An Audiovisual Corpus of Guided Tours in Cultural Sites: Data Collection protocols in the CHROME Project

Antonio Origlia
URBAN/ECO Research Center, University of Naples, "Federico II"
Naples, Italy
antonio.origlia@unina.it

Renata Savy
Department of Humanities Studies, University of Salerno
Salerno, Italy
rsavy@unisa.it

Isabella Poggi
Department of Philosophy, Communication and Performing Arts, Roma Tre University
Rome, Italy
isabella.poggi@uniroma3.it

Francesco Cutugno
Department of Electrical Engineering and Information Technology, University of Naples "Federico II"
Naples, Italy
cutugno@unina.it

Iolanda Alfano
Department of Humanities Studies, University of Salerno
Salerno, Italy
ialfano@unisa.it

Francesca D'Errico
Department of Philosophy, Communication and Performing Arts, Roma Tre University
Rome, Italy
francesca.derrico@uniroma3.it

Laura Vincze
Department of Philosophy, Communication and Performing Arts, Roma Tre University
Rome, Italy
Laura.vincze@gmail.com

Violetta Cataldo
Department of Humanities Studies, University of Salerno
Salerno, Italy
violetta.cataldo@live.itt

ABSTRACT

Creating interfaces for cultural heritage access is considered a fundamental research field because of the many beneficial effects it has on society. In this era of significant advances towards natural interaction with machines and deeper understanding of social communication nuances, it is important to investigate the communicative strategies human experts adopt when delivering contents to the visitors of cultural sites, as this allows the creation of a strong theoretical background for the development of efficient conversational agents. In this work, we present the data collection and annotation protocols adopted for the ongoing creation of the reference material to be used in the Cultural Heritage Resources Orienting Multimodal Experiences (CHROME) project to accomplish that goal.

CCS CONCEPTS

• **Human-centered computing → User studies**; **HCI theory, concepts and models**; **User models**;

KEYWORDS

Corpus collection, guided tours, social signal processing

ACM Reference Format:
Antonio Origlia, Renata Savy, Isabella Poggi, Francesco Cutugno, Iolanda Alfano, Francesca D'Errico, Laura Vincze, and Violetta Cataldo. 2018. An Audiovisual Corpus of Guided Tours in Cultural Sites: Data Collection protocols in the CHROME Project. In *Proceedings of 2nd Workshop on Advanced Visual Interfaces for Cultural Heritage (AVI-CH 2018).* Vol. 2091. CEUR-WS.org, Article 8. http://ceur-ws.org/Vol-2091/paper8.pdf, 4 pages.

AVI-CH 2018, May 29, 2018, Castiglione della Pescaia, Italy

1 INTRODUCTION

Developing Social Signal Processing [15] techniques for advanced, natural interfaces requires a significant analysis effort on multiple aspects of communication between individuals engaging in social activity. Collecting meaningful corpora to document the multimodal signals people exchange during these activities has been the subject of a large amount of research. Among others, available corpora document meetings [6], intercultural dynamics of first acquaintance [1], phone calls between non acquainted subjects [10], and two-person dialogues [14]. The Italian national project CHROME aims at developing a data collection and annotation procedure to support the development of new interactive technologies for cultural heritage. The project concentrates on the three Campanian Charterhouses: an integrated description of these from different point of views (textual, behavioural, geometrical, etc…) is being developed.

In this paper, we present the data collection and annotation protocols adopted in the CHROME project to obtain reference material of expert gatekeepers, intended as holders of knowledge for others to refer to, accompanying visitors of cultural sites. This data will be used to investigate the social communication strategies adopted by the considered experts to deliver information to different groups of visitors. By comparing different experts (inter-subject comparisons) and different groups accompanied by the same expert (intra-subject comparison) a Gatekeeper Computational Model will be obtained and, on the basis of this model, a socially aware conversational agent, in the form of a 3D avatar, will be developed. This is expected to improve the capabilities of an interactive agent to involve people in engaging presentations of cultural heritage. These will make use of the 3D reconstructions of the three Campanian Charterhouses, also collected in the framework of the CHROME project.

.

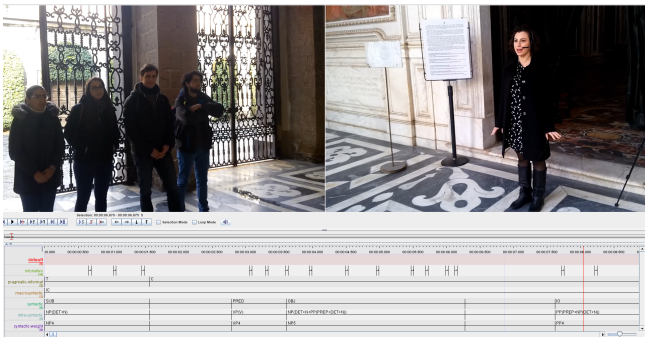

Figure 1: A screenshot of the ELAN interface showing the synchronised videos of the expert and of the audience, together with an example annotation.

