Did Somebody Say 'Gest-IT'?
A Pilot Exploration of Multimodal Data Management

Ludovica **Pannitto**[1,*], Lorenzo **Albanesi**[1], Laura **Marion**[1], Federica Maria **Martines**[1], Carmelo **Caruso**[1], Claudia S. **Bianchini**[2], Francesca **Masini**[1] and Caterina **Mauri**[1]

[1]*Alma Mater Studiorum - University of Bologna*

[2]*University of Poitiers, FoReLLIS Laboratory*

Abstract

The paper presents a pilot exploration of the construction, management and analysis of a multimodal corpus. Through a three-layer annotation that provides orthographic, prosodic, and gestural transcriptions, the *Gest-IT* resource allows to investigate the variation of gesture-making patterns in conversations between sighted people and people with visual impairment. After discussing the transcription methods and technical procedures employed in our study, we propose a unified CoNLL-U corpus and indicate our future steps.

Keywords

Corpora, Multimodality, Gestuality, Blindness, Universal Dependencies

1. Introduction

Corpora represent the main tool for linguists to observe language in its real use and verify its general trends on both a quantitative and qualitative basis [1]. Today, written language corpora are the most used, thanks to the greater availability of written data and the ease of processing. However, in speech, speakers appeal to numerous semiotic sources (e.g., spoken channel, gestures, proxemics, facial expressions, etc.) to create and convey meaning, and written corpora fail to account for this richness of modalities. To effectively study how language works, one should observe these different semiotic sources independently of each other and take their interactions into account [2]. To capture this complexity, it is necessary to go beyond written data and use multimodal corpora, namely collections of audio-visual linguistic data that allow to both hear and see linguistic productions.

Multimodal corpora can be used to analyze a wide variety of linguistic phenomena, especially those related to the use of body in human communication and interaction, and to the way bodily communication and spoken language interact to generate meaning. They are, therefore, the primary sources for the analysis of co-speech gestures [3] and Sign Languages. Following [4, 5], we define a multimodal corpus as 'an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language'. Multimodal corpora come in various shapes, depending on the nature of the communication channel captured by the resource. For the sake of this article, we specifically restrict our attention to resources including both a video and audio recording of linguistic content[1]. The long tradition of analyzing written and spoken data has, over time, led to the development of transcription systems (such as IPA, orthography, and various prosodic conventions), which have become recognized standards within the linguists' community. These systems allow for an "objective" description of speech, independently of any considerations on the functions or the meanings of the described elements. For gestural data, some transcription systems have also been proposed, such as the Linguistic Annotation System for Gestures (LASG; [6]), but none of them has attained enough acceptance to qualify as a standard [4]. In the absence of an effective transcription system, hybrid solutions are often employed, situated between transcription and annotation, where the choices for describing gestural forms reflect their attributed functions (e.g.: a "shrug" is a shoulder movement, but this

✉ ludovica.pannitto@unibo.it (L. Pannitto);
lorenzo.albanesi@studio.unibo.it (L. Albanesi);
laura.marion@studio.unibo.it (L. Marion);
federica.martines2@studio.unibo.it (F. M. Martines);
carmelo.caruso@unibo.it (C. Caruso);
claudia.savina.bianchini@univ-poitiers.fr (C. S. Bianchini);
francesca.masini@unibo.it (F. Masini); caterina.mauri@unibo.it
(C. Mauri)

🆔 0000-0002-5760-6447 (L. Pannitto); 0000-0002-4783-1202
(C. S. Bianchini); 0000-0002-2354-4612 (F. Masini);
0000-0002-8226-7846 (C. Mauri)

---

[1]There exists a wide variety of multimodal resources for spoken and signed language, many of them openly available to the community through initiatives such as CLARIN-ERIC (https://www.clarin.eu/). For a collection of available multimodal resources see https://www.clarin.eu/resource-families/multimodal-corpora (spoken language) and https://www.clarin.eu/resource-families/sign-language-resources (sign language), while a list of audio-only resources of spoken language can be found at https://www.clarin.eu/resource-families/spoken-corpora

label is often used to refer to any pragmatic gesture of epistemic denial performed by moving the shoulders, without specifying either the characteristics of the shoulder movement or the movements of other body parts that may have contributed to the execution of the gesture). Similar challenges arise in studies on Sign Languages. Although there is a larger number of transcription systems for the latter (see, for instance, the review in [7]), none of them has achieved the status of a universal standard. Additionally, attempts to adapt these systems to the transcription of gestures have so far been limited and not particularly successful. The lack of a transcription standard for gestures, that describes them independently of their function or meaning, hinders the ability to precisely investigate the relationship between speech and gesture.

Another aspect concerns the nature of language data captured in multimodal resources: as the collection and standardization process for this kind of linguistic data is, by its very nature, much more complex, resources are often tailored to specific purposes and therefore involve task-oriented interactions (e.g., describing objects as in the *NM-MoCap-Corpus* [8]; spatial comunication tasks as in the *SaGA Corpus* [9]), thus capturing interactions that may be naturalistic but are inherently non-ecological, i.e. not naturally-occurring [10, 11]. Often, participants are asked to wear special devices such as headsets or trackers during the recordings [12, 13], clearly altering the spontaneity of the interaction.

