CHROME Corpus
Multi-level annotation of an audiovisual resource[1]

Iolanda Alfano°, Violetta Cataldo°,*, Loredana Schettino°,^

°Università degli Studi di Salerno, *Ghent University, ^Università degli Studi di Napoli Federico II

In this paper we describe the methodology employed for the annotation of a resource developed within the CHROME (Cultural Heritage Resources Orienting Multimodal Experience) project, aimed at the protection and promotion of cultural heritage. More specifically, the ultimate goal of the project is the modelling of multimodal data (including speech features and gestures) for the design of a virtual agent serving in museums and capable of communicating in intelligible as well as effective and natural way. In order to grasp the relationship between linguistic and gestural behaviours, multi-level annotation systems have been developed and implemented for the labelling of linguistic and gesture features on different levels of analysis. This article is dedicated to a general presentation of the corpus and to the description of the different levels of linguistic annotation; then, the final section, reports conclusive remarks considering the applications of the described methodology. The CHROME corpus and the mark-up methodology described in this work represent valuable multimodal resources for investigations on communicative dynamics which may offer valid support for both theoretical and practical applications.

Keywords: corpus collection; guided tours; linguistic annotation; multimodal annotation

[1] This paper is the result of the collaboration between the authors. Nevertheless, for academic purposes only, we attribute to Iolanda Alfano §§ 3.1.1, 3.1.2 and 3.1.3; to Violetta Cataldo §§ 2 and 3; to Loredana Schettino §§ 1 and 3.1.4. The authors have joint responsibility for § 4.

CHIMERA. Romance Corpora and Linguistic Studies 10 (2023), 135-153. ISSN 2386-2629

1. Introduction

Human communication may be described as a complex activity whereby human beings can transmit messages using multiple channels, such as the phonic-auditory channel for speech signs, or the gestural-visual channel for gesture signs, or the graphic-visual channel for written communication (Voghera 2017). The specific conditions of production linked to the different channels excert a strong influence on linguistic uses and correlate with the characteristics of different modes of communication, for example written and spoken language preferably rely on different uses of linguistic structures.
Moreover, in spontaneous contexts, spoken communication can rely on the cooperative use of different channels. In particular, acoustic and visual information may be integrated in that gestures (the so-called "co-speech gestures") accompany speech to define, complete or reinforce the meaning of spoken utterances (see Campisi 2018). Hence, analysing the ways in which speech and gesture may be combined and contribute to the communicative effort is fundamental for an encompassing understanding of human behaviours in spoken communication.

However, speech and gesture signs share the basic features of being temporally defined, transient objects, inextricably tied to the moment of production, which has made difficult to conduct scientific analyses on spoken language, co-speech gesture and their relationship in a systematic way until suitable recording tools became available (Voghera 2017). Thereafter, further technological advancements have prompted the development of *Corpus Linguistics* which has supported the empirical study of attested language uses based on the collection and annotation of data (McEnery & Gabrielatos 2006). In particular, the operation of data annotation mainly consists of providing information on linguistic phenomena by labelling items according to theoretically defined categories. Moreover, in the case of speech data, the annotation may be aligned to the signal and multiple annotation layers may be considered in order to account for simultaneous phenomena and their characteristics on different levels of analysis.

So, the last few decades have seen a growing interest for collecting corpora of audiovisual recordings of communicative exchanges and defining reliable protocols for the time-aligned multi-level annotation of the temporally-determined and multifaceted phenomena that characterize speech and gesture. Main purposes of this endeavour are the conduction of more complete studies of the construction of meaning and understanding in face-to-face interaction and the implementation of these observations in technological applications such as the development of

embodied conversational agents or avatars that are able to reproduce human-like communicative behaviours (Allwood 2008).

In this paper, we describe the methodology employed for the annotation of a resource developed within the Italian national CHROME (Cultural Heritage Resources Orienting Multimodal Experience) project (Origlia *et al*. 2018)[2], generally aimed at the protection and promotion of cultural heritage. More specifically, the ultimate goal of the project is the modelling of multimodal data (including speech features and gestures) for the design of a virtual agent serving in museums and capable of communicating in intelligible as well as effective and natural ways. To obtain insights on human linguistic and gestural behaviours and their relationship, multi-level annotation systems have been developed and implemented for the independent labelling of linguistic and gestural features on different levels of analysis.

The article is structured as follows: section 2 is dedicated to a general presentation of the CHROME corpus; section 3 delves into the description of the different levels of linguistic annotation; then, the final section, reports conclusive remarks considering the applications of the described methodology.

2. Data collection

The multimodal CHROME resource involves two corpora: a written corpus (§ 2.1) and an oral corpus (§ 2.2).

Both written and oral texts were selected and collected for their relevance in describing and promoting the cultural sites represented by the charterhouses present in Campania, which are the core of the cultural heritage selected for the CHROME project[3].

2.1 Written corpus

The written corpus is composed of written texts belonging to six main textual types distinguished by different degrees of complexity identified according to the target audience. Specifically, the audience types are placed along a continuum ranging from experts, specialists in the field of cultural heritage, to non-experts, hence a more heterogeneous and non-selected audience of tourists.

