Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 628–633
Marseille, 11–16 May 2020

PACO : A corpus to analyze the impact of common ground in spontaneous face-to-face interaction

Mary Amoyal, Béatrice Priego-Valverde, Stéphane Rauzy
Aix Marseille Université, CNRS, Laboratoire Parole et Langage UMR 7309
5 avenue Pasteur, BP 80975, 13604 Aix-en-Provence, FRANCE
{mary.amoyal, beatrice.priego-valverde, stephane.rauzy}@univ-amu.fr

Abstract

PACO is a French audio-video conversational corpus made of 15 face-to-face dyadic interactions, lasting around 20 min each. This compared corpus has been created in order to explore the impact of the lack of personal common ground (Clark, 1996) on participants collaboration during conversation and specifically on their smile during topic transitions. We have constituted this conversational corpus PACO by replicating the experimental protocol of "Cheese!" (Priego-Valverde et al., 2018). The only difference that distinguishes these two corpora is the degree of CG of the interlocutors: in Cheese! interlocutors are friends, while in PACO they do not know each other. This experimental protocol allows to analyze how the participants are getting acquainted. This study brings two main contributions. First, the PACO conversational corpus enables to compare the impact of the interlocutors' common ground. Second, the semi-automatic smile annotation protocol allows to obtain reliable and reproducible smile annotations while reducing the annotation time by a factor 10.

Keywords: Common ground, spontaneous interaction, smile, automatic detection

1. Introduction

This work is part of a larger on-going study of smiling during topic transitions and its evolution depending on the interlocutors' shared knowledge in French conversations. There are evidence that face-to-face interactions are "collaborative productions" (Sacks et al., 1978). This cooperation will be studied in specific moments that particularly require the interlocutors adjustment : topic transitions. These thematic movements are moments when interlocutors switch from a discursive topic to another (Garcia and Joanette, 1997). This shift can lead to a negotiation of the next topic to be addressed and the interlocutors must jointly approve the next topic of discussion (Berthoud and Mondada, 1995). This required adaptation can lead to the mobilization of different modalities (Riou, 2015). Among the several modalities used, we pay a particular attention to smile. This facial expression involves several face muscles such as (among others) the zygomatics and the orbicular. The solicitation of these different muscles and the differences observed between a low or a high smile (El Haddad et al., 2019) provides evidence that it is important to take into account the different intensities of smiles. Theses evidences has led us to no longer approach this gesture in a binary way (e.g. presence/absence). The following study provides a smile representation that considers the smile granularity (Gironzetti et al., 2016) which results in a scale of smile intensity (from 0 to 4). This methodology allows to be interested in the smiles evolution as this could be associated to a specific pragmatic interpretation. Smile is herein considered as an "interactive gesture" (Bavelas and Gerwing, 2007). There is some evidence that smiling is a social behavior intended for communication as smiling is determined and better predicted by the social context rather than by inner emotions (Fridlund, 1994). Up until now, smile in interaction has not been sufficiently documented, as it has mainly been treated in a perception perspective (Aubergé and Cathiard, 2003; Fagel, 2010) and in relation to its emotional aspects (Ekman and Friesen, 1975; Tartter, 1980).

In a previous study, we have shown that participants who are friends are more likely to decrease their smile when they initiate a transition with a verbal marker. On the contrary, when they do not use any verbal marker to realize a topic shift, they are more likely to increase their smile (Amoyal and Priego-Valverde, 2019). Nevertheless, these preliminary results need to be deepened based on a larger collection of conversational data. In a second stage, we plan to compare these results with the ones found in conversations involving interlocutors who do not know each other.

Whether or not the participants know each other is of great importance as it has been shown that interlocutors rely on their "common ground" (Clark, 1996) - henceforth CG - to jointly construct a conversation. This CG could come from different domains (Clark, 1996):

- Incremental : the construction of common ground during the interaction (i.e. the process of grounding).
- Communal: the knowledge shared by interlocutors from the same culture (e.g. the language).
- Personal : the knowledge shared by the participants (e.g. their conversational history (Golopentja, 1988)) as a result of their prior common experience.