The aim of the *Gest-IT* project is to build a multimodal corpus of ecological data, allowing for the integrated analysis of verbal and gestural communication in spontaneous interactions. In this paper, we will focus on the protocol of multimodal data management that we tested for this resource. We will first discuss the main existing multimodal resources (Section 2), showing how, as of today, there doesn't seem to be any ecological, accessible, multimodal corpus for Italian. We will then introduce the *Gest-IT* pilot resource and present its main features with respect to existing resources (Section 3). Section 4 outlines the main design choices taken for the creation of our resource and Section 5 describes the path ahead.

2. Multimodal resources: problems and overview

Multimodal corpus research faces two major problems: (i) the lack of existing transcription and annotation standards (tools, formats and schemes), especially for coding nonverbal behavior [4]; and (ii) the time consuming nature of transcription and annotation process, which is responsible for the relatively small sizes of searchable multimodal corpora that are currently available.

Specifically with respect to point (i), a major problem concerning available resources is the non-separation between the identification and description of gestures on the one hand, and their interpretation on the other. Indeed, in many resources and studies a particular gestural pattern is transcribed based on its *function*, i.e. its interpretation, rather than on a description of the 'objective' aspects that characterize its 'form'. However, if we aim to provide an integrated analysis of verbal and nonverbal communication, it is crucial that – just as we employ IPA or simplified orthographic transcriptions for verbal signs – we establish a standard to transcribe nonverbal signs in order to then annotate and interpret them. Furthermore, in most resources gesture is transcribed only with reference to verbal behaviour: the very identification of the gestures depends on their association, according to the annotator's subjective filtering, to an identifiable verbal sequence. In the *PoliModal corpus*[2] [14], a resource including transcripts of 14 hours of TV face-to-face interviews from the Italian political talk show *Mezz'ora in più*, for instance, gestures are annotated if they are judged as having a communicative intention [15] (displayed or signalled), or a noticeable effect on the recipient. Once a gesture has been selected, it is annotated with functional values, as well as features that describe its behavioural shape and dynamics. The descriptions provided for gesture annotation, moreover, seem to be an approximation of the movement: gestures are often described relying on the annotator's categorization and not using meaningful and objective parameters. For example, in the MUMIN coding [16] scheme used in the *PoliModal Corpus* and reported in Table 1, a number of possible values for each behaviour attribute are defined, but these fail to describe the entire range of possibilities (i.e., only three values are provided for face movements) or excessively simplify the description (i.e., the value *complex* is used to capture movements where several trajectories are combined, thus leaving unspecified whether they combine sequentially or in a non-linear trajectory for instance). Similar code schemes are used in the *Corpus d'interactions dialogales* (CID, [12]) and in the *Hungarian Multimodal Corpus* [17].

For resources such as *Natural Media Motion-Capture Corpus* (NM-MoCap-Corpus [8]), *Bielefeld Speech and Gesture Alignment Corpus* (SaGA [9]) and *BAS SmartKom Public Video and Gesture corpus* (SKP [18]), researchers decided to adopt McNeill's categories [19] or a schema inspired by them [20, 21]. In addition, some of the Swedish data in the *Thai/Swedish child data corpus* [22] were partially annotated thanks to the standard notation CHAT [23]. In the CORMIP [24] resource, instead, each gesture is segmented according to gesture phrases and gesture units [25]. Gestures are then classified solely based on iconicity, classifying them as 'Pictorial', 'Non-Pictorial' or 'Conventional'. While they claim to avoid

---

[2]https://github.com/dhfbk/InMezzoraDataset

Table 1

MUNIN [16] coding scheme

Behaviour attribute	Behaviour value
General face	Smile, Laugh, Scowl, Other
Eyebrow movement	Frown, Raise, Other
Eye movement	Extra-Open, Close-Both, Close-One, Close-Repeated, Other
Gaze direction	Towards-Interlocutor, Up, Down, Sideways, Other
Mouth openness	Open mouth, Closed mouth
Lip position	Corners up, Corners down, Protruded, Retracted
Head movement	Down, Down-Repeated, BackUp, BackUpRepeated, BackUp-Slow, Forward, Back, Side-Tilt, Side-TiltRepeated, Side-Turn, Side-Turn-Repeated, Waggle, Other
Handedness	Both hands, Single hands
Hand movement trajectory	Up, Down, Sideways, Complex, Other
Body posture	Towards-Interlocutor, Up, Down, Sideways, Other

Table 2

Gestures classification in CORMIP [26]

Pictorial	image-like shapes, or boundaries of a real-world object or action.
Non-Pictorial	rythmic movements (i.e., batonic) or geometric forms. Deictic gestures also fall within this category.
Conventional	gestures with a degree of conventionality that allows to associate, in a specific linguistic system, a semantic value to tehm (e.g., the 'okay' sign).

categorization of gesture functions or conventionality, the description of their lables (see Table 2) seems to contradict this statement [26]. Lastly, as far as Italian is concerned, the *Padova Multimodal Corpus* [27, 28] has to be mentioned, where textual transcriptions are enriched with annotations about a number of non-verbal components, there including also aspects such as gaze and gestures. The MultiModal MultiDimensional (M3D) labelling scheme[3] [29, 30] tries to decouple gesture transcription in the three different dimensions of its form, its relation to spoken prosody and its semantic or pragmatic functions. As reported in their manual, however, the transcriber is required to make choices, on the *form* layer, such as which is the predominant articulator (e.g., left or right hand, or both) or to choose for the articulator one of the provided forms, one of which is labeled as *iconic OK shape.*

The challenge of transcription becomes even more significant when dealing with multimodal corpora representing sign language. Typically, this issue is addressed using glosses, a form of sign-to-word translation that provides information about the meaning of signs without indicating their form [7]. However, over the years, some systems have been developed to represent the shape of signs. Most of these systems focus primarily on the hands [31], which are only a small part of the articulators contributing to meaning. Among these systems, Typannot [32] stands out as it offers a comprehensive

description of the entire set of body parts— from fingers to toes, including the head and torso — used to transcribe both sign languages and co-verbal gestures.