In an expert-to-tourist order, here the textual types are listed:

[2] http://www.chrome.unina.it.

[3] Upon completion of the project, the dataset will be made available for the scientific community (https://live.european-language-grid.eu/).

- Scientific essays (20) – specialists, scholars;
- Specialistic catalogues (3) – specialists, researchers, specialized audience;
- Informative catalogues (5) – specialized audience;
- Specialistic travel guides (15) – specialized/interested audience;
- Brochure and web pages (47) – non-selected audience;
- Explicative/explanatory kits (4) – non-selected audience, tourists.

We collected 94 texts (about 15664 lemmas and 271930 word tokens), all in Italian, dealing with the three cultural sites of the Campanian Charterhouses of San Martino in Naples, San Giacomo in Capri and San Lorenzo in Padula[4].

2.2 Oral corpus

The oral corpus corresponds to about 11 hours of speech. Like the written texts, issues of charterhouses' cultural contents are addressed in the oral texts as well.

This corpus consists of audiovisual recordings of tourist visits led in Italian in the cultural site of San Martino Charterhouse in Naples. Specifically, we collected recordings of three female expert guides[5]. Each guide conducted four visits lasting about an hour each, with a small group of tourists (four people) for every visit. Audiovisual recordings were collected separately for both the guide and the audience, who knew that the visit was recorded but were not aware of the scopes of the research. Hence, a fixed shot of the guide and another of the visitors was used in order to obtain two Full-HD video recordings; moreover, the guide's voice was recorded using a close range digital microphone. The syinchronization of videos and audio was carried out through a visual-acoustic marker (Origlia *et al.* 2018).

The guided tours followed a route consisting of six main points of interest (POIs) of the Charterhouse, identified on the basis of their historical-artistic value: 1) *Pronaos*; 2) *Great cloister*; 3) *Parlor*; 4) *Chapter hall*; 5) *Wooden choir*; 6) *Treasure hall*.

Visitors were free to ask any question any time. However, the collected speech of tourist guides can not be defined as completely pre-formulated and learned-by-heart productions, but rather approximate semi-spontaneous and semi-

[4] The written corpus collects the available texts on themes that are relevant for the project, i.e. descriptions of the three Charterhouses. Accordingly, it is not balanced across the six different textual types. However, as we state throughout the paper, the core of the CHROME corpus is the oral corpus, which has been indeed employed for the linguistic analyses reported in § 3.

[5] Potential variation due to the gender factors falls outside of the scopes of this research.

monological speech (Cataldo *et al.* 2019), as they are characterized by high degree of discourse planning, high selective attention of the audience, quite low degree of interlocutors' explicit dialogic interaction and close integration between verbal and non-verbal elements (Voghera 2017).

3. Data annotation

The written and the oral corpora were analyzed and annotated on a number of linguistic levels. In the light of the main goals of the CHROME project, greater attention was devoted to the analysis of the oral corpus in order to investigate guides' multimodal communication. Accordingly, in this paper we focus on the description of linguistic analysis and annotaton of speech data.

Multimodal and multi-layered annotations were performed in *Praat* (Boersma & Weenink 2019) and *ELAN* (Sloetjes & Wittenburg 2008) with the aim of investigating the guides' linguistic and gestural behavior and the benefit of allowing cross-domain research (Origlia *et al*. 2018).

In a preliminary phase, an orthographic transcription of the collected guides' speech was carried out, following the guidelines provided in Savy (2005a) for the orthographic transcription of oral texts. Hence, such a transcription takes into account all lexical elements as well as long and short silent and filled pauses, false starts, non-verbal elements, such as laughts and coughts.

Then, the orthografic transcription was time-aligned to the signal using *Webmaus* (Kisler *et al*. 2017). The speech-text alignment was carried out in *Praat* (Boersma & Weenink 2019). As a result, Praat textgrids with separate annotation levels were obtained, with the levels being:

- *Orthographic level*, segmented in lexical elements, fragments of words, pauses, non-verbal elements;
- *Phonetic level*, segmented in phones, silences, non-verbal vocalizations;
- *Syllabic level*, segmented in phonetic syllables, silences, non-verbal vocalizations.

The semi-automatic alignment was then manually checked. In particular, for the phonetic and syllabic levels we followed the indications reported in Savy (2005b) in order to consider also coarticulation phenomena. These three levels of analysis served as a basis for subsequent analyses and annotations described below.

3.1 Linguistic annotation

The linguistic annotation concerns different levels of analysis, to which we devote the following sections. As we will see, the units of analysis of each level were defined according to principles referring to that specific level. Moreover, the annotation of each level was executed separately, in order to keep the annotations as independent as possible from one another. The output can provide new data on the mapping between prosodic structure (§ 3.1.1), information structure (§ 3.1.2), and syntactic structure (§ 3.1.3). Moreover, disfluency phenomena are considered too (§ 3.1.4).

3.1.1 *Intonation*

On the first *ELAN* level (Sloetjes & Wittenburg 2008), called *Demarcative*, a speech segmentation in major prosodic units is provided.