This last point is our focus since, to our knowledge, few studies have examined the impact of the personal CG on the mechanism of conversational interactions. This impact has been further studied on controlled interactions (see (Holler and Bavelas, 2017) for an overview). Therefore, this study was conducted with the aim of providing spontaneous data that allows a systematic analysis of the role of personal CG in the conversation construction. In the long term, the objective of this study is to answer the following question: What is the role of the smile in conversational interactions? More specifically, how does the smile evolve according to the topic transitions and the degree of the interlocutors' CG ? In the shorter term, this study focuses on methodological aspects necessary to explore this research question. First, we will present "PACO", the collection of conversational corpus (5h) of face-to-face interactions in French that allows to compare the interaction organization in regard with the personal CG (controlled factor) of the interlocutors. Second, we will present the methodology developed in order to analyze this corpus. This methodology includes basic levels such as the transcription of the speech signal but also a more original and specific enrichment for our variable of interest which concerns smiles. We will thus present our semi-automatic smile annotation protocol, followed by the future works planned in order to enrich and analyze our corpus.

2. Collection of data

PACO is an audio-video corpus recorded in 2018 at the LPL - Laboratoire Parole et Langage, Aix-en-Provence, France, at the Centre for Speech Experimentation (CEP)[1]. This corpus is composed of 15 face-to-face dyadic interactions lasting around 20 min each, which represent a dataset of more than 5 hours of conversational data.

2.1. Participants

PACO brings together 30 French native students (24 female and 6 male) from 20 to 35 years old (mean= 22, sd= 5). Out of the 15 interactions, 6 are mixed (F and M) and 9 are non-mixed (F and F). All the participants were volunteers and one third of them have received a financial compensation. Before recording, they were all informed of their image rights as well as their retraction rights. Thus, every participants shown in this study agreed to be recorded in audio and video. They all signed a written consent form before the recordings but none of them knew the scope of the recordings in order not to bias their behaviors. At the end of the data collection, we explained to the participants that smile and the conversation organization were going to be analyzed. Each recording combines two participants that met for the first time in the anechoic room. As we wanted to control the factor "common ground", the lack of interpersonal knowledge was the specific and decisive characteristic for the participants recruitment.

2.2. Experimental design

PACO has been recorded in respect with the Cheese! protocol (Priego-Valverde et al., 2018) where participants were friends. Indeed, the principal aim of this data collection is to compare the impact of the presence or absence of the interlocutors Common ground. In that aim, the Cheese! protocol has been replicated with the only difference that participants did not know each other, other things being equal. Participants were recorded in a soundproof room where they were seated face-to-face. They were each fitted with a headset microphone (Sennheiser HSP4 EW) optimally positioned in order not to hide their mouth. Two cameras (CANON XF105) were positioned in such a way that each participant was filmed from the front (see figure 1).

[1]CEP is a shared experimental platform for the collection and analysis of data for the study of speech production and perception.

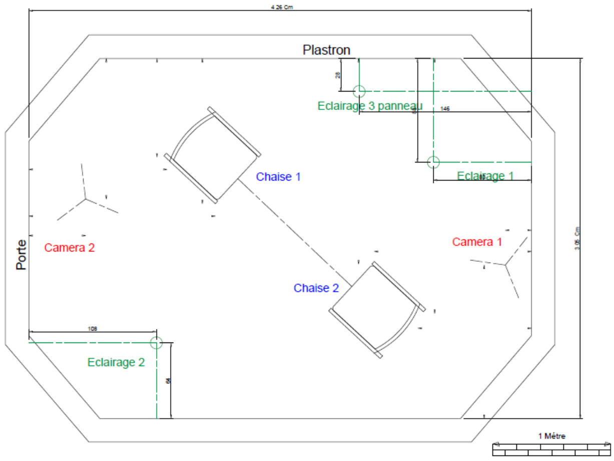

Figure 1: Experimental design of the soundproof room with the position and distance (in meter) of the chairs, the micros, cameras and the lights

Two tasks were delivered to the participants: first, they were asked to read each other a canned joke (Attardo et al., 2011); second, they were asked to converse as freely as they wished for the rest of the interaction. Our double instruction constitutes a strong advantage. The reading task allows us to have similar linguistic material for all speakers. This allows us to identify how the first thematic transition to the second conversation task is negotiated and to analyze how speakers manage their first speech turn (auto/hetero selection) (Sacks et al., 1978) to the conversational sequence. The second conversational task without any predefined topic allows to observe how the conversational topic are negotiated depending on whether or not the participants know each other. This experimental protocol leads to a very high quality data-set in term of video and audio records. Indeed, as each participant is recorded in one audio channel with a sampling frequency of 48000Hz, it facilitates the transcription steps and eventually the automatic annotations based on the audio signal (`.wav`). Moreover as each participant is filmed with one camera (see figure 2), it is possible to consider a precise gesture annotation and eventually an automatic post-treatment (see section 4.1).