3. Towards the *Gest-IT* corpus: blind and sighted speakers

We aim at building a corpus consisting of maximally ecological interactions, transcribed on three separate layers aligned to each other: (i) an orthographic transcription; (ii) a prosodic transcription, and (iii) a gestural transcription. At present, we are still in an initial, exploratory phase, but we already addressed the most important decisions to be made.

The first decision concerned the informants to be recorded. In order to be able to investigate whether the ability to see and the perception of being seen during a communicative exchange can influence gesture production, we decided to take into consideration both sighted and visually impaired L1 speakers in dialogical situations. Gesture is indeed closely linked not only to intersubjective needs, connected to clarity, efficiency and attention-getting functions, but also to cognitive needs: speakers recur to gestures both when the interlocutor is not visible [33] and when the speaker is visually impaired [34], thus independently of the interlocutors' ability to see and interpret them. Yet, the actual relation and reciprocal influence between gestures and the perception of being seen has received little attention so far.

We included in the study 6 blind and 8 sighted participants, recruited on a voluntary basis and through a protocol that has been evaluated as compliant with GDPR and ethical requirements[4]. The blind group included speakers who were born blind, who acquired blindness later and who are partially-sighted. The total average age of the participants is mean = 39 years old (sd = ±18.7). The average age of the PG is mean = 55.8 years (sd = ±18), while the control group has an average age of mean = ±26 years old (sd = ±3.9). The total gender distribution is 85.7% F and 14.2% M. In the blind goup (BG) 100% of the participants are F. In the sighted group (SG) 75% are F and 25% are M. The total average educational level distribution shows that 64.2% of participants has a bachelor's degree, while 35.7% has a high school diploma. In the BG 83.3% of participants has a high school diploma, while 16.7% of the participants has a bachelor's degree. In the SG 100% of participants in the control group has a bachelor's degree.

All participants were paired and later involved in 30-minutes seated conversations, to elicit samples of spontaneous speech. As the participants to each dialogue were
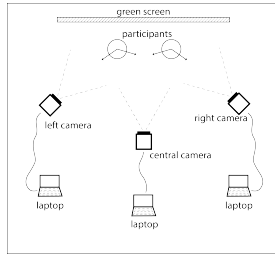
Figure 1: Room setting for the recording sessions, Logitech Brio Stream Webcam were employed for recordings

unlikely to know each other, in order to avoid moments of silence, some questions were prepared to enhance spontaneous conversations (see Appendix A). Interestingly, speakers recurred to these prompts only in few cases: the interactions developed very spontaneously despite the absence of previous contacts among the interlocutors.

We built the pairs and the interactional setting according to two parameters:

- speakers could belong to the same category of participant (both blind, both sighted) or different categories. We coded these two situations as S (*same*, blind-blind or sighted-sighted conversation) or D (*different*, blind-sighted conversation);
- speakers could be facing each other or be seated back-to-back, to ensure that participants could not perceive the other's nonverbal communication. We coded these two situations as M (*masked*, back-to-back situation) or U (*unmasked*, facing situation).

We recorded 13 conversations, for a total of roughly 7 hours (428.15 minutes), from three points of view: the central camera faced the couple, whereas the other two recorded the left side and the right side (the left and right cameras were located so that they could capture the participants frontally, see Figures 1 and 2). The goal was to take the participants' gestures from all possible perspectives. Recordings took place over two days. Some details about the 13 recorded conversations are available in Table 3[5].

4. The *Gest-IT* corpus schema

The other decisions that we had to make from the very beginning concerned the repository, the archiving protocol, and the standards for transcribing the three layers we aim to represent (orthographic, prosodic and gestural).

---

[5]Interactions involving only blind speakers did not require the masked setting, which was aimed to let sighted speakers experience a sight impairment of some sort during in-presence communication.

Table 3
The number of recording sessions in each considered condition (i.e., M=masked conversational situation, U=unmasked conversational situation, S=same sight conditions, D=different sight conditions). The total amount of recorded speech is also reported for each condition (in minutes).

	M	U		mins
S	3	4	7	225.44
D	3	3	6	202.71
mins	6	7		
	198.1	225.4		

(a) Unmasked scenario, Same sight conditions

(b) Masked scenario, Different sight conditions

Figure 2: Recording scenarios: unmasked and masked, same and different sight conditions

The next Sections are devoted to discuss these aspects in detail.

4.1. Data repository

Resource building is a team enterprise, performed asynchronously by a number of different people (i.e., PIs, interns, technicians etc.), often with different levels of technical expertise and background knowledge about the genesis of the data. Our project is no exception.