The first step of the annotation was to decide the criteria that were relevant for the segmentation of the prosodic flow. The basic unit of reference in the intonation level, variously referred to as intonation unit, prosodic phrase, tone unit (TU), etc., is generally defined most saliently by a single and coherent intonational contour. There has been extensive work on TUs in several languages, combining both phonological and phonetic approaches (Savy 1999: 172-178). Phonological approaches claim that prosodic phrasing mainly depends on syntactic, and hence metric, structure[6], while phonetic approaches try to segment the speech flow using acoustic correlates of boundaries, relying on a combination of cues (Albano Leoni & Maturi 2002: 125-128). As is now widely known, TUs can show tremendous variety of intonation patterns, and high variability across languages, speakers, speech style, communicative situation and so on. The demarcation of TUs is not always an easy task, since there can be a certain degree of ambiguity and a disagreement as to what criteria are relevant for this phrasing. One runs the risk of relying on syntactic, semantic or pragmatic criteria.

Our phrasing was based on phonetic criteria, by combing both perceptive and acoustic analysis. A major TU was isolated when a number of phonetic boundary markers co-occurred, on the basis of the acoustic signal alone, i.e. presence of a (potential) final pause; f_0 declination of both f_0 and energy; parametrical reset at the beginning of a new TU; prepausal lengthening. Following Degand & Simon (2009), which consider a three-level prosodic segmentation procedure for major,

[6] Whether, and to what extent, prosodic constituents are isomorphic with syntactic ones is a central research question, to which theories of syntactic structure and prosodic structure give different answers. For an overview of theoretical advances in research on the syntax-prosody interface, see Elfner (2018).

intermediate, and minor units, we separated major units. We segmented following perceptive criteria and, when in doubt, we moved from *ELAN* to *Praat* and followed manually the rules identified to split in major TUs: i) Do not assign a boundary to a final syllable when the syllable marks a hesitation. ii) Assign a major boundary to a final syllable: when syllable duration prominence > 3 (i.e. 3 times as long as the context mean); or, when this syllable is followed by a pause ≥ 200 ms; or, when the intra-syllabic pitch rise ≥ 4 semitones (ST) and the syllable mean pitch prominence ≥ 5 ST (i.e. 5 ST higher than the context mean).

Often major TUs are characterized by an initial f_0 reset, followed by a declination, an overall fall in f_0 (and in intensity), and a variety of final contours. These units are often delimited by real or potential silent pauses.

In a second step, tonal events were semi-automatically labelled following the INTSINT system (Hirst & Di Cristo 1998), but not using the MOMEL algorithm. In our analysis, we used the *Prosomarker* program (see Origlia & Alfano 2012; Alfano 2019), which contains a stylization algorithm and an annotation module based on the INTSINT system.

Thus, we used the following tag-set:

– T (*Top*), the point corresponding to the highest value of f_0;
– M (*Mid*), the initial point in the TU;
– B (*Bottom*), the point corresponding to the lowest value of f_0;
– U (*Up*), the point in a rising sequence or peak;
– D (*Down*), the point in a falling sequence or valley;
– H (*Higher*), peak;
– L (*Lower*), valley;
– S (*Same*), the point with the same value as the preceding target point.

As far as the threshold to determine the tags of relative tones is concerned, the *Prosomarker* algorithm works as follows. It assigns the tag:

– S, if the difference between the target point indicated by the tag and the previous target point is lower than 1.5 ST;

– D or U, if the difference between the target point indicated by the tag and the previous target point is higher than 1.5 ST, but lower than 3 ST;

– L or H, if the difference between the target point indicated by the tag and the previous target point is higher than 3 ST;

– D, if the value of the target point indicated by the tag is lower than the value of the previous target point, but higher than the value of the following target point;

– U, if the value of the target point indicated by the tag is higher than the value of the previous target point, but lower than the value of the following target point.

Finally, we introduce a new element in the annotation by treating the absolute points T and B as relative points too, which indicates that the relative tone between

parentheses corresponds to the absolute tone. Indeed, a rise that culminates in the maximum peak of the TU may be more or less steep and, as such, this difference is expressed by the tag used in the annotation corresponding to T(H) or T(U), respectively. Figure 1 displays an example of a TU, corresponding to *La certosa di san Martino* in *La Certosa di san Martino, qui a Napoli, ha almeno due anime,* 'San Martino Charterhouse, here in Naples, has two souls at least'. These TUs are delimited by silent pauses and are characterized by an initial f_0 reset.

