SMYLE: A new multimodal resource of talk-in-interaction including neuro-physiological signal

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Matthis Houlès, Thierry Legou, Magalie Ochs, Philippe Blache

SMYLE: A new multimodal resource of talk-in-interaction including neuro-physiological signal

Auriane Boudin
Institute of Language Communication and the Brain
Laboratoire Parole et Langage
Laboratoire d'Informatique et Systèmes
Aix-en-Provence, France
auriane.boudin@univ-amu.fr

Roxane Bertrand
Laboratoire Parole et Langage
Institute of Language Communication and the Brain
Aix-en-Provence, France
roxane.bertrand@univ-amu.fr

Stéphane Rauzy
Laboratoire Parole et Langage
Institute of Language Communication and the Brain
Aix-en-Provence, France
stephane.rauzy@univ-amu.fr

Matthis Houlès
Institute of Language Communication and the Brain
IMT Atlantique
Aix-en-Provence, Nantes, France
matthis.houles@gmail.com

Thierry Legou
Institute of Language Communication and the Brain
Laboratoire Parole et Langage
Aix-en-Provence, France
thierry.legou@univ-amu.fr

Magalie Ochs
Institute of Language Communication and the Brain
Laboratoire d'Informatique et Systèmes
Marseille, France
magalie.ochs@lis-lab.fr

Philippe Blache
Institute of Language Communication and the Brain
Laboratoire Parole et Langage
Aix-en-Provence, France
blache@ilcb.fr

ABSTRACT

This article presents the SMYLE corpus, the first multimodal corpus in French (16h) including neuro-physiological data from 60 participants engaged in face-to-face storytelling (8.2h) and free conversation tasks (7.8h). The originality of this corpus lies first in the fact that it bears all modalities, precisely synchronized and second in the addition for the first time at this scale of neuro-physiological modalities. It constitutes the first corpus of this size offering the opportunity to investigate cognitive characteristics of spontaneous conversation including at the brain level. The storytelling task comprises two conditions: a storyteller talking with a "normal" or a "distracted" listener. Contrasting normal and disrupted conversations allows to study at a behavioral, linguistic and cognitive levels the complex characteristics and organization of conversations.

In this article, we present first the methodology developed to acquire and synchronize the different sources and types of signal. In a second part, we detail the large set of automatic, semi-automatic and manual annotations of the complete dataset. In a last section, we illustrate one application of the corpus by providing preliminary analyses of the annotated data, that reveal the impact of distracted listener's on his/her feedbacks and the quality of the narration.

ICMI '23 Companion, October 9–13, 2023, Paris, France

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3610661.3616188

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Multimodal dataset, Interaction, Automatic annotation

ACM Reference Format:
Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Matthis Houlès, Thierry Legou, Magalie Ochs, and Philippe Blache. 2023. SMYLE: A new multimodal resource of talk-in-interaction including neuro-physiological signal. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3610661.3616188

1 INTRODUCTION

Conversation is a spontaneous activity, based on implicit rules, which make it (almost always) successful. However, what are the mechanisms involved in making it effective? What do the interlocutors do to understand each other? What happens when some of these mechanisms are disrupted?

Multimodal datasets of spontaneous conversations are essential for improving our understanding of human communication. Although several such datasets exist, they often lack comprehensive

annotations at all levels, which is due to the significant resources required to acquire them. Furthermore, to study cognitive processes during conversation, access to brain signals is crucial, yet this has been minimally explored in the context of spontaneous conversations. Therefore, it is essential to develop large-scale datasets of spontaneous conversations with long exchanges and a diverse group of participants. In addition, detailed annotations of these datasets must be made available to enable research in several areas, including natural language processing, speech recognition, and cognitive neuroscience.

This paper aims to present the precise experimental setup, specifically developed to collect this dataset (see 3.3). In terms of instrumentation, we used specific devices for the different modalities, including Empatica recorders for physiological signal and an EEG hyperscanning setup with 64-channels Biosemi system.