Therefore, in order to ensure data consistency and maintenance, a specific workflow has been put in place. More specifically, a central `git` repository[6] keeps track of the status of the resource. The `main` branch contains the last, released version of the corpus while the `dev` branch is used for development in between releases (versions are numbered according to semantic versioning standards).

Each participant and each conversation is defined through a `.yaml` file (Appendix B), allowing for a number of CI/CD practices to be put in place: each time a new conversation description file is pushed to the repository, for instance, a table summarizing the full status of the resource is generated. Similarly, automatic checks are performed each time a transcription is updated to ensure the consistency of the overall resource: for instance, a script makes sure that names of layers in the transcription correspond to participants, that jeffersonian notation (see Section 4.2) is well formed, etc.

---

[6]https://github.com/LaboratorioSperimentale/Gest-IT

Table 4
Jefferson prosodic conventions

Symbol	Description	Symbol	Description
.	Descending intonation	?	Rising intonation
(.)	Short pause	cia-	Interrupted word
>ciao<	Faster pronunciation	=	Prosodically bound units
,	Weakly rising intonation	<ciao>	Slower pronunciation
°ciao°	Lower volume	:	Prolonged sound
[ciao]	Overlap between speakers	CIAO	Louder volume

Data pertaining to each conversation is constituted by a set of digital objects, that represent different layers of information attached to the same recording. These include: (i) three video tracks and one audio track; (ii) a verbal transcription layer, which was initially automatically created with the `whisper` ASR toolkit [35] and then revised at the ortographic and prosodic level (Section 4.2); (iii) gesture transcription, starting from video sources (Section 4.3); (iv) UD annotation layers.

Transcriptions are maintained in CoNLL-U format[7], with specific MISC features for the gesture component. This will allow, in the future, to enrich the resource with additional annotation layers.

4.2. Verbal language transcription

As regards verbal communication, we decided to adopt the standards of the KIParla corpus [36], a corpus of spoken Italian that allows full access to audio files and transcriptions of roughly 153 hours of spontaneous speech [8].

Once the recordings were acquired, the transcription process began. In accordance with the KIParla protocol, it was agreed to use the ELAN software [37], which allows for time alignment of videos, audio files and transcriptions. In practice, the speech was segmented into transcription units identified on a perceptual basis, especially by reference to prosodic unit boundaries. The transcription process involved two steps:

- *orthographic transcription*, which included anonimization, turn assignment, and nonverbal behaviours. Whenever the annotator didn't understand, they could either choose 'xxx' or type their hypothesis in parentheses;
- *prosodic transcription*, following a simplification of the Jefferson system [38], widely shared by the scientific community [39]. The employed conventions [40] are reported in Table 4.

Transcriptions are thus available in two formats: they can be read as simple orthographic texts, or they can be read as enriched texts with prosodic and interactional information (such as overlaps, speed alterations, ascending or descending intonation, pauses, etc., as in example below). In both cases, it is possible to directly relate the transcription unit to the audiovisual unit. A further revision step will be done once the corpus will be fully transcribed, in order to make sure that notation is consistent throughout the resource.

(1) S001   l'ultimo che ho fatt:o   allora sono
speak-id -   -   -   Prolonged -   -
stata a siviglia:,   p[er   natale]
-    - Prolonged+Ascending overlap overlap
'my last one well I was in Seville for Christmas'

(2) B001   [che   io   ador]o che [io
speak-id overlap overlap overlap -   overlap
adoro]
overlap
'(Seville) which I love'

(3) S001   [bellissima po]i   a [natale è
speak-id overlap   overlap - overlap overlap
stato   mag]ico
overlap
'wonderful Christmas was magic'

(4) B001   [siviglia meravigliosa]
speak-id overlap   overlap
'wonderful Seville'

4.3. Gesture transcription

In order to provide also a transcription of gestures, as objective and interpretation-independent as possible, we decided to employ Typannot.

Typannot is a typographic system for the representation of sign languages, a project in development since 2013 by the *Gestual Script* research group, composed of linguists, graphic designers, typographers, and computer scientists. Its articulatory description of the body, independent of the language studied, allows it to be adapted

sent_id	unique identifier for transcritpion unit
text	space-separated sequence of token forms
conversation_id	unique identifier of conversatin
speaker_id	unique identifier of conversation participant
duration	duration of segment
overlaps	space-separated list of other sent_ids
text_jefferson	for speech units, original prosodic/jeffersonian transcription
type	for gestural units, identifier of the articulator

Table 5

Metadata describing each transcription (tu-xxx) or gestural unit (gu-xxx) in our corpus. `sent_id` and `text` are derived from the UD format, while others are introduced for the purpose of this resource.

to the study of gestures as well. Typannot proposes to analyze gestures and signs as realizations of the body and not just the hands: to facilitate analysis, the body is divided into different Articulatory Systems (AS), covering every body part from the hands to the feet and includes a description of facial expressions. For the purpose of the *Gest-IT* project, only three will be considered:

Finger (F): the dynamics of the fingers of the hand (thumb, index, middle, ring, and little finger). Furthermore, the distinction between the fingers of the right hand and those of the left hand will be considered and referred to respectively as RH and LH;

UpperLimb (UL): the dynamics of the upper limbs (arm, forearm, hand);

UpperBody (UB): the dynamics of the segments that make up torso (hip, spine and shoulder), neck and head.

In this system, the sign's form is seen as a set of articulatory body information (we extend this view to gestures). Currently, the generic characters that make up the graphic inventory of Typannot are used to describe the dynamics of all body segments.