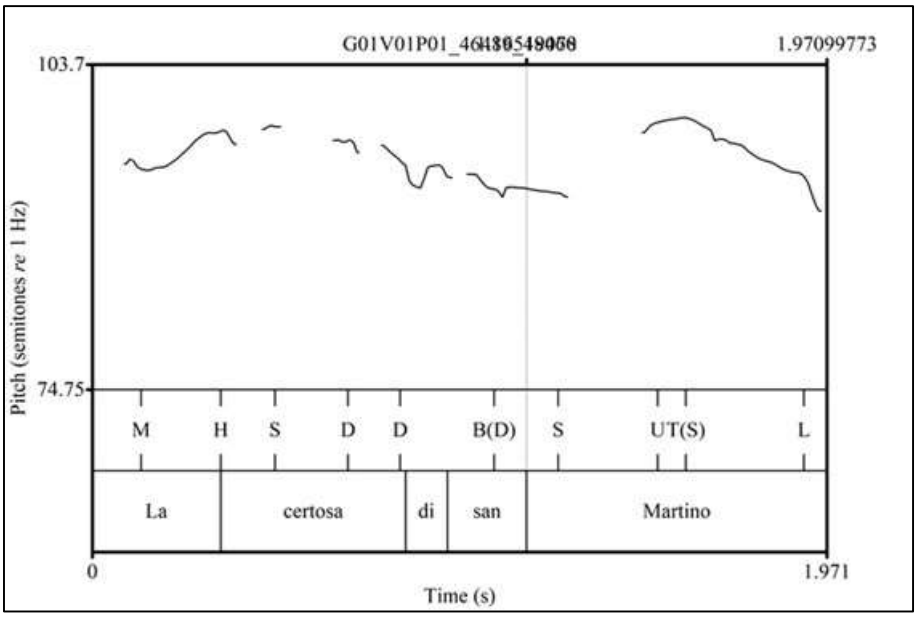

Figure 1. An example of TU, *La Certosa di san Martino* 'San Martino Charterhouse' [G01V01P01].

3.1.2 *Information structure*

On the second *ELAN* level, called *Informative*, a segmentation in informative units (IUs) is provided.

In this level we distinguished three major IUs: Topic (T), Comment (C) and insertions (INS), made by parenthetical elements.

As far as the T is concerned, we considered it as what the sentence is about (Reinhart 1981, Gundel 1988, Lambrecht 1994, Krifka 2008), frequently identified as the single most salient given referent in an utterance. Following Gundel's topic and comment definition, "An entity, E, is the topic of a sentence, S, iff in using S the speaker intends to increase the addressee's knowledge about, request

information about, or otherwise get the addressee to act with respect to E. A predication, P, is the comment of a sentence, S, iff in using S the speaker intends P to be assessed relative to the topic of S." (Gundel 1988: 210).

Within the "Language into Act Theory" (Cresti & Moneglia 2018), topics are seen as units that develop the information function of field of application for the illocutionary force (expressed by the comment unit). Therefore, they do not convey the illocution of the utterance; they always precede the comment and can be identified in speech only considering their prosodic performance[7].

Despite the differences between the various theoretical proposals, we can state that they all consider the topic as the basis for what is said or the frame for the most relevant part of the message. For the identification of topics in this work, we assumed that sentence topics do not have to be referential, since they can also express situations or states of affairs, are optional, do not necessarily occur in a fixed position in the utterance and are not necessarily given. Finally, there can be several topics in one utterance.

As far as INSs are concerned, these units were identified following the functional criteria exposed in Firenzuoli & Tucci (2003). The following utterance displays an example for each unit:

(1) La Certosa (T) ha un'origine trecentesca (C), come vi dicevo (INS)
'The charterhouse (T) has 14th-century origins(C), as I told you (INS)'

3.1.3 *Syntax*

On the following *ELAN* levels, clause types (*Macrosyntactic* level), intraclausal syntactic functions (*Syntactic* level), types of phrases corresponding to the intraclausal nodes (*Grouping* level), and the composition of the phrases of the previous level (*Intrasyntactic* level) are tagged. Finally, on the *Syntactic Weight* level, for each phrase the "weight" is indicated.

On the *Macrosyntactic* level clauses are tagged according to the tag-set exposed in Table 1.

[7] The units in this framework have two types of function: textual, in the case that they participate in the composition of the utterance semantics, such as the Comment or the T, and dialogical, if they assist in the exchange with the addressee and signal that the speech flow will continue, i.e., for instance, *Incipit*, *Phatic* or *Allocutive*. We did not consider this kind of units for a matter of tag-set economy. Since our kind of speech is semi-monological, dialogic functions are very limited with respect to conversational speech.

Table 1. The *Macrosyntactic* level tag-set.

Clause	Tags
Independent Clause	IC
Finite Argument clause	FAC
Finite Circumstantial Clause	FCC
Finite Relative Clause	FRC
Nonfinite Argument Clause	nonFAC
Nonfinite Circumstantial Clause	nonFCC
Nonfinite Relative Clause	nonFRC
Coordinate Independent Clause	CoordIC
Coordinate Dependent Clause	CoordDC

On the S*yntactic* level, syntactic functions are tagged according to the tag-set in Table 2.

Table 2. The *Syntactic* level tag-set.

Function	Tags
Predicate	PRE
Subject	SUB
Direct Object	OBJ
IndirectObject	IO
Circumstantial elements	CE
Nominal Sentence	NS
Isolated elements	ISO

On the *Grouping* and the *Intrasyntactic* levels, the types of phrases corresponding to the intraclausal nodes and simple phrases are tagged according to the tag-set showed in Table 3.

Table 3. The *Grouping* and *Intrasyntactic* level tag-set.