Upon completion of the project, the dataset will be made freely available for the scientific community.

In the next sections, we will present the data collection protocol, highlighting the chosen recording positions in the site of interest and the recording setup. We will, then, present the multimodal annotation protocol, designed to provide a formal description of how the guide makes use of social signals exchange to adapt the presentation and to effectively support the verbal transfer of cultural contents. Next, we will describe the informative, syntactic and prosodic annotations documenting the linguistic behaviour that characterises the domain expert. The transcribed recordings, together with the produced annotations, will be compared with a corpus of textual resources describing the objects of interest. This will support the development of a synthetic voice model for 3D avatars designed to extract cultural contents from textual databases and deliver them using social communication strategies. To improve the quality of the model, the linguistic analysis will also include a detailed annotation of disfluency phenomena, which are important to produce a *natural* sounding voice.

2 DATA COLLECTION

The data collection plan foresees a campaign of audiovisual recordings involving four art historians with strong experience in accompanying groups of visitors. Given the limited number of gatekeepers considered in the CHROME project, only female experts were recruited to remove gender effects in multimodal and linguistic analysis. Future extensions of the corpus will include male experts as well.

Recorded data include two Full-HD video recordings: the first one is a fixed shot of the gatekeeper, taken from a position immediately next to the attending group while the second one is a fixed shot of the visitors. A close range digital microphone with background noise cancellation is used to record the gatekeeper's voice. Immediately after the visit, the recruited visitors compile a questionnaire composed of 23 items including both Likert scale evaluations and open answer questions. The items are designed to collect anagraphic data, a self-evaluation of artistic competence, an evaluation of personal satisfaction after the visit and an evaluation of the gatekeeper's performance. These data will be used to weight objective measures of social behaviour.

Each recruited expert accompanies four groups of four people in an hour long guided tour at the San Martino Charterhouse in Naples. Recruited members of the audience vary on a socio-demographic basis and each group is gender balanced. The visit is divided into six points of interest (POIs), selected as the most relevant parts of the Charterhouse from an architectural and artistic point of view:

- *Pronaos*: outside the doorstep of the church. The introductory part of the visit is recorded in this POI. Environmental elements mainly consist of architectural details;
- *Great cloister*: a large external place, near the monks' cemetery. Further details about the monks' life are given. Environmental elements consist of the natural setting of a large garden and of the cemetery elements (e.g. *memento mori*);
- *Parlor*: the first internal setting. Specific details about the Charthusians' rules are given here. Environmental elements mainly consist of frescoes;
- *Chapter hall*: next to the parlor. Specific details about the Charthusians' order are given here. Environmental elements mainly consist of frescoes;
- *Wooden choir*: inside the church, behind the altar. The history of the church decoration process is given here. Environmental elements consist of both architectural details (e.g. the choir and the harmonic chassis) and artistic elements (frescoes and statues);
- *Treasure hall*: deeper inside the complex. Details about the relationship between the monks and the different governing parties in Naples are given. Environmental elements mainly consist of architectural details.

The selected POIs allow us to capture the social behaviour visitors and gatekeepers exhibit to negotiate the approach to the visit and to document postural and gestural behaviour of an art historian presenting a complex environment.

Videos and audio recordings are synchronised *a posteriori* using a visual-acoustic marker. Linguistic and multimodal annotations, performed on the synchronised versions of the collected material, will be merged using the ELAN software [17]. An ELAN project file will be produced for each POI visit in order to allow cross-domain research and closed vocabularies for the label sets belonging to each annotation domain will be used to ensure consistency. An example of the ELAN interface showing the two video shots and a sample annotation tier is shown in Figure 1.

3 MULTIMODAL ANNOTATION

The video recording of the expert gatekeeper is annotated as to the structure of verbal discourse and to body communicative behaviour. The *discourse structure* point of view is based on a previous analysis on videos of Art Commentators (ACs), that is, both museums gatekeepers and art historians illustrating artworks in tv, where a general *script* was extracted of what the AC can /should say in one's work. This allowed to outline the typical discourse structure of any AC which, based on the analysis of discourse as a hierarchy of goals [8], distinguishes four main goals pursued by the gatekeeper: a general goal of *cultural elevation*; encompassing favouring aesthetic enjoyment, imagination and emotion triggering, and, subsumed to it; the *textual goals* of providing information about the opera, its history, function, cultural milieu, and the author; the corresponding

Table 1: An example of multimodality annotation.