4.4. Towards a unified CoNLL-U corpus

The resulting corpus is composed of verbal-prosodic units and gestural units, with information about their overlaps[9]. Each unit is described by the metadata listed in Table 5. In case of non verbal units, the text is filled with a placeholder token (EMPTY) and relevant information is contained in the MISC column, where the following features are introduced, `meta` for para-verbal information (such as laughs, coughs...) and `gesture` for Typannot codes (see Appendix C).

---

[9]At the moment of writing, 1 minute of pilot transcription has been produced.

5. Future steps

The aim of this paper is to share with the scientific community the protocol developed to build a multimodal resource for the Italian language in terms of data collection (design, ethic issues, practicalities); data management and curation; data transcription, annotation and analysis. In doing so, we contribute to the debate on multimodal resource building, which is still lacking an established standard. In particular, our contribution in this respect is twofold.

Firstly, our study suggests to adopt a three-layer transcription where the three layers (i.e., the orthographic transcription, the prosodic/interactional transcription, and the gestural transcription) align to each other, by using ELAN as a tool for transcribing and CoNLL-X as an interoperable output format. This has the advantage of grounding gestures as an integrated semiotic source within verbal conversation and ultimately allows to unveil gesture-speech regularities.

Secondly, we propose an innovative approach for the annotation of gesture data. By relying on common practices in the field of sign languages, we suggest that gesture transcription should follow the same rationale of phonetic transcription, with a method that describes 'objective' aspects that characterize the 'form' of the gesture, thus allowing for an interpretation-independent annotation.

Clearly, the project is still at a very preliminary stage. Next steps will include the complete orthographic, prosodic and gesture transcription of the recordings; a thorough revision and pseudoanymization.

References

[1] A. Lüdeling, M. Kytö, Corpus Linguistics: An International Handbook, De Gruyter Mouton,

---

[10]https://site.unibo.it/laboratorio-sperimentale/
[11]https://www.cavazza.it/

2009. URL: https://www.degruyter.com/database/COGBIB/entry/cogbib.7917/html.

[2] J. Bezemer, C. Jewitt, Multimodality: A guide for linguists, Research methods in linguistics 28 (2018).

[3] N. Abner, K. Cooperrider, S. Goldin-Meadow, Gesture for linguists: A handy primer, Language and Linguistics Compass 9 (2015) 437–451. doi:10.1111/lnc3.12168.

[4] Á. Abuczki, E. B. Ghazaleh, An overview of multimodal corpora, annotation tools and schemes, Argumentum 9 (2013) 86–98.

[5] M. E. Foster, J. Oberlander, Corpus-based generation of head and eyebrow motion for an embodied conversational agent, Language Resources and Evaluation 41 (2007). doi:10.1007/s10579-007-9055-3.

[6] J. Bressem, S. H. Ladewig, C. Müller, 71. Linguistic Annotation System for Gestures, De Gruyter Mouton, Berlin, Boston, 2013, pp. 1098–1124. URL: https://doi.org/10.1515/9783110261318.1098. doi:doi:10.1515/9783110261318.1098.

[7] C. S. Bianchini, (D)écrire les Langues des Signes: une approche grapholinguistique aux Langues des Signes, number 8 in Grapholinguistics and its Applications, Fluxus Editions, 2024. URL: https://hal.science/hal-04602726. doi:10.36824/2024-bianchini, iSSN 2681-8566 & eISSN 2534-5192; EAN 9782487055025; CrossRef 1612798840.

[8] F. Freigang, M. A. Priesters, R. Nishio, K. Bergmann, Your data at the center of attention: A metadata session profile for multimodal corpora, in: Proceedings of the CLARIN Annual Conference, volume 2014, 2014.

[9] A. Lücking, K. Bergmann, F. Hahn, S. Kopp, H. Rieser, The bielefeld speech and gesture alignment corpus (saga), in: LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality, 2010.

[10] J. Du Bois, G. Troiani, Typology and its data: functional monoculture or structural diversity?, presented at Naturally occurring data in and beyond linguistic typology, 2023.

[11] G. Troiani, Representing a language in use: corpus construction, prosody, and grammar in Kazakh, Ph.D. thesis, UC Santa Barbara, 2023.

[12] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, S. Rauzy, Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle, Revue TAL: traitement automatique des langues 49 (2008) pp–105.

[13] D. Knight, S. Adolphs, P. Tennent, R. Carter, The nottingham multi-modal corpus: A demonstration, in: Programme of the Workshop on Multimodal Corpora, 2009, p. 64.

[14] D. Trotta, A. Palmero Aprosio, S. Tonelli, A. Elia, Adding gesture, posture and facial displays to the polimodal corpus of political interviews, in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), European Language Resources Association, 2020, pp. 4320–4326.

[15] J. Allwood, Capturing differences between social activities in spoken language, Pragmatics and Beyond New Series (2001) 301–320.

[16] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, P. Paggio, The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena, Language Resources and Evaluation 41 (2007) 273–287.

[17] K. Pápay, S. Szeghalmy, I. Szekrényes, Hucomtech multimodal corpus annotation, Argumentum 7 (2011) 330–347.

[18] F. Schiel, S. Steininger, U. Türk, The smartkom multimodal corpus at bas., in: LREC, Citeseer, 2002.