Phrase	Tags
Noun Phrase	NP
Prepositional Phrase	PP
Verb Phrase	VP
Predicative Noun Phrase	PNP
Adverbial Phrase	ADVP

Finally, on the *Syntactic Weight* level, for each phrase the "weight" is calculated according to a scale that takes both phrases' structure and expansion into account. This scale considers several levels of weight according to the presence/absence of determiners (± det) and modifiers (± mod), whether the head is a noun or a pronoun (pro), and whether the verb is saturated or not (± sat). The scale originates from Voghera & Turco (2008); in addition, the '+' and '-' symbols following the phrase indicate a different degree of heaviness (in the sense of "segmental lightness or heaviness") and/or the number (one or more than one) of determiners or modifiers. An example for each case is provided in Table 4.

Table 4. The *Syntactic* weight level tag-set.

Phrase	**Tags**	**Presence of dets and mods/type of head**	**Example**
NP	NP1	[+ pro] [- det] [- mod]	*Egli* 'He'
	NP2	[- det] [- mod]	*Ragazzi* 'Guys'
	NP3-	[- det] [+ mod]	*Ragazzi simpatici* 'Nice guys'
	NP3+	[- det] [+ mod]	*Ragazzi molto simpatici* 'Very nice guys'
	NP4-	[+ det] [- mod]	*Un ragazzo* 'A guy'
	NP4+	[+ det] [- mod]	*Qualunque ragazzo* 'Any guy'
	NP5-	[+ det] [+ mod]	*Un ragazzo simpatico* 'A nice guy'
	NP5+	[+ det] [+ mod]	*Qualunque ragazzo molto simpatico* 'Any very nice guy'
PP	PP1-	[+pro] [- det] [- mod]	*Per me* 'For me'
	PP1+	[+pro] [- det] [- mod]	*Attraverso quello* 'Through that'
	PP2-	[- det] [- mod]	*Da profano* 'As a layman'
	PP2+	[- det] [- mod]	*Sotto terra* 'Below ground'

	PP3-	[- det] [+ mod]	*In casa mia* 'In my house'
	PP3+	[- det] [+ mod]	*In casi poco chiari* 'In unclear cases'
	PP4-	[+ det] [- mod]	*Nella vita* 'In the life'
	PP4+	[+ det] [- mod]	*In qualsiasi momento* 'At any time'
	PP5-	[+ det] [+ mod]	*Nell'ipotetico caso* 'In the hypothetical case'
	PP5+	[+ det] [+ mod]	*Mediante un espediente letterario* 'Through a literary device'
VP	VP1	[+ impersonal] [- sat] [- mod]	*Ha piovuto* 'It rained'
	VP2	[- sat] [- mod]	*Siamo arrivati* 'We have arrived'
	VP3-	[- sat] [+ mod]	*Andiamo piano* 'We go slowly'
	VP3+	[- sat] [+ mod]	*Andiamo molto lentamente* 'We go very slowly'
	VP4	[+ sat] [- mod]	Maria *ti ha visto* 'Maria saw yoy'
	VP5-	[+ sat] [+ mod]	Maria *gioca bene* 'Maria plays well'
	VP5+	[+ sat] [+ mod]	Maria *gioca veramente sempre bene* 'Maria always plays really well'
	VP6-	[servile or phraseological and causative verbs] [+ sat] [+ mod]	Noi *dobbiamo immaginare* la Napoli del secolo scorso 'We need to imagine Naples in the last century'
	VP6+	[servile or phraseological and causative verbs] [+ sat] [+ mod]	Se tutti noi *provassimo ad immaginare* la Napoli del secolo scorso 'If we all tried to imagine the Naples of the last century'
PNP	PNP1	[+ pro] [- det] [- mod]	È *quello* 'It is that one'

PNP2	[- det] [- mod]	È *perfetto* 'It is perfect'
PNP3-	[- det] [+ mod]	Sono *affari miei* 'It is my business'
PNP3+	[- det] [+ mod]	È *particolarmente simpatico* '(S)he is really nice'
PNP4-	[+ det] [- mod]	È *una meraviglia* 'It is wonderful'
PNP4+	[+ det] [- mod]	È *altrettanto simpatico* '(S)he is just as nice'
PNP5-	[+ det] [+ mod]	È *un bel ragazzo* 'He is a good-looking guy'
PNP5+	[+ det] [+ mod]	È *un mio carissimo amico* 'He is a very dear friend of mine'

Figure 2 provides an example of the annotation in *ELAN* of the syntactic levels, in addition to the first two levels corresponding to the annotations of TUs (*Demarcative*) and IUs (*Informative*).

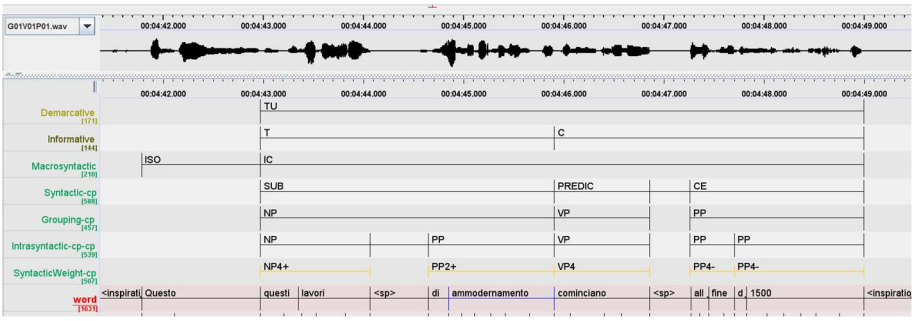

Figure 2. An example of syntactic levels annotation, *Questo questi lavori di ammodernamento cominciano alla fine del Millecinquecento* 'This These renovation works began at the end of the 16th century' [G01V01P01].