Figure 2: Example of the scene configuration for one participant of PACO.

Figure 3: Example of the merged video for one interaction of PACO.

As a result, we obtained 30 videos of each participant in `mxf` format that we have converted in `mp4` for an easier post treatment (Elan software among others (Sloetjes and Wittenburg, 2008)). The image resolution is 1920x1080 px. A video editing software (Adobe Premiere Pro CC) was used to merge the two videos of each interlocutor of an interaction into a single one (see figure 3) in order to consider a multimodal annotation of the whole interaction. In the scope of being shared, the audio and the video files will be publicly available in the Open Resources and TOols for LANGuage (ortolang.fr) repository.

3. Enrichment of data

At the time of writing this paper, 10 dialogues out of 15 are transcribed following the subsequent procedure.

3.1. Inter-Pausal Units (IPU)

Inter-Pausal Units (IPU) are speech blocks separated by 200 ms silent pauses. The IPU segmentation has been commonly used for large corpora as it facilitates sound and transcription alignment. From the audio file of each speaker, IPUs were automatically extracted with SPPAS software (Bigi, 2015) on every dialogue of PACO 's Corpus. The following parameters were used:

- Minimum silence duration is 200 ms (a common value for French langage).
- Minimum IPU duration is 100 ms (appropriate to properly find the isolated feedbacks like “mh”).
- Shift-left the begin of the IPUs is 20 ms allows to not truncate first word starting.
- Shift-right the end of the IPUs is 20ms to not truncate last-word endings.

The resulting annotations were manually verified with Praat (Boersma and Weenink, 2018). It resulted in 30 files (15 dialogues) with the expected IPUs and 30 files with the IPUs automatically found by SPPAS. The automatic system has been satisfactorily assessed on its ability to meet the expected result (Bigi and Meunier, 2018).

3.2. Enriched Orthographic Transcription

Then the orthographic transcription was done manually on 10 dialogues of the corpus, following the Enriched Orthographic Transcription convention (Bertrand et al., 2008) and corrected by one of the authors. The transcription convention used give the opportunity to report typical oral phenomena such as filled pauses (”uh”, ”hum”), false starts, initiators, truncated words, repetitions (among others).
Those two basics levels of enrichment (IPU and transcription) allow to analyze the discursive organisation of the conversations. At the moment, only speech levels are annotated but other linguistic levels could be considered in the future. More specifically and thanks to the audio record in one channel by participant, it is possible to automatically annotate several linguistic levels with SPPAS software (Bigi, 2015) such as : tokens, phonemes, syllables and morphosyntactic categories using the MarsaTag tagger (Rauzy et al., 2014). As this study, in the longer term, focuses on the smile role during topic transitions, the whole corpus will be parsed in topic and then the frontiers of them will be annotated.
In order to analyze the conversation organization and interlocutors’ smile in regard with their common ground, the next section is dedicated to the appropriate methodology of smile annotation developed on that purpose.

4. Smile annotation protocol

Smile annotation is commonly performed manually while we are interested in a systematic and reproducible methodology for dynamic interaction. Moreover, as any other gesture, smile annotation is a time consuming task: we are then interested in reducing this annotation time. Below, we will detail the smile annotation protocol developed to consider those two methodological aspects. First, we will present the automatic tool that annotates smile in its dynamic evolution during conversation. Second, we will present the procedure to manually correct the automatic smile outputs.

4.1. The Smiling Intensity Scale

As smile is considered as a facial expression that involves physiological fine degree of changes, smile will be described thanks to the “Smiling Intensity Scale” (SIS) (Gironzetti et al., 2016). This scale describes the different smile intensity levels gradually from 0 (neutral face) to 4 (laughter), see table 1. Each smile intensity category involves a specific combination of Action Units (AUs) detailed by the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). This scale gives the opportunity to consider smile in its evolution and not anymore in a binary way (presence/absence). The interlocutors’ smiles could then be analyzed in a very fine degree of granularity which will allow to reveal the role of the different smiles in different conversational sequences.

