Exploring Attention in a Multimodal Corpus of Guided Tours

Andrea Amelio Ravelli[◇], **Antonio Origlia**[•], **Felice Dell'Orletta**[◇]
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{andreaamelio.ravelli, felice.dellorletta}@ilc.cnr.it
[•] University of Naples "Federico II"
antonio.origlia@unina.it

Abstract

This paper explores the possibility to annotate engagement as an extra-linguistic information in a multimodal corpus of guided tours in cultural sites. Engagement has been annotated in terms of gain or loss of perceived attention from the audience, and this information has been aligned to the transcription of the speech from the guide. A preliminary analysis suggests that the level of engagement correlates with some specific linguistic features, opening up to possible future exploitation.

1 Introduction

Understanding a message expressed through the speech channel in face-to-face interactions involves more than the ability to decipher a string of characters and to assign a meaning to words and sentences. The linguistic information conveyed by lexicon is only the tip of the iceberg: intonation, gesture, facial expression, gaze, body movement play a key role in spoken communication. By summing the information in all these complementary modalities acquired through different channels (i.e. auditory and visual systems), the human brain is capable to analyse and decode a message not only on the basis of the words it contains. Moreover, the vision modality enables the speaker to evaluate the effectiveness of his/her message on the audience. In fact, face-to-face interactions offer the possibility to have an on-line feedback from the addressee even without an ongoing active dialogue. Simply by interpreting unconscious signals accessible from the vision modality, such as body postures and movements, facial expressions, eye-gazes, the speaker can understand if the addressee is engaged with the discourse, and continuously fine-tune his/her communication strategy in order to keep the attention high in the audience.

Engagement can be explained as the process by which two or more actors establish, maintain and end their perceived connection during interactions they jointly undertake (Rich et al., 2010). It is composed of a series of verbal and non verbal behaviours, useful to understand the involvement between the actors, and specifically between the actors and the content of their communication scene, and it can be used to provide evidence of the waning of connectedness (Sidner et al., 2005).

In this work we describe a pilot annotation of audience engagement during guided tours in cultural sites, by evaluating the observable behaviours of the visitors in response to the speech from the guide. The main goal is to trace the level of attention of the visitors. Engagement is defined as a multidimensional meta-construct (Fredricks et al., 2004), and attention is considered a component of its the visible cues.[1] The paper is organised as follows: section 2 introduces the CHROME project and its multimodal corpus; section 3 describes the visual annotation; section 4 reports the results of the annotation in terms of agreement and some linguistic analysis on the available set of aligned transcriptions; section 5 concludes with some discussions on possible future works and exploitation for this kind of resource.

2 The CHROME Project

The Italian national project Cultural Heritage Resources Orienting Multimodal Experience (Origlia et al., 2018) aims at developing a data collection and annotation procedure to support the develop-

[1] Per definition, cognitive engagement refers to internal processes, whereas only the emotional and behavioral components are manifested in visible cues. Nevertheless, all engagement elements are highly interrelated and do not occur in isolation (Fredricks et al., 2004). Thus, attention plays a crucial role (Goldberg et al., 2019).

ment 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 the framework of this project, a data collection campaign to document how professional guides present architectural heritage contents when on-site was defined.

2.1 The CHROME multimodal corpus

The collected data consist of audiovisual recordings involving three art historians with strong experience in accompanying groups of visitors. Given the limited number of informants considered in the CHROME project, only female experts were recruited to remove gender effects in multimodal and linguistic analysis.

Recorded data include two Full-HD video recordings: the first one is a fixed shot of the art historian, taken from a position immediately next to the attending group, while the second one is a fixed shot of the group of recruited visitors. A close-range digital microphone with background noise cancellation is used to record the guide's voice.

Each recruited expert accompanied 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, are merged and aligned using the ELAN software (Wittenburg et al., 2006). An ELAN project file is produced for each POI visit in order to allow cross-domain research and closed vocabularies for the label sets belonging to each annotation domain are used to ensure consistency. Specifically about linguistic annotations, the considered levels consisting of word, syllable and phone level transcriptions are obtained using WebMAUS (Kisler et al., 2017) and manually checked by human experts. Also, tonal units are manually marked by a human expert, as well as syntactic structures.