3.1.4 *Disfluencies*

Multiple *ELAN* levels are employed for the annotation of disfluency phenomena as well. These are defined as phenomena produced by speakers in order to efficiently manage and control their speech either by editing already uttered segments or temporarily suspending speech delivery due to planning demands. The annotation scheme (described in Schettino 2022) is a revised version of the one tested in

a previous pilot study (Cataldo *et al*. 2019) and consists of different levels so as to account for both disfluencies' formal structures and their functions in the context of occurrence.

The *Disfluency Model (dml)* level concerns the labelling of the macro-structure of disfluencies (see Shriberg 1994). Namely, the region to be repaired (*Reparandum*), the repaired one (*Reparans*), repaired regions that are to be further repaired (*Chained Repair*), the region in which the delay occurs (*Interregnum*), the ones that precede (*Original Utterance*) and follow (*Continuation*) a delay (see Table 5).

Table 5. The *Disfluency Model* level tag-set.

Macro-structure	**Tags**
Reparandum	RM
Interregnum	IM
Reparans	RS
Chained Repair	RS_RM
Original utterance	OU
Continuation	CNT

The *Disfluency Structure (dstr)* level serves for labeling the micro-structure embodying the disfluency. Here, disfluent items are categorized as Insertion, Deletion, Substitution, Repetition, Silent Pause, Lengthening, Filled Pause, Lexicalized Filled Pause (Eklund 2004; see Table 6). In particular, not every instance of silence, segmental prolongation and pragmatic markers are identified and tagged as, respectively, Silent Pause, Lengthening and Lexicalized Filled Pause, but only those that in the given context can be identified as disfluency phenomena according to the provided definition, in this case, covering for speech planning time.

Table 6. The *Disfluency Structure* level tag-set.

Structure	**Tags**
Deletion	DEL
Substitution	SUB
Insertion	INS
Repetition	REP
Silent Pause	SP
Filled Pause	FP
Lexicalized Filled Pause	LFP
Lengthening	LEN

On the *Disfluency Function (dfn)* level, each item is assigned its macro-function, Backward-Looking or Forward-Looking (Ginzburg *et al*. 2014; see Table 7).

Table 7. The *Disfluency Function* level tag-set.

Function	Tags
Backward Looking Disfluency	BLD
Forward Looking Disfluency	FLD

Finally, on the *Hesitation Function (hfn)*, Forward-Looking items – also defined in literature as *hesitations* marking a temporary delay in speech – are associated with more specific functions regarding their co-text (see Table 8). Namely, *Word Searching*, when disfluencies are involved in lexical retrieval or lexical selection purposes (Tottie 2020); *Structuring*, for disfluencies occurring at the boundaries of syntactic or information structure, e.g., clauses and topic-comment, respectively; *Focusing*, associated to disfluencies marking upcoming "semantically heavy concepts or words" (Kjellmer 2003)[8]; *Hesitative*, for disfluencies' occurrences that play none of the preceding sub-functions and are triggered by only broad speech planning.

Given that hesitation phenomena may carry out more than one function, on the fourth functional level, categories exposed in Table 8 are not mutually exclusive, unlike the categories considered for the first three levels.

Table 8. The *Hesitation Function* level tag-set.

Function	Tags
Word Searching	WS
Structuring	STR
Focusing	FOC
Hesitative	HES

Figure 3 provides an example of the multi-level disfluency annotation in ELAN.

[8] Note that this label was not assigned to phenomena identified as signal of properly focalized elements, but rather to items that are involved in the planning and production of following key information, e.g., new or emphasized elements, independently from syntactic structures.

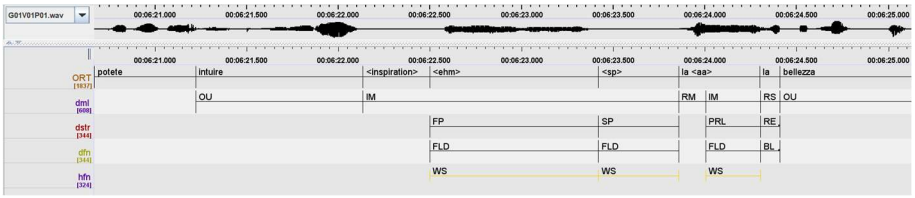

Figure 3. An example of disfluency levels annotation: *potete intuire <inspiration> <ehm> <sp> la <aa> la bellezza,* 'You can grasp <*inspiration*> <*ehm*> <*sp*> the<ee> the beauty' [G01V01P01].

4. Conclusions

The article presents the CHROME Corpus as a multimodal resource consisting in written texts and audio-video recordings of spoken productions that can be considered as a basis for investigations on communicative behaviours in the domain of cultural heritage popularization.