[19] D. McNeill, Gesture and Thought, University of Chicago Press, 2013. doi:10.7208/chicago/9780226514642.001.0001.

[20] I. Mlakar, M. Rojc, Capturing form of non-verbal conversational behavior for recreation on synthetic conversational agent eva, WSEAS Trans. Comput.[Print ed.] 11 (2012) 218–226.

[21] I. Mlakar, D. Verdonik, S. Majhenič, M. Rojc, Towards pragmatic understanding of conversational intent: A multimodal annotation approach to multiparty informal interaction–the eva corpus, in: Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings 7, Springer, 2019, pp. 19–30.

[22] D. Fišer, J. Lenardič, Overview of multimodal corpora in the clarin (2020).

[23] B. MacWhinney, Computational transcript analysis and language disorders, in: Handbook of Neurolinguistics, Elsevier, 1998, pp. 599–616.

[24] L. Lo Re, Prosody and gestures to modelling multimodal interaction: Constructing an italian pilot corpus, IJCoL. Italian Journal of Computational Linguistics 7 (2021) 33–44.

[25] A. Kendon, et al., Gesticulation and speech: Two aspects of the process of utterance, The relationship of verbal and nonverbal communication 25 (1980) 207–227.

[26] L. Lo Re, Corpus multimodale dell'italiano parlato: basi metodologiche per la creazione di un prototipo, Ph.D. thesis, University of Firenze, 2022.

[27] K. Ackerley, F. Coccetta, Enriching language learning through a multimedia corpus, ReCALL 19 (2007) 351–370. doi:10.1017/S0958344007000730.

[28] F. Coccetta, et al., Multimodal functional-notional concordancing, New Trends in Corpora and Lan-

guage Learning. London: Continuum (2011) 121–138.

[29] P. L. Rohrer, I. Vilà-Giménez, J. Florit-Pons, N. Esteve-Gibert, A. Ren, S. Shattuck-Hufnagel, P. Prieto, The multimodal multidimensional (m3d) labelling scheme for the annotation of audiovisual corpora, Gesture and Speech in Interaction (GESPIN) (2020).

[30] P. L. Rohrer, E. Delais-Roussarie, P. Prieto, Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in english academic discourses, Lingua 293 (2023) 103583.

[31] L. Chevrefils, C. Danet, P. Doan, C. Thomas, M. Rébulard, C. Adrien, J.-F. Dauphin, C. S. Bianchini, The body between meaning and form: kinesiological analysis and typographical representation of movement in sign languages, Languages and Modalities 1 (2021) 49–63.

[32] D. Boutet, P. Doan, C. Danet, C. S. Bianchini, T. Goguely, A. Contesse, M. Rébulard, Systèmes graphématiques et écritures des langues signées, Signata. Annales des sémiotiques/Annals of Semiotics (2018) 391–426.

[33] M. W. Alibali, D. C. Heath, H. J. Myers, Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen, Journal of Memory and Language 44 (2001) 169–188. doi:10.1006/JMLA.2000.2752.

[34] J. M. Iverson, S. Goldin-Meadow, Why people gesture when they speak, Nature 1998 396:6708 396 (1998) 228–228. URL: https://www.nature.com/articles/24300. doi:10.1038/24300.

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2023.

[36] C. Mauri, S. Ballarè, E. Goria, M. Cerruti, F. Suriano, et al., Kiparla corpus: a new resource for spoken italian, in: CEUR WORKSHOP PROCEEDINGS, SunSITE Central Europe, 2019, pp. 1–7.

[37] H. Sloetjes, P. Wittenburg, Annotation by category-elan and iso dcr, in: 6th international Conference on Language Resources and Evaluation (LREC 2008), 2008.

[38] G. Jefferson, et al., Glossary of transcript symbols with an introduction, Conversation analysis (2004) 13–31.

[39] S. Slembrouck, Transcription—the extended directions of data histories: a response to m. bucholtz's' variation in transcription', Discourse Studies 9 (2007) 822–827.

[40] S. Ballarè, E. Goria, C. Mauri, Italiano parlato e variazione linguistica: Teoria e prassi nella costruzione del corpus KIParla, Pàtron editore, 2022.

A. Prompts for conversation

1. Sai che a Bologna c'è questa storia dell'"umarel"? Sai cos'è un umarel?

2. Sai quante lingue si insegnano al LILEC, il dipartimento di lingue dell'università di Bologna? Sapresti elencarle?

3. C'è una lingua che hai sempre voluto imparare? E una che invece proprio non ti ha mai incuriosito?

4. Secondo te quante sono le lingue parlate nel mondo?

5. Alcune espressioni sono veramente curiose: per esempio, hai mai pensato come mai la "zuppa inglese" si chiama così?

6. Parli un dialetto? Con chi lo parli? Quando lo parli?

7. Alcune espressioni sono veramente curiose: per esempio, hai mai pensato come mai si dice "fumare come un turco"?

8. I tortellini bolognesi: ti piacciono o no? Ma perché costano così tanto?

9. Qual è un piatto della tua infanzia che ricordi sempre con piacere?

10. Quale piatto cucini più spesso? Come lo prepari? Che ingredienti usi?

11. In che zona di Bologna vivi? Ti piace? Perché?

12. Secondo te, possono esistere lingue con massimo due parole per indicare i colori?

13. Che differenza c'è tra un dialetto e una lingua?

14. Ma perché si dice "chi va a Roma perde la poltrona"?

15. Credi che Bologna sia una città sicura dove vivere? Quali sono i suoi pro e i suoi contro?

16. In Italia il dialetto è un vero e proprio simbolo identitario. E tu che rapporto hai con il dialetto? Lo parli spesso? E con chi?