3 Engagement annotation

A subset of data from the CHROME Project has been used for this work. More specifically, we acquired data for one guide accompanying four different groups of visitors in the Charterhouse of St. Martin in Naples, consisting in 24 video couples (aligned videos of both the guide and the audience, one couple for each POI). Annotation has been performed by two annotators by means of PAGAN annotation web-based platform (Melhart et al., 2019), which enables the users to easily align and play two videos. Annotators have been

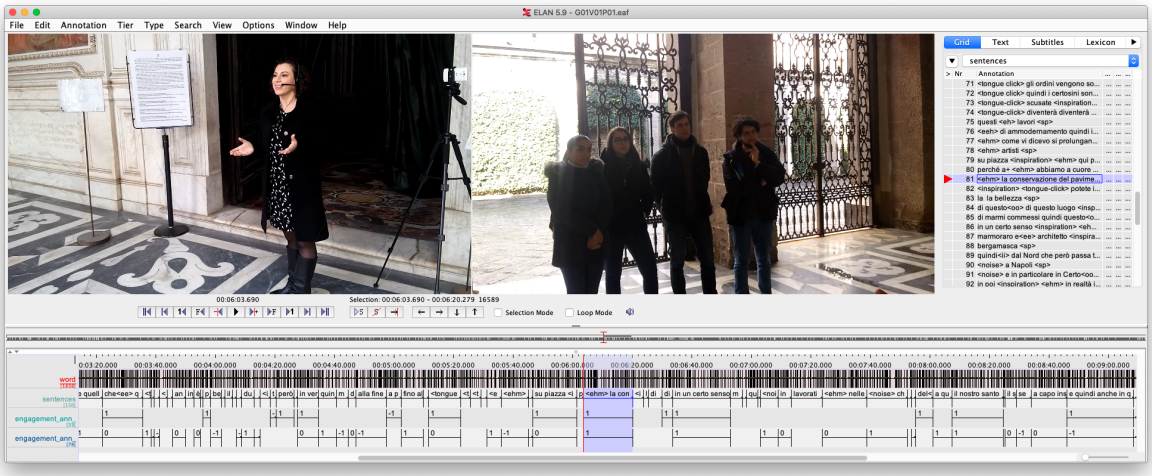

Figure 1: Elan project file with the alignment of the annotations.

asked to recognise signals of gain or loss of attention in the audience, and they recorded their observations through simple interactions with the up and down keys of the keyboard, where up stands for a gain and down for a loss in attention. Given the nature of the annotation (and the scope of this pilot work), no strict instructions have been delivered to the annotators. They based their judgement on visible cues of perceivable variation in the level of attention from the group of visitors, such as gaze following a deictic gestures, facial expressions as feedback to the guide's speech, head movements, pose and so on. The interactions in PAGAN are recorded using RankTrace framework (Lopes et al., 2017), and the whole annotation session is exported as a tab-separated file containing continuous series of milliseconds and values for each interaction. In total, the set of videos consists of ∼3:20 hours, with an average length of ∼8:40 minutes per point of interest.

For 3 of these videos it was already available[2] the ELAN project file containing the orthographic transcription of the guide's speech (more specifically, the speech from the visit in the POI 1 with the first three groups), thus it has been possible to automatically align the visually-derived annotation, using the pympi-ling Python Module (Lubbers and Torreira, 2018).

Figure 1 shows an example of the alignment for one of the videos in an ELAN project file. Using these alignments it has been possible to investigate if any correlation exists between linguistic features extracted from the guide's speech and engagement from the visitors.

[2]The transcription and annotation of the whole corpus of the CHROME Project is still an ongoing work, thus completely annotated and aligned data is still limited.

4 Evaluation of the corpus

Video	Length	Spearman's rho
Group 1		
POI 1	00:11:33	0.94
POI 2	00:08:42	0.83
POI 3	00:05:17	0.70
POI 4	00:05:46	0.87
POI 5	00:06:47	0.72
POI 6	00:10:08	0.94
Group 2		
POI 1	00:13:12	0.98
POI 2	00:08:45	0.91
POI 3	00:05:24	0.39
POI 4	00:06:25	0.92
POI 5	00:08:09	0.83
POI 6	00:12:08	0.43
Group 3		
POI 1	00:16:18	0.98
POI 2	00:10:43	0.98
POI 3	00:07:38	0.98
POI 4	00:08:43	0.90
POI 5	00:05:40	0.89
POI 6	00:13:07	0.99
Group 4		
POI 1	00:02:35	0.93
POI 2	00:10:20	0.98
POI 3	00:07:17	0.89
POI 4	00:07:21	0.97
POI 5	00:05:52	0.98
POI 6	00:11:10	0.98
AVG	**00:08:42**	**0.87**

Table 1: Correlations on the annotations for each video.