4.2. Automatic annotation of smile with SMAD

The fields of machine learning and computer vision have experienced a rapid advance during this last decade (see for example (Martinez et al., 2019)) and it is indeed possible to automatically annotate facial gestures such as smiles. We

Neutral facial expression (S0): No smile, no flexing of the zygomaticus (no AU12), may show dimpling (AU14), but no raised side of the mouth, the mouth may be closed or open (AU25 or AU26). AU concerned : 12,14,25,26	
Closed mouth smile (S1): Flexing of the zygomaticus (AU12), may show dimpling (AU14), may show flexing of the orbicularis oculi (caused by AU6 or AU7). AU concerned : 12 (6,7,14)	
Open mouth smile (S2): Showing upper teeth (AU25), flexing of the zygomaticus (AU12), may show dimpling (AU14), may show flexing of the orbicularis oculi (caused by AU6 or AU7). AU concerned : 25,12 (14,6,7)	
Wide open mouth smile (S3): Showing lower and upper teeth (AU25), or a gap between upper and lower teeth (AU25 and AU26), flexing of the zygomaticus (AU12), may show dimpling (AU14) and flexing of the orbicularis oculi (AU6 or AU7). AU concerned : 12,6,7,25,26 (14)	
Laughing smile (S4): Jaw dropped (AU25 and AU26 or AU27), showing lower and upper teeth, flexing zygomaticus (AU12), flexing of the orbicularis oculi (AU6 or AU7), dimpling (AU14). AU concerned : 25,26,27,12,6,7,14	

Table 1: Description of the Smiling Intensity Scale (SIS) of (Gironzetti et al., 2016) for annotating smile activity. The illustrations of each level are pictures extracted from the CHEESE! corpus.

developed a tool, the SMAD[2] software (Rauzy & Amoyal, submitted to JMUI), which allows to automatically annotate a video record following the SIS system. The output of SMAD consists in a sequence of contiguous time intervals labeled by smile intensity varying from the neutral facial expression S0 to the laughing smile S4. The SMAD predictions are based on the AU intensities measured by the OpenFace software (Baltrušaitis et al., 2018). The OpenFace toolkit is an open source project which performs several tasks: head tracking, facial landmark detection, head pose estimation, facial action unit recognition and eye-gaze estimation (see figure 4 for an illustration of frame capture of an OpenFace processed video). Our automatic smile annotation tool SMAD relies on a stochastic model which has been trained on manual reliable smile annotations (following the SIS guideline) on 1 hour of videos from CHEESE!. The SMAD ouput proposes also an "X" label which identifies time areas where the prediction is made insecure because of the low confidence level associated with the OpenFace AU measurements. These events occur for example when the track of the face is temporarily interrupted due to rapid movements or face occlusions. The time intervals labeled "X" by SMAD will have to receive a manual annotation.

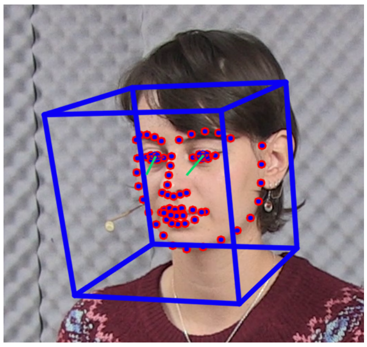

Figure 4: Example of a frame capture of the OpenFace processed video on one participant of PACO. The 68 landmark positions are pictured by the red-blue points, the head pose is traced by the projected blue cube edges, and the two green segments show the eye-gaze direction.

The SMAD software is freely available and can be downloaded at the HMAD open source project url `https://github.com/srauzy/HMAD`. SMAD proposes an output format which can be edited using the Elan software (i.e. `eaf` extension, see an illustration figure 5).

It should be noted that the experimental protocol with regard to the camera positions and the scene configuration is at this stage of primary importance. The quality of the SMAD ouput depends indeed crucially on the facial movements detected by the OpenFace software. Some specific requirements are thus to be fulfilled in order to warrant an optimal treatment (good illumination conditions, no face occlusion by the second participant, short distance between the camera and the participant, ...).

4.3. Manual correction

As any automatic annotation, it requires a manual procedure in order to check and possibly to correct the proposed outputs. Thus the automatic detection of smile ran with SMAD is followed by a manual check of the outputs and if necessary a correction. The manual correction concerns the smile intensities (from S0 to S4) and the smile boundaries. It means that the annotator has to decide for each predicted interval whether the smile intensity matches with the participant's smile and whether each smile starts and ends effectively at the time predicted. As mentioned in the

[2]The acronym SMAD stands for Smile Movement Automatic Detection.