The paper is structured as follows: we first present the aims and scope of this work including the related existing resources in spontaneous conversations proposing the same set of multimodal data. The methodological section presents the experimental setup, participants, tasks and procedure. It also describes the instrumentation and the specific question of data synchronisation. The 4 section presents the automatic and semi-automatic tools used to perform the verbal, vocal and visual annotations as well as their manual corrections. We present finally some preliminary investigations.

2 AIMS AND SCOPE

Many multimodal corpora already exist. However, none of them gather all possible modalities, recorded in natural interactions, in a sufficient size, and including neuro-physiological signal. We briefly present in this section the main goals of such corpora and gives a focus on the particular question of datasets including the brain signal in conversation.

2.1 A rich multimodal synchronized dataset for studying conversations

In several models [21, 22, 32] interlocutors achieve a common understanding by aligning their behavior and linguistic representations. Dialogues are therefore considered as a collaborative effort between the participants both at the production and perception levels. As a result, phonetic, syntactic, semantic and pragmatic information progressively becomes aligned. The hypothesis is that listeners actively process simultaneously these sources of information in order to understand the main speaker's production and be prepared to produce a feedback[1] (or take the turn) appropriately. However, it should be noted that this hypothesis has yet to be definitively proven through empirical research during spontaneous conversations.

In recent years, many works have focused on the quality and the success of conversational interactions. In particular, these studies have addressed a certain level of synchronization between participants during a joint action, including at the brain level [10, 25, 30]. Nevertheless, only few works have been done using corpora during which the dynamics of the interaction has been analyzed from audio, video and neurophysiological perspectives, including by altering it.

[1]Feedback, also referred to as a backchannel, is defined as the reactions produced by a listener in response to the main speaker's discourse. [36, 38].

We present in this paper the corpus *"Show Me You are Listening"* (SMYLE), a first audio-visual and neuro-physiological dataset of 30 spontaneous face-to-face interactions in French. The dataset is composed of two conversational tasks: a controlled narrative one and a free discussion. Inspired from [5], the first storytelling task is conducted under two conditions: 1/ a *control* one involving the storyteller and an *attentive listener* and 2/ a *distracted* condition involving the storyteller and a *distracted listener*. The second part of SMYLE is made of free conversations, no specific instructions being given to the interlocutors. This part is important by offering the possibility to compare controlled and unrestricted productions.

SMYLE has been initially designed to study both the production and perception of speakers in their discursive role, as well as the quality of conversations as a function of feedback production. As proactive reactions produced by listeners, feedbacks play a crucial role in structuring conversations and promoting alignment between interlocutors. They involve both backward and forward action: feedbacks are triggered by the previous context (i.e. multimodal cues from the main speaker) [16, 17, 28, 29] and have an impact on the dynamics of the conversation [6, 9, 21].

However, SMYLE goes far beyond the study of feedbacks by providing an extremely rich source of data for studying a wide range of conversational phenomena.

2.2 Complete multimodality: inclusion of the brain signal

In recent years, there has been a growing interest for studying neural synchrony between individuals during social interactions, to understand how speakers reach a common understanding (see [26] for a detailed review).

Focusing more specifically on social aspects, [27] compared brain-to-brain synchrony between romantic couples and strangers, studying male-female interactions using a hyperscanning EEG setup. A 32-electrodes cap was used to record the cerebral signal for each participants and video signal with one adjacent camera. The study included 52 romantic couples and 52 strangers dyads, who were asked to plan a fun day to spend together during spontaneous conversation while seated face-to-face. A significant brain-to-brain synchrony in gamma frequency-bands localized in the temporal-parietal area has been found among couples but not for strangers dyads. The neural synchrony was observed during moments of social gaze and positive affect but appeared to be unrelated to speech/non speech activity. While this corpus has a large number of speakers, the conversations it contains are relatively short, lasting only about 5 minutes.