To evaluate the agreement and thus the reliability of the annotation, we calculated the Spearman's rho for the continuous series of values from the two annotators. Table 1 reports the results of the correlations: the overall agreement is significantly high, with a average correlation between the two series of 0.87. Figure 2 and 3 shows respectively the plot for highest and lowest correlation.

Linguistic Feature	Ann_1		Ann_2	
	Positive	Null	Positive	Null
	Avg (St.Dev)	Avg (St.Dev)	Avg (St.Dev)	Avg (St.Dev)
n_tokens	19.78 (14.63)**	10.42 (9.79)**	16.68 (13.78)**	5.68 (5.72)**
% NOUN	15.97 (9.69)	17.32 (14.32)	16.6 (10.09)	16.54 (16.98)
% PROPN	4.48 (11.7)*	4.24 (9.9)*	4.99 (11.24)**	4.12 (11.82)**
% PRON	7.65 (8.04)**	6.77 (11.85)**	8.13 (12.65)**	4.58 (9.88)**
% VERB	11.33 (9.2)*	12.2 (18.07)*	11.12 (9.71)*	16.91 (27.33)*
% AUX	5.87 (7.19)**	5.07 (12.12)**	5.73 (11.64)**	4.63 (13.38)**
% ADJ	3.94 (5.04)	5.06 (10.91)	4.69 (6.43)**	4.07 (14.18)**
% ADV	14.14 (13.49)**	13.55 (20.19)**	13.12 (14.75)**	12.97 (22.68)**
% DET	15.49 (13.99)	14.74 (12.73)	15.75 (10.42)**	13.66 (17.5)**
% NUM	0.32 (1.35)	0.42 (2.45)	0.52 (2.36)**	0.21 (2.29)**
% CCONJ	4.85 (15.34)**	2.48 (8.16)**	2.7 (5.77)**	4.21 (16.25)**
% SCONJ	2.48 (3.52)**	3 (11.46)**	2.05 (4.29)**	2.97 (11.97)**

Table 2: Average and Standard deviation of Profiling-UD linguistic features for POS distributions (in percentage) and number of tokens per sentence. * p<0.05; ** p<0.01.

Such information can be used to extract meaningful segments concerning the level of attention (e.g for machine learning purposes).

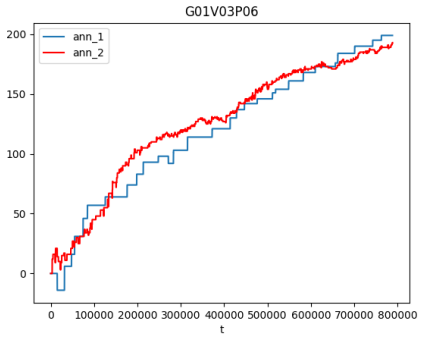

Figure 2: Plot of annotations for the video with the highest correlation (Spearman's rho: 0.99).

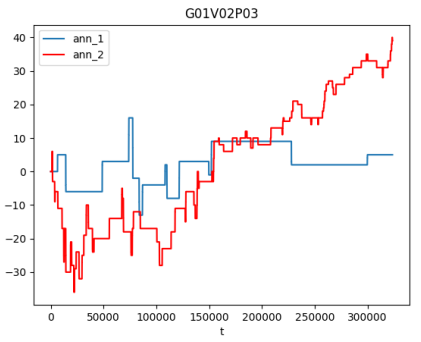

Figure 3: Plot of annotations for the video with the lowest correlation (Spearman's rho: 0.39).

4.1 Linguistic features correlation

As briefly mentioned before, we exploited the corpus composed of available orthographic transcriptions to carry out some analysis about the possible correlation between content of the speech and the perceivable engagement of the audience. To do so, we considered pause tags, i.e. short and long pauses (respectively, <*sp*> and <*lp*>), as boundaries for sentence-like units of text to be processed along with the corresponding engagement value. We are aware that breath groups cannot be considered as reference units for the analysis of speech,[3] and that applying written language methodologies and tools to spoken modality is biasing (Linell, 2005; Linell, 2019), but for the scope of the present work it has been necessary to make use of the available segmentation.