17. Qual è viaggio ti ha lasciato il ricordo più bello?

18. Chi è il tuo/la tua cantante preferito/a? Hai mai avuto modo di assistere a un suo concerto?

19. Secondo te esistono lingue più facili o più difficili da imparare, che per te suonano meglio o peggio? Quali e perché?

20. Hai qualche sogno o obiettivo che stai cercando di realizzare?

21. Se potessi vivere in un'altra città, quale sarebbe e perché?

22. C'è una lingua che avresti sempre voluto imparare, ma non hai mai studiato? Cosa ti attrae di questa lingua?

23. Credi che l'apprendimento di una nuova lingua possa influenzare il modo in cui vedi il mondo? In che modo?

24. Hai mai avuto difficoltà a comprendere gli accenti regionali o le varietà linguistiche?

25. Qual è il tuo modo preferito per rilassarti dopo una lunga giornata?
26. Se dovessi spiegare un modo di dire italiano a qualcuno che non lo conosce, quale sceglieresti e come lo spiegheresti?
27. Qual è il modo di dire italiano che trovi particolarmente divertente o curioso?
28. Cosa pensi del dibattito sull'influenza dell'inglese sull'italiano contemporaneo? È una minaccia o un arricchimento?
29. La lingua italiana è considerata una delle più musicali al mondo. Secondo te è vero? Quali sono, secondo te, altre lingue particolarmente musicali? E quali invece non lo sono affatto?
30. Hai mai avuto l'occasione di assaggiare la cucina tipica di un'altra nazione? Quale piatto ti è piaciuto in particolar modo e quale invece non ti ha convinto a pieno?

B. Metadata schemata

For both participants (see Subsection B.1) and conversations Subsection B.2), metadata is collected and maintained in `.yaml` files, with the following formats

B.1. Participants metadata

```
Code: # 4-char string composed by
    either S (Sighted) or B (
    Blind) and an integer padded
    with 0s

Gender: # either F (Female) or M
    (Male)

Age: # age range of the
    participant expressed as 5-
    years bins (0-5, 6-10,
    11-20,...)

Region: # 1 of the 20 italian
    regions (typing conventions
    provided)

First language: # upper cased iso
    -693-3 code of mother tongue

Education level: # one value in (
    Primaria, Medie inferiori,
    Medie superiori, Laurea, PhD)
```

```
Profession: # istat-derived
    category for profession (list
    provided)

Notes on sight-related
    disabilities: # any relevant
    annotation on sight-related
    conditions declared by the
    participant
```

B.2. Conversation metadata

```
Code: # 11-char string composed
    by [D|S] (same-condition or
    different-condition
    participant) + [M|U] (masked
    or unmasked conversation) + [L
    |S] (code associated to room
    where the conversation was
    recorded) + [DDMMhhmm]

Participants:
    - [participant_code_1] # code
        of participant sitting on
        left side
    - [participant_code_2] # code
        of participant sitting on
        right side

Facing: # M (Masked) or U (
    unmasked) depending on type of
    conversation

Data:
    - Video:
        - Left: path/to/left/camera/
            recording
        - Centre: path/to/central/
            camera/recording
        - Right: path/to/right/camera
            /recording
    - Audio: path/to/audio/file
    - Transcription:
        - Automatic: path/to/
            automatic/transcription
        - Manually revised: path/to/
            manually/revised/
            transcription
        - Prosodic: path/to/prosodic/
            transcription
        - Gestual: path/to/gestual/
            transcription
```
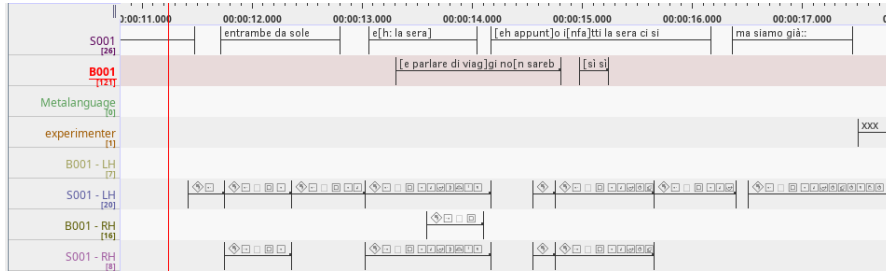
Figure 3: ELAN tiers showing verbal and gestural transcriptions.

C. Integrated transcription in ELAN

Figure 3 shows an example of the collaborative ELAN environment where transcriptions are developed. The picture shows 8 tiers: the first two (S001 and B001) refer to the verbal-prosodic transcription; an additional experimenter tier is used to take care of verbal productions of the experimenter, in case they occurr; the *metalanguage* tier encodes non-verbal acts such laughs, noises etc.; the remaining tiers encode the Typannot-based transcriptions for finger articulators (F:LH and F:RH stand for *left hand* and *right hand* respectively).

The full CoNLL-U data can be consulted at https://github.com/LaboratorioSperimentale/gest-IT/blob/dev/data/conll/DUC22051430.annotated.conll, a small portion is reported in Figure 4.