More specifically, the core of this work is the description of the collected data and the methodological principles defined and adopted for the coding of linguistic information on different levels of analysis.

The resource has been developed within the context of the CHROME project which concentrated on observing the communication processes between tourist guides and visitors of the Neapolitan San Martino Charterhouse to support the development of interactive technologies, such as speaking Virtual Agents, for cultural heritage. Hence, major interest was devoted to modelling linguistic features that characterize spoken language.

Different levels of analysis were considered, i.e., prosodic, syntactic and pragmatic levels, and for each individual layer, a specific annotation scheme reflecting the theoretical model of reference was defined. Multilevel labelling systems based on the principle of independence, i.e. with each layer relying on specific criteria which are independent from each other and from the data, allow for theoretically coherent analyses and reduce the risk of circularity in studies on speech phenomena, either concerning single levels or mapping information on different ones. Indeed, the relevance of this method lies in allowing both the consideration of individual levels and the study of correlations between levels related to different domains.

The described annotation schemes proved useful for modelling communicative behaviours on the basis of observations of patterns of specific linguistic structures or phenomena in speech, such as topic units or speech

disfluencies. One example is provided by the study of the phonetic realization of sentence topics as a function of syntactic and textual-pragmatic features which was conducted mapping information on the intonation, informative, syntactic and disfluency levels of annotation (Alfano *et al.* 2021). Other examples come from the number of studies conducted on disfluency phenomena, which concerned the analysis of how the phonetic characteristics of specific phenomena (silent pauses, prolongations, lexical and non-verbal fillers) could vary as a function of structural and functional features (Cataldo *et al*. 2019; Schettino & Cataldo 2019; Schettino *et al.* 2021a; Schettino *et al.* 2021b; Schettino 2022).

Beyond the labelling of information on speech elements, further level of annotation could be introduced to account for co-speech gestures (see Campisi 2018) or structures in written texts in order to conduct analyses in multimodal perspectives. For example, insights on multimodal communicative dynamics in the interaction between tourist guides and visitors could be obtained mapping the annotation on the speech disfluency levels onto the ones related to hand gestures, which yielded characteristic patterns of co-occurrence between disfluency phenomena and gestures (Cataldo *et al.* 2019; Origlia *et al.* 2019; Chiera *et al.* 2023). Moreover, the corpus includes both written and oral texts concerning the same domain and could provide a valuable resouce for comparing the uses of specific linguistic structures in written and spoken language.

To conclude, the CHROME corpus and the mark-up methodology described in this work represent multimodal resources for linguistic (and gestural) analyses supporting both theoretical and practical applications. On the one hand, the detailed annotation systems provide a valid support for investigations aimed at deepening our understanding of communicative dynamics; on the other hand, this insight may find application in speech technologies such as the development of (possibly) natural-sounding and efficient interactive Embodied Virtual Agents.

Acknowledgments

Work funded by the Italian PRIN project Cultural Heritage Resources Orienting Multimodal Experience (CHROME) #B52F15000450001.

References

Albano Leoni, F. & Maturi, P. 2002. *Manuale di fonetica*. Roma: Carocci.

Alfano, I. 2019. *Methodological and practical issues in studying intonation. The case of requests in Italian and Spanish task-oriented dialogues.* Studi AISV 5. Milano: Officinaventuno.

Alfano, I., Cataldo, V. Orrico, R. & Schettino, L. 2021. Sentence topics in Italian: An analysis on the CHROME Corpus. *Loquens* 8(1-2): e083.

Allwood, J. 2008. Multimodal Corpora. In A. Lüdeling & M. Kytö (eds), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 207-225.

Boersma, P. & Weenink. 2019. Praat: doing phonetics by computer [Computer program]. Version 6.1.08. http://www.praat.org/ (accessed December 5, 2019)

Campisi, E., 2018. *Che cos'è la gestualità*. Roma: Carocci.

Chiera, A., Ansani, A., Sessa, I., Cataldo, V., Schettino, L. & Poggi, I. 2023. Gestures and pauses to help thought: hands, voice, and silence in the tourist guide's speech. *Cognitive Processing* 24: 25-41.

Cataldo, V., Schettino, L., Savy, R., Poggi, I., Origlia, A., Ansani, A., Sessa, I. & Chiera, A. 2019. Phonetic and functional features of pauses, and concurrent gestures, in tourist guides' speech. In D. Piccardi, F. Ardolino & S. Calamai (eds), *Atti del XV Convegno Nazionale AISV. Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale*. Studi AISV 6, 205-231.

Cresti, E. & Moneglia, M. 2018. The illocutionary basis of Information Structure. Language into Act Theory (L-AcT). In E. Adamou, K. Haude, & M. Vanhove (eds), *Information Structure in Lesser-described Languages: Studies in Prosody and Syntax.* Amsterdam: John Benjamins, 359-401.

Degand, L. & Simon, A.C. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours* 4. http://discours.revues.org/index5852.html (accessed January 7, 2022).

Elfner, E. 2018. The syntax-prosody interface: Current theoretical approaches and outstanding questions. *Linguistics Vanguard* 4(1): 1-14.