Creating a full multimodal dataset is therefore crucial for examining the mechanisms involved in spontaneous and natural conversations. To address this need, [12] developed a new protocol for recording audio-visual and neuro-physiological data. This corpus involved 5 Spanish dyads engaged in spontaneous conversations and recorded in 3 sessions, each lasting 45 minutes and spaced four days apart. During the 3 sessions, participants were given a creative task and a moral dilemma to discuss and potentially resolve. In total, 12.30 hours of data were collected, and several automatic annotations were performed from the audio and video signal. These annotations include transcription, token alignment,

phonemes alignment, part of speech, nods and smile. Nonetheless, at this stage, annotations have not been corrected and the number of speakers is limited.

At a more specific level, [31] successfully detected inter-brain neural coupling during naturalistic interactions during non-visual interactions, with no overlap between the speaker and the listener. Additionally, [27] observed a neural synchrony only during phases of prolonged social gaze between couples but not during periods without. These studies differ in terms of the tasks performed, modalities recorded, number of subjects and duration of the interactions. Taken together, the literature suggests that there is a complex relationship between brain-to-brain synchrony and the characteristics of the social interaction, and that further research is needed to elucidate these relationships.

3 THE SMYLE CORPUS

The goal of SMYLE is to address unexplored aspects of conversation such as the cognitive processes involved to build a common ground (the shared knowledge between participants [19, 24]) and achieve mutual understanding. [5] found that a decrease in the feedback frequency is correlated to the quality of the storytelling. [5, 8, 37] show that listeners provide different types of responses depending on the sequential organization of the narrative. However, some aspects of social communication remain poorly described or subject to debate in the literature. The specific ways in which the flow of information is transmitted, processed and integrated by the interlocutors in the conversation, as well as the necessary attentional mechanisms, remain unclear.

SMYLE presents each pair of participants in both controlled and uncontrolled context with pre-defined discursive roles for the narration but free ones for conversation. The experimental design and protocol allows us to question the level of predictability and automaticity of feedbacks and to compute predictive models of feedback and turn-taking organization. The corpus is available to the scientific community on https://hdl.handle.net/11403/smyle [15].

3.1 An original task for disturbing the listener behavior

In their work, [5] proposed an original experiment to examine the role of the listener by contrasting attentive and distracted listeners. Thirty-four dyads of participants were recorded during a face-to-face storytelling of a near-miss accident. Half of the dyads were in normal condition, where the listener has to summarize the story. The other half of the dyads were in distracted condition, where a *t-counting task* was give to the listener without the storyteller being aware of it. The t-counting task, consisted of counting words uttered with a *t* as a first letter. The quality of the story ending has then been assessed by third-person analysis.

Distracted listeners gave significantly less feedback than attentive listeners, which had an impact on the quality of the story. This effect is significantly demonstrated by the story ending evaluation scores.

One limitation of the present study is that only behavioral data can be analyzed, as there is no accompanying neural signal data. This lack of neural data precludes the investigation of the cognitive impact on speakers when conversations are disrupted. Additionally, automaticity and predictability aspects of conversations could not be investigated.

3.2 Participants

Sixty participants took part in the experiment (mean age = 22.77, sd = 3.29 , min = 18, max = 36). Forty-three participants were female and 17 participants were male. Fifty-four participants were students of different levels and fields, five were employed and one was unemployed. Participants were recruited from the Aix-Marseille University and from the mailing lists of Laboratoire Parole et Langage. All participants were native French speakers, right-handed, and reported no neurological or language disorders. The experiment was conducted in November at Laboratoire Parole et Langage at Aix-en-Provence, France. The participants received a compensation of 30€. None of the participant dyads knew each other before the experiment. Sixty participants were recorded in all modalities, except one participant with no EEG data and one storyteller with no video for Task 1 due to recording failure, both in the normal condition.