Figure 4: CoNLL-U extract

```
# sent_id = tu0005
# overlaps = gu0003 gu0004 gu0005 gu0006
# conversation = DUC22051430
# speaker_id = S001
# duration = 1.088
# text_jefferson = entrambe da sole
# text = entrambe da sole
1   entrambe  entrambi  PRON  _   Gender=Fem|Number=Plur|PronType=Ind   0   root   _   AlignBegin=11.704
2   da   da   ADP   _   _   3   case   _   _
3   sole   solo   ADJ   _   Gender=Fem|Number=Plur   1   nmod   _   AlignEnd=12.792

# sent_id = tu0006
# overlaps = tu0007 gu0007 gu0008 gu0009
# conversation = DUC22051430
# speaker_id = S001
# duration = 0.987
# text_jefferson = e[h: la sera]
# text = eh la sera
1   eh   eh   INTJ   _   _   3   discourse   _   AlignBegin=13.047|Overlap=B:tu0007|ProlongedSound=eh:
2   la   il   DET   _   Definite=Def|Gender=Fem|Number=Sing   3   det   _   Overlap=I
3   sera   sera   NOUN   _   Gender=Fem|Number=Sing   0   root   _   AlignEnd=14.034|Overlap=I

# sent_id = tu0007
# overlaps = tu0006 tu0008 gu0007 gu0008 gu0009 gu0010 gu0011 gu0012 gu0013
# conversation = DUC22051430
# speaker_id = B001
# duration = 1.500
# text_jefferson = [e parlare di viag]gi no[n sarebbe male]
# text = e parlare di viaggi non sarebbe male
1   e   e   CCONJ   _   _   7   cc   _   AlignBegin=13.3|Overlap=B:tu0006
2   parlare   parlare   VERB   _   VerbForm=Inf   7   csubj   _   Overlap=I
3   di   di   ADP   _   _   4   case   _   Overlap=I
4   viaggi   viaggio   NOUN   _   Gender=Masc|Number=Plur   2   obl   _   Overlap=I
5   non   non   ADV   _   _   7   advmod   _   Overlap=B:tu0008
6   sarebbe   essere   AUX   _   Mood=Cnd|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   7   cop   _   Overlap=I
7   male   male   ADV   _   _   0   root   _   AlignEnd=14.8|Overlap=I

# conversation = DUC22051430
# sent_id = gu0003
# overlaps = tu0004 tu0005
# speaker = S001
# duration = 0.330
# text = EMPTY
# type = F:LH
1   EMPTY   EMPTY   X   _   _   0   root   _   AlignBegin=11.410|AlignEnd=11.740|gesture='\ue5de\ue002 [ \uf197 \ue008\uf19f\
    ue5ea\ue5ef\ue5e8\ue5ef\uf1a0\ue5fe\ue5ee - \ue004\ue005\ue006\ue007\uf19f\ue5fe\ue5ee\ue5e8\ue5ef\uf1a0\ue5e7
    \ue5ef ] [ \uf198\ue001 ]'

# conversation = DUC22051430
# sent_id = gu0004
# overlaps = tu0005 gu0005
# speaker = S001
# duration = 0.610
# text = EMPTY
# type = F:LH
1   EMPTY   EMPTY   X   _   _   0   root   _   AlignBegin=11.740|AlignEnd=12.350|gesture='\ue5de\ue002 [ \uf197 \ue008\uf19f\
    ue5ff\ue5ee\ue5fb\ue5ee\uf1a0\ue5fe\ue5ee - \ue004\ue005\ue006\ue007\uf19f\ue5fe\ue5ee\ue5fb\ue5ee\uf1a0\ue5fd
    \ue5ee ] [ \uf198\ue001 ]'

# conversation = DUC22051430
# sent_id = gu0005
# overlaps = tu0005 gu0004
# speaker = S001
# duration = 0.610
# text = EMPTY
# type = F:RH
1   EMPTY   EMPTY   X   _   _   0   root   _   AlignBegin=11.740|AlignEnd=12.350|gesture='\ue5de\ue003 [ \uf197 \ue008\uf19f\
    ue5ff\ue5ee\ue5fb\ue5ee\uf1a0\ue5fe\ue5ee - \ue004\ue005\ue006\ue007\uf19f\ue5fe\ue5ee\ue5fb\ue5ee\uf1a0\ue5fd
    \ue5ee ] [ \uf198\ue001 ]'

# conversation = DUC22051430
# sent_id = gu0006
# overlaps = tu0005
# speaker = S001
# duration = 0.670
# text = EMPTY
# type = F:LH
1   EMPTY   EMPTY   X   _   _   0   root   _   AlignBegin=12.350|AlignEnd=13.020|gesture='\ue5de\ue002 [ \uf197 \ue008\uf19f\
    ue5ea\ue5ef\ue5fb\ue5ee\uf1a0\ue5fe\ue5ee - \ue004\uf19f\ue5e7\ue5ef\ue5fb\ue5ee\uf1a0\ue5fd\ue5ee - \ue005\
    ue006\ue007\uf19f\ue5fe\ue5ee\ue5fb\ue5ee\uf1a0\ue5fd\ue5ee ] [ \uf198 ( \ue00e\ue001 ) ( \ue00b \ue00c\ue005\
    ue006\ue007 - \uf196\ue5ee\ue001 ) ]'
```