Kirkwood A. *Intonation and Grammar in British English*. The Hague: Mouton, 1967.

21. 't Hart, J., Collier, R., Cohen, S., *A Perceptual Study of Intonation. An Experimental-Phonetic Approach to Speech Melody.* Cambridge: Cambridge University Press, 1990.

22. Helmic I, Lausberg H. *Hand movements with a phase structure and gestures that depict action stem from a left hemispheric system of conceptualization*. Experimental Brain Research, 232(10): 3159–3173, 2014.

23. Holle H, Rein R. *The Modified Cohen's Kappa: Calculating Interrater Agreement for Segmentation and Annotation. In Lausberg H.* (ed.), Understanding Body Movement. A Guide to Empirical Research on Nonverbal Behaviour. With an Introduction to the NEUROGES Coding System. Frankfurt am Main: Peter Lang, 2013, 261–275.

24. Izre'el S, Mettouchi A. *Representation of Speech in CorpAfroAs*. Transcriptional Strategies and Prosodic Units. In Mettouchi A, Vanhove M, Caubet D. (eds) Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages. Amsterdam: Benjamins, 2015, 13–41.

25. Izre'el S, Mello H, Panunzi A, Raso T. (eds), *In Search of Basic Units of Spoken Language*. Amsterdam: Benjamins, 2020.

26. Karcevsky S. *Sur la phonologie de la phrase*. Travaux du Cercle linguistique de Prague, IV: 188–228, 1931.

27. Kendon A. *Some relation between body motion and speech: An analysis of an example*. In Siegman, AW, Pope B. (eds), Studies in Dyadic Communication New York: Elsevier, 1972, 177–210.

28. Kendon A. *Gesticulation and speech: Two aspects of the process of Utterance.* In Key MR. (ed.), Nonverbal Communication and Language. The Hague: Mouton, 1980, 207–227.

29. Kendon A. *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press, 2004.

30. Kita S, van Gijn I, van der Hulst H. *Movement phases in signs and co-speech gestures, and their transcription by human coders*. In Wachsmuth I, Fröhlich M. (eds), Gesture and Sign Language in Human-Computer Interaction Berlin: Springer, 1998, 23–35.

31. Krippendorff K. *Content Analysis: an introduction to its Methodology*. Thousand Oaks, CA: Sage Publications, 1980.

32. Ladewig SH, Bressem J. *A linguistic perspective on the notation of gesture phases.* In Müller C, Cienki A, Fricke E, Ladewig SH, McNeill D, Teßendorf S. (eds), Body – Language – Communication: An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38) Vol. 1, Berlin: De Gruyter Mouton, 2013, 1060–1079.

33. Lausberg H. *Understanding Body Movement. A Guide to Empirical Research on Nonverbal Behaviour, With an Introduction to the NEUROGES Coding System*. Frankfurt am Main: Peter Lang, 2013.

34. Loehr D. *Gesture and prosody.* In Müller C, Cienki A, Fricke H, Ladewig SH, McNeill D. (eds), Body – Language – Communication: An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38, Vol. 2. Berlin: De Gruyter Mouton, 2014, 1381–1391.

35. McClave E. *Intonation and Gesture.* Ph.D. dissertation. Washington, DC: Georgetown University, 1991.

36. McNeill D. *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press, 1992.

37. McNeill D. *Gesture and Thought.* Chicago: University of Chicago Press, 2005.

38. McNeill D. *Why We Gesture: The Surprising Role of Hand Movements in Communication*. Cambridge: Cambridge University Press, 2016.

39. Moneglia M, Cresti E. *Intonazione e criteri di trascrizione del parlato*. In Bortolini U, Pizzuto E. (a cura di) *Il progetto CHILDES Italia.* Pisa: Del Cerro, 1997, 57–90.

40. Moneglia M, Raso T, Malvessi-Mittmann M, Mello H. *Challenging the perceptual relevance of prosodic breaks in multilingual spontaneous speech corpora: C-ORAL-BRASIL / C-ORAL-ROM*. In Speech Prosody 2010, W1.09, Satellite workshop on Prosodic Prominence: Perceptual, Automatic Identification. Chicago. Retrieved from: https://www.isca-speech.org/archive/sp2010/sp10_2010.html, 2010.

41. Moneglia M, Raso T. *Notes on the Language into Act Theory.* In Raso T, Mello H. (eds), Spoken corpora and linguistics studies. Amsterdam: Benjamins, 2014, 468–494.

42. Nencioni G. *Parlato-parlato, parlato-scritto, parlato-recitato*. Strumenti critici, LX, 1–56, 1976.

43. Panunzi A, Gregori L, Rocha B. *Comparing annotations for the prosodic segmentation of spontaneous speech: Focus on reference units*. In Izre'el S, Mello H, Panunzi A, Raso T. (eds), In Search of Basic Units of Spoken Language. A corpus-driven approach. Amsterdam: John Benjamins, 2020, 403–431.

44. Quirk R, Greenbaum S, Leech G, Svartvik J. *A Comprehensive Grammar of the English Language*. London/New York: Longman, 1985.

45. Raso T, Mello H. (eds), *C-ORAL-BRASIL I: Corpus de referência de português brasileiro falado informal*. Belo Horizonte: Editora UFMG, 2012.

46. Raso T, Barbosa P, Cavalcante F, Malvessi-Mittmann M. *Segmentation and analysis of the two English excerpts.* In Izre'el S., Mello H., Panunzi A., Raso T. (eds), In Search of Basic Units of Spoken Language. A corpus-driven approach. Amsterdam: John Benjamins, 2020, 309–325.

47. Saccone V, Vieira M, Panunzi A. *Complex Illocutive Unit in Language into Act Theory: an analysis of non-terminal prosodic breaks of Bound Comments and Lists*. Journal of Speech Sciences, 7.2. 51–64, 2018.

48. Sorianello P. *Per una definizione fonetica dei confini prosodici*. In Pettorino M, Giannini A, Savy R. (eds) Atti del Convegno Internazionale, La comunicazione parlata. Napoli: Liguori, 2006, 310–330.

49. Spoken Dutch Corpus: available at: http://lands.let.ru.nl/cgn/doc_English/topics/project/pro_info.htm
50. WINPITCH. *Software for prosodic research, with on the fly aligner, real-time spectrograph, multitracking F0 analysis, video and audio analysis, and much more (Free installation password required after 30 days of use).* Available at: https://www.winpitch.com/.