Eklund, R. 2004. Disfluency in Swedish human-human and human-machine travel booking dialogues. PhD diss., Linköping University: Electronic Press.

Firenzuoli, V. & Tucci, I. 2003. L'unità informativa di inciso: correlati intonativi. In G. Marotta & N. Nocchi (eds), *La coarticolazione. Atti delle XIII giornate di studio del Gruppo di fonetica sperimentale (AIA)*, Pisa: ETS, 185-192.

Ginzburg, J., Fernández, R. & Schlangen, D. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics* 7(9): 1-64.

Gundel, J. 1988. Universals of topic-comment structure. In M. Hammond, E. Moravcsik, & J. Wirth (eds), *Studies in Syntactic Typology*. Amsterdam: John Benjamins, 209-239.

Hirst, D. & Di Cristo, A. 1998. A survey of intonation systems. In D. Hirst & A. Di Cristo (eds), *Intonation systems: a survey of twenty languages*. Cambridge: Cambridge University Press, 1-44.

Kisler, T., Reichel, U. & Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326-347.

Kjellmer. G. 2003. Hesitation. in defence of er and erm. In *English Studies* 84(2): 170-198.

Krifka, M. 2008. Basic notions of information structure. *Acta Linguistica Hungarica* 55(3-4): 243-276.

Lambrecht, K. 1994. *Information Structure and Sentence Form:Topic Focus and the Mental Representation of Discourse Referents*. Cambridge: Cambridge University Press.

McEnery T. & Gabrielatos C. 2006. English Corpus Linguistics. In B. Aarts & A. McMahon (eds), *The Handbook of English Linguistics*. Oxford: Blackwell, 33-71.

Origlia, A. & Alfano, I. 2012. Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds), *Proceedings of LREC 2012*, Istanbul, Turkey, 21-27 May 2012, 997-1002.

Origlia, A., Savy, R., Poggi, I., Cutugno, F., Alfano, I., D'Errico, F., Vincze, L., & Cataldo, V. 2018. An audiovisual corpus of guided tours in cultural sites: Data collection protocols in the chrome project. In *Proceedigs of the AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage*. Grosseto, Italy.

Origlia, A., Savy, R., Cataldo, V., Schettino, L., Ansani, A., Sessa, I., Chiera, A. & Poggi, I. 2019. Human, all too human. Towards a disfluent Virtual Tourist Guide. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization.* Larnaca, Cyprus, 9-12 June 2019, 393-399.

Reinhart, T. 1981. Pragmatics and linguistics: An analysis of sentence topics in pragmatics and philosophy. *Philosophica* 27(1): 53-94.

Savy, R. 1999. *Riduzioni foniche nel parlato spontaneo: il ruolo della morfologia flessiva nell'interpretazione del messaggio e nella comunicazione*. PhD diss., University of Roma Tre.

Savy, R. 2005a. Specifiche per la trascrizione ortografica annotata dei testi. In F. Albano Leoni & R. Giordano (eds), *Italiano Parlato. Analisi di un dialogo*. Napoli: Liguori.

Savy, R. 2005b. Specifiche per l'etichettatura dei livelli segmentali. In F. Albano Leoni & R. Giordano (eds), *Italiano Parlato. Analisi di un dialogo*. Napoli: Liguori.

Schettino, L., Betz, S., Cutugno, F., Wagner, P. 2021a. Hesitations and individual variability in Italian tourist guides' speech. In C. Bernardasci, D. Dipino, D. Garassino, S. Negrinelli, E. Pellegrino & S. Schmid (eds), *Atti del XVII Convegno Nazionale AISV. Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*. Studi AISV 8, 243-262.

Schettino, L., Betz, S., & Wagner, P. 2021b. Hesitations distribution in italian discourse. *Proceedings of the 10th Workshop on Disfluency in Spontaneous Speech (DiSS 2021)*, 29-34.

Schettino, L. 2022. *The Role of Disfluencies in Italian Discourse. Modelling and Speech Synthesis Applications*. PhD diss., University of Salerno.

Schettino, L., & Cataldo, V. 2019. Lexicalized pauses in Italian. In A. Botinis (ed.), *Proceedings of the 10th International Conference of Experimental Linguistics (ExLing 2019)*, 189-192.

Sloetjes, H. & Wittenburg, P. 2008. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. https://tla.mpi.nl/tools/tla-tools/elan/ (accessed November 16, 2018)

Tottie, G. 2020. Word-Search As Word-Formation?: The Case Of "Uh" And "Um". In P. Núñez-Pertejo, M.J. López-Couso, B. Méndez-Naya & J. Pérez-Guerra (eds), *Crossing linguistic boundaries: systemic, synchronic and diachronic variation in English*. London: Bloomsbury Academic, 29-42.

Voghera, M. & Turco, G. 2008. Il peso del parlare e dello scrivere. In M. Pettorino, A. Giannini, M. Vallone & R. Savy (eds), *La comunicazione parlata*. Napoli: Liguori, 727-760.

Voghera, M. 2017. *Dal parlato alla grammatica*. Roma: Carocci.