Verbal Text	Discourse function	Gesture	Meaning
The Saint Martin's Charterhouse here in Naples has at least two souls	Textual goal: Information on the artwork	*hands, palms to each other, like framing something*	I am framing the object of discourse, Metacognitive gesture
Nowadays it is not only a Charterhouse	Textual goal: Information on the identity of the artwork	*Left hand moves to left, Metadiscursive gesture*	I locate the identity *Charterhouse* on my left → I build the first entity
but it is also a national museum	Textual goal: Information on the identity of the artwork	*Right hand moves to right. Metadiscursive gesture*	I locate the identity known as *Museum* on my right -> I build the second entity
So try to imagine Naples 700 years ago	Emotional goal: Solicit imagination	-	-

modal goals of attracting and sustaining attention, favouring comprehension and inferential connections with the tourists' previous knowledge; *interactional goals* such as tuning, setting empathic connection with tourists. Each particular performance of a gatekeeper or other AC can be analysed in terms of this abstract script, and this allows, among other things, to distinguish the idiosyncratic styles of different ACs in terms of which nodes of the structure they prefer to expand. Some mainly focus on the author and his life, some on the deep symbolic meanings of the artwork, some on the author's style and the surrounding cultural milieu, and so on.

The analysis of the gatekeeper's *multimodal communication* takes into account the following body communicative modalities: gestures, postures, head movements, facial expression, gaze communication. For each communicative item in each modality, the signal is annotated in ELAN in terms of a detailed description of its production: gestures are described according to their parameters of hand configuration, location, orientation and movement; gaze in terms of eye direction, eyebrows and eyelids movements; face in terms of Ekman's FACS; head movements in term of head nod, shake, toss, canting; postures in terms of leg and trunk movements. Then, for the signal described in this way, a verbal phrasing of its meaning is provided (after [9]). Based on this meaning, the item is classified as to its role and function within the gatekeeper's discourse structure. An example of multimodal annotation is shown in Table 1.

Table 2: Disfluency annotation levels

Disfluency Type	Type of disfluent phenomenon
Disfluency Function	Pragmatic function of the disfluent phenomenon
Disfluency Model	Model of occurrence
Disfluency Components	Internal regions of the phenomenon

4 LINGUISTIC ANNOTATION

Using the close-mic recordings, speech produced by the expert gatekeeper is analysed and annotated on different levels. From the *informative-syntactic* point of view, an *orthographic* level is produced on the basis of the indications provided by [11]. This level involves the transcription of a number of elements: lexical elements, silent and filled pauses, noises, vocal (nonverbal) phenomena, truncated words, interrupted words, false starts and lapsus linguae. A *phonetic* level is included to store the phonetic transcription of the utterances and markers of phonetic phenomena like coarticulation, following the indications found in [12]. A *syllabic* level is produced to allow speech fluency and speech rate analyses. A *disfluency* level, involving the annotation of disfluency phenomena [3, 13], is also included. This analysis level consists of four annotation tiers, detailed in Table 2.

To document the prosodic component of the experts' linguistic behaviour, a multilevel annotation, structured in different tiers, has been produced. The considered aspects include: an *intonative level*, using the INTSINT coding scheme [4, 5], providing a labels sequence representing the f0 curve, obtained with the Prosomarker tool [7]; a *pragmatic - informative level*, providing an analysis of information structure considering topic (preposed or postposed) and comment units [2]; a *macro-syntactic level*, indicating the types of clauses dividing independent clauses from dependent clauses and specifying the type of subordination; a *syntactic level*, describing the main syntactic functions; an *intra-syntactic level*, labelling the type of phrase and its composition (between parenthesis); a measure of *syntactic weight*, based on [16], which takes into account both the structure and the length of constituents. It considers the following features: ± presence of determiners, ± presence of modifiers, ± presence of pronouns, ± verbal valency saturation. An annotation example is shown in Figure 2.