3.3 Experimental set-up and equipment

The participants were installed in a soundproof room and seated face-to-face on chairs place 130 centimeters apart. To ensure a frontal view of the participants, the chairs were shifted by 30 centimeters to the left for participant A and to the right for participant B. This allowed the cameras (CANON XF105) to be positioned almost directly behind the chair for a full frontal framing, as show in figure 2. The cameras recorded 5-minutes rushes in .mxf format at a resolution of 1920x1080 and 25 fps. To capture clear audio, each participant was fitted with a headset microphone (AKG-C520), which was connected via XLR to the faceplate of the soundproof room and fed back to the audio-computer in the control room using a RME FireFace UC (the same sound configuration than [2, 33]). The gain for each participant was adjustable using the Fireface UC and the signal was monitored in real-time using Audacity. The sound was then routed back to the respective cameras of each participant via XLR cables. Both speakers' microphones were recorded in stereo on the same .wav file with a sampling rate of 48,000 Hz and 16 bits signed PCM per sample. To ensure a neutral background for analysis with automated software like OpenFace or FaceWare, a green background was placed behind the chairs and the chairs was also covered themselves. This also helped to hide the camera and make the participants feel more comfortable during the experiment. To provide optimal lighting, 3 spotlights were used, with two small one directed towards each participant and a larger one directed towards the center. Next to each participant, the EEG AD-BOX was placed on a small table. In the center of the two tables, the two AD-BOXES were connected to a beacon which was in turn connected via USB to the control room, see figure 3.

The electrophysiological signal was recorded using two 64 Biosemi active electrodes with a standard 10/20 positioning. Both Biosemi systems were configured for hyperscanning of two participants with a temporal resolution for signal acquisition of 2048 Hz. Three additional electrodes were used. Two on the left and right mastoids for referencing and one under the left eye to monitor eye blinks.

Different caps were used depending on the size of the participant's head. Electrode impedances were controlled and the signal was recorded using ActiView software. The physiological signal was recorded using two Empatica E4 wristbands. The E4 is equipped with various sensors (PPG sensor, EDA sensor, 3-axis accelerometer, Infrared Thermopile) that measures Blood Volume Pulse (BVP), heart rate, electrodermal fluctuation, peripheral skin temperature and motion-based activity. The E4 also has an event marker button that allows for signal synchronization.

Figure 1: Picture of the set-up with Participant A (listener) on the left and Participant B (main speaker) on the right.

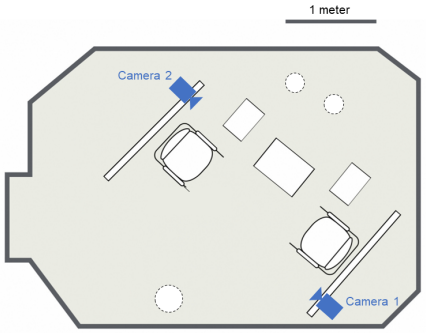

Figure 2: A plan of the soundproof room installation with the position and distance of the chairs, cameras and the lights.

3.4 Signal synchronisation

Recording multimodal and dyadic corpus involves to synchronize all audio, video, EEG and physiological signals together, both within and between the two speakers. To achieve this, we used an Arduino prototyping to generate a visual, auditory and EEG trigger on a beacon for signal synchronization. The two cameras were synchronized using a synchronisation cable and a supplementary movie clapper was used at the beginning of the session recording. We used Sony Vegas Pro 18.0 software to assemble the rushes from each session and to merge the videos from both participants into one. The videos of the two speakers were automatically synchronized based on the time code. We then imported the audio recorded on Audacity from the audio control computer and manually aligned it to the video using the sound wave of the clapper. We used the sound recorded on Audacity from the audio computer due to its better quality. The videos were then trimmed to keep only tasks 1 and 2 based on the visual trigger. The EEG systems are synchronised with a daisy-chain and recorded in a single .bdf file. The Empatica E4s are synchronized based on the visual trigger generated by the wristbands.