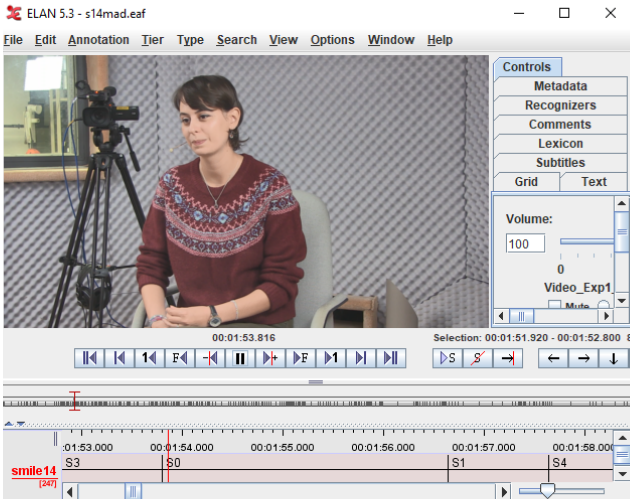

Figure 5: The automatic smile annotation generated by the SMAD tool (Elan format).

previous subsection, this correction step will also concern the "X" intervals for which the annotator will have necessarily to attribute a smile intensity or to move the predicted boundaries.

The first advantage of this protocol is the improvement in reliability and in reproducibility. Indeed, the analyze of automatic outputs compared to the one corrected reveals that the automatic tool is a good predictor as the results show that 73% of the smile where correctly predicted[3]. A second advantage of this automatic tool is the gain in annotation time. Manually correcting 10 min of video record requires 1h of work (for an expert judge of the SIS). This time spent manually correcting the automatic smile is 10 time less important than when our procedure was fully manual. Indeed, it required 10h for 10 min of video record to manually annotate smiles for a participant. The improvement of smile annotation reliability and the time saved are the two main methodological contribution of this study. This will allow to explore larger corpora as it is often the annotation cost that restraint a broader analyze.

5. Future work

The data has been enriched at several levels, and the perspective of enrichment and analysis are multiple.

The first step will be to transcribe the 5 interactions left in the PACO Corpus. Based on theses transcriptions, an analysis of hetero-repetition could be performed using SPPAS software (Bigi, 2015) and would allow to highlight if interlocutors use this linguistic process to get acquainted. Every topic discussed will be parsed, based on the audio signal and the transcription and in order to analyze the discursive organization of the interactions. We will then analyze the type of topic transitions used by the interlocutors. Based on the video records, automatic annotation of smiling will be carried out according to our smile annotation protocol (see section 4). Thanks to the SMAD model, the automatic annotation will be done followed by a manual correction of those smiles. Those corrected and reliable smile annotations will then enrich our training corpus which will lead to improve the robustness of the SMAD model. Topic transitions and smile evolution will be then compared in the two conditions : when participants know each other well (Cheese!) versus when they meet for the first time (PACO). In the perspective of an open science, every further enrichment and analysis on PACO will be soon publicly available.

[3] For comparison, the evaluation performed on our training corpus reveals an observed agreement of 68% between the fully manual annotations and the automatic SMAD outputs (Rauzy & Amoyal, submitted to JMUI).

6. Conclusions and perspectives

PACO is a 5h audio video corpus created in order to compare the conversations organization and the smile of interlocutors that know each other well (Cheese!) to interlocutors that just met each other (PACO). It has been shown that the experimental design is optimal in several aspects. First, the good quality audio file per participant allows an easier and reliable transcription. Second, the video record per participant allows a post-treatment, specifically using the SMAD model. Third, the double instructions given to the participants is a strong strength as it provides the same linguistic material for every participant of every condition (with/without CG) and it also enables participants to discuss as freely as they want without any topic or task constraint. Fourthly, the controlled factor "Common Ground" will enable a comparison all things being equal. This data set will allow to compare the smiles of the interlocutors according to whether they know each other or not. Thus, we would like to highlight the possibility that this facial gesture may depend on the interlocutors' relationship. We will therefore conduct an analysis of the duration of the different smiles in the two corpora as well as their location in the conversation.