5 CONCLUSIONS AND FUTURE WORK

We have presented the data collection and annotations protocols for a work in progress on an audiovisual corpus documenting how cultural heritage gatekeepers support people in accessing architectural heritage and consists of both video and audio recordings to capture the social interaction process taking place between the

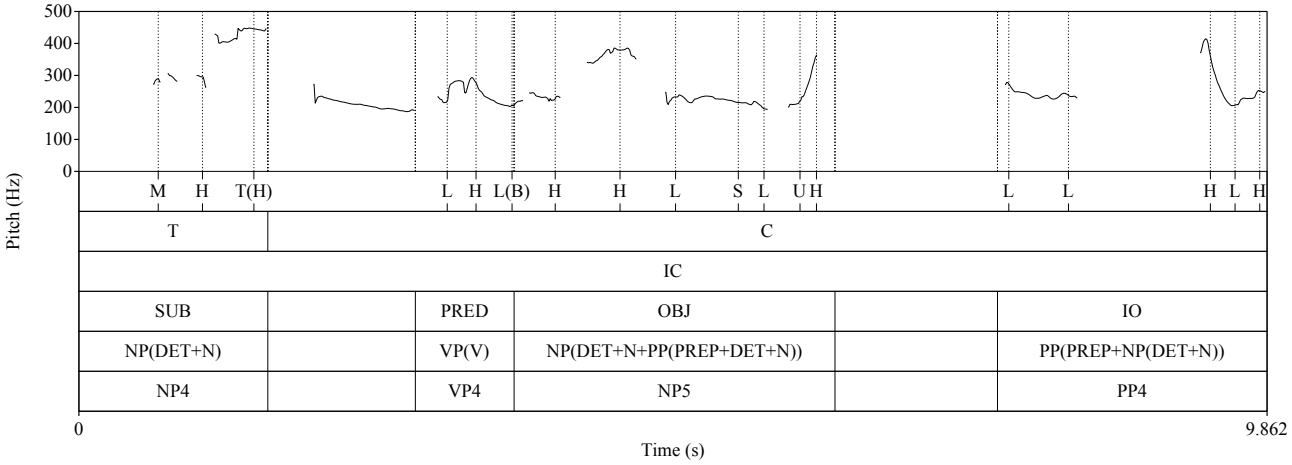

Figure 2: An example on the utterance *I certosini devono la fondazione del loro ordine a un uomo* (*Carthusians due their order's foundation to a man*). The order of the annotation tiers is the one found in the text.

group guide and the attending audience. Annotation levels cover linguistic and multimodal aspects of communication to allow a multi-faceted investigation of the ongoing communicative process. The collected material will be used as reference to build a computational model of a 3D virtual character presenting reconstructions of architectural heritage sites.

6 ACKNOWLEDGMENTS

Antonio Origlia's work is funded by the Italian PRIN project *Cultural Heritage Resources Orienting Multimodal Experience* (CHROME) #B52F15000450001.

REFERENCES

[1] Jens Allwood, Nataliya Berbyuk Lindström, and Jia Lu. 2011. Intercultural dynamics of fist acquaintance: comparative study of swedish, chinese and swedish-chinese first time encounters. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 12–21.

[2] Jeanette K Gundel. 1988. Universals of topic-comment structure. *Studies in syntactic typology* 17 (1988), 209–239.

[3] Adolf E Hieke. 1981. A content-processing view of hesitation phenomena. *Language and Speech* 24, 2 (1981), 147–160.

[4] Daniel Hirst and Albert Di Cristo. 1998. A survey of intonation systems. *Intonation systems: A survey of twenty languages* (1998), 1–44.

[5] Daniel Hirst, Albert Di Cristo, and Robert Espesser. 2000. Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and experiment*. Springer, 51–87.

[6] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, and others. 2005. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, Vol. 88. 100.

[7] Antonio Origlia and Iolanda Alfano. 2012. Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabi fication.. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*. 997–1002.

[8] Domenico Parisi and Cristiano Castelfranchi. 1976. *The discourse as a hierarchy of goals*. Centro Internazionale di Semiotica e di Linguistica, Università di Urbino.

[9] Isabella Poggi. 2007. *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler.

[10] Hugues Salamin, Anna Polychroniou, and Alessandro Vinciarelli. 2013. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 4282–4287.

[11] Renata Savy. 2005. Specifiche per la trascrizione ortografica annotata dei testi. *Italiano Parlato, Analisi di un dialogo* (2005), 1–28.

[12] Renata Savy. 2005. Specifiche per l'etichettatura dei livelli segmentali. *Italiano Parlato. Analisi di un dialogo. Napoli: Liguori* (2005).

[13] Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. Dissertation. Citeseer.

[14] Yasir Tahir, Debsubhra Chakraborty, Tomasz Maszczyk, Shoko Dauwels, Justin Dauwels, Nadia Thalmann, and Daniel Thalmann. 2015. Real-time sociometrics from audio-visual features for two-person dialogs. In *Digital Signal Processing (DSP), 2015 IEEE International Conference on*. IEEE, 823–827.

[15] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[16] Miriam Voghera and Giuseppina Turco. 2007. Il peso del parlare e dello scrivere. In *Proc. of International Conf.* Il Parlato Italiano, *Liguori, Napoli*.

[17] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*. 1556–1559.