Even if we had few text available (3 transcriptions, for a total of 5,648 tokens in 464 sentences; ∼12 tokens per sentence), we analysed the corpus using Profiling-UD[4] (Brunato et al., 2020), a web–based application that performs linguistic profiling of a given text. The output of Profiling-UD is a tab-separated file, with one row per document (one for sentence, in this case) and one column for each of the 122 linguistic features analysed by the system. The objective is to investigate

[3]Segmentation of speech in basic units is still an open challenge in spoken language studies, as recently testified by Izre'el et al. (2020) and Mello et al. (2020).

[4]`http://linguistic-profiling.italianlp.it`

if any relation could be traced between the perceived attention from the audience and the linguistic features extracted from the guide's speech. We observed the scores for the sentences marked with a gain of attention against those for which annotators did not interact with the platform (i.e. those sentences that, aligned with time stamps to the series of the annotations, was not marked as gain or loss of attention). We performed the Wilcoxon rank sum test on features values for the two groups of sentences (positive vs. null) for both the annotators.

Table 2 reports average and standard deviation for the linguistic features with $p<0.05$ for at least one annotator.[5] It is possible to notice that, among positive and null marked sentences in both the annotator's data, the feature that significantly varies more than the other are the length of sentences (n_tokens) and the distribution of auxiliars.

The correlation between length and attention is not surprising, since longer sentences are likely to be more informative and thus probably more engaging. Even if sentence length is normally associated to a higher sentence complexity (Brunato et al., 2018), other typical features of complexity are not appreciably, given that subordinative conjunctions (SCONJ) are sensibly lower in higher attention marked sentences, while coordinative conjunctions (CCONJ) shows opposite trend in the two groups. For both the groups proper names (PROPN) and pronouns (PRON) seem to characterise engaging sentences.

5 Conclusions and Future Works

In this work we introduced a pilot annotation of visually perceivable attention, meant as a component of engagement, and its alignment in a multimodal corpus of guided tours in cultural sites. Moreover, we analysed the available speech transcription for 3 of the 24 videos and, notwithstanding the small dimension of the corpus ($\sim$5K tokens), some signal of the connection between attention and specific lexical features emerges, and it would be interesting to augment data in terms of annotations and alignment in order to extensively verify these correlations. Much more reliable analysis may be carried on by exploiting better textual segmentation, e.g. tonal units, and fine-tuning the feature extraction procedure in order to better handle spoken language. In this way, it would be possible to account also spoken-specific peculiarities and correlate them to audience engagement.

Finally, in the specific context of hosting and guiding visitors in cultural sites, the possibility to trace the level of engagement during tours can open up to interesting outcomes. In this regard, aligning speech transcription with attention tracking and other data, such as gaze, intonation, gesture, facial expression, body movement (for both the speaker and the addressee), would be particularly useful to train a classifier to recognise engaging information both in spoken language and in videos.

[5] In this analysis we consider exclusively features on sentence length and part-of-speech distributions. Profiling-UD is a tool designed for written text and not trained to work on speech transcriptions, thus any significance on syntactic features is not reliable.

References

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151.

Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109.

Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2019. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educational Psychology Review*, 35(1):463–23, January.

Shlomo Izre'el, Heliana Mello, Alessandro Panunzi, and Tommaso Raso. 2020. *In Search of Basic Units of Spoken Language*, volume 94 of *A corpus-driven approach*. John Benjamins Publishing Company, Amsterdam, June.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.

Per Linell. 2005. *The Written Language Bias in Linguistics. Its Nature, Origins and Transformations*. Routledge.

Per Linell. 2019. The Written Language Bias (WLB) in linguistics 40 years after. *Language Sciences*, 76:101230.

Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. Ranktrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 158–163. IEEE.

Mart Lubbers and Francisco Torreira. 2018. pympi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files. `https://pypi.python.org/pypi/pympi-ling`. Version 1.69.

David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. Pagan: Video affect annotation made easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 130–136. IEEE.

Heliana Mello, Lúcia Ferrari, and Bruno Rocha. 2020. Multimodality, Segmentation and Prominence in Speech. *Journal of Speech Sciences*, 9:1–6.

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. *JOWO*, 2091.

Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. 2010. Recognizing engagement in human-robot interaction. In *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382. IEEE.

Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.

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)*, pages 1556–1559.