Concerning the data enrichment, we have pointed out several methodological contributions. This semi-automatic smile annotation protocol is a reliable and time saving solution for any analysts that are interested in smile during a dynamic video. The training corpus of the SMAD model will be enriched so the model could be even more robust. Further perspectives could also be engaged. As every participant is recorded in one channel, many other annotation levels of the speech signal could be analyzed. For example, a prosody analysis could be performed as everyone read the same text. The video recording is of great interest for everyone interested in facial mimicry but also hand gestures studies and postural studies. Finally, without being exhaustive on the possibilities given by this corpus, the video records could constitute the material of a perception study of the different roles of smile.

In the perspective of an open science, the PACO corpus and the SMAD model are both in open access.

7. Acknowledgements

Data were filmed and recorded in the soundproof room of the CEP experimental platform which allows to gather a wide range of audio and video data in its anechoic room (LPL, AMU-CNRS, Aix-en-Provence, France). A special

thanks goes to Carine André who managed the audio video collection and treatment.

8. Bibliographical References

Amoyal, M. and Priego-Valverde, B. (2019). Smiling for negociating topic transitions in French conversation. In *Gesture and Speech in Interaction, GeSpIn 2019*, Paderborn, Germany, September.

Attardo, S., Pickering, L., and Baker, A. (2011). Prosodic and multimodal markers of humor in conversation. *Pragmatics & Cognition*, 19(2):224–247.

Aubergé, V. and Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication*, 40(1-2):87–97.

Baltrušaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, May.

Bavelas, J. B. and Gerwing, J. (2007). Conversational hand gestures and facial displays in face-to-face dialogue. *Social communication*, pages 283–308.

Berthoud, A.-C. and Mondada, L. (1995). Traitement du topic, processus énonciatifs et séquences conversationnelles. *Cahiers de linguistique française*, 17:205–228.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., and Rauzy, S. (2008). Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, 49(3):pp–105.

Bigi, B. and Meunier, C. (2018). Automatic segmentation of spontaneous speech. *Revista de Estudos da Linguagem. International Thematic Issue: Speech Segmentation*, 26(4).

Bigi, B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, 111–112:54–69.

Boersma, P. and Weenink, D. (2018). Praat: doing phonetics by computer [computer program], version 6.0.37, retrieved 14 march 2018 from http://www.praat.org/.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Ekman, P. and Friesen, W. (1975). *Unmasking the face : a guide to recognizing emotions from facial clues*. Englewood Cliffs : Prentice-hall.

Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System: Manual*. Number vol. 1 à 2. Consulting Psychologists Press.

El Haddad, K., Chakravarthula, S. N., and Kennedy, J. (2019). Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles. In *2019 International Conference on Multimodal Interaction*, pages 259–263. ACM.

Fagel, S. (2010). Effects of smiling on articulation: Lips, larynx and acoustics. In *Development of multimodal interfaces: active listening and synchrony*, pages 294–303. Springer.

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. Academic Press.

Garcia, L. J. and Joanette, Y. (1997). Analysis of conversational topic shifts: A multiple case study. *Brain and Language*, 58(1):92–114.

Gironzetti, E., Attardo, S., and Pickering, L. (2016). Smiling, gaze, and humor in conversation: A pilot study. In Leonor Ruiz-Gurillo, editor, *Metapragmatics of Humor: Current research trends*, pages 235 – 254.

Golopentja, S. (1988). Interaction et histoire conversationnelle. *Échanges sur la conversation. Paris: CNRS*, pages 69–81.

Holler, J. and Bavelas, J. (2017). Multi-modal communication of common ground. *Why gesture*, pages 213–240.

Martinez, B., Valstar, M., Jiang, B., and Pantic, M. (2019). Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 10(3):325–347.

Priego-Valverde, B., Bigi, B., Attardo, S., Pickering, L., and Gironzetti, E. (2018). Is smiling during humor so obvious? A cross-cultural comparison of smiling behavior in humorous sequences in American English and French interactions. *Intercultural Pragmatics*.

Rauzy, S., Montcheuil, G., and Blache, P. (2014). MarsaTag, a tagger for French written texts and speech transcriptions. In *Second Asian Pacific Corpus linguistics Conference*, pages 220–220, Hong Kong, China.

Riou, M. (2015). A methodology for the identification of topic transitions in interaction. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (16).

Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & psychophysics*, 27(1):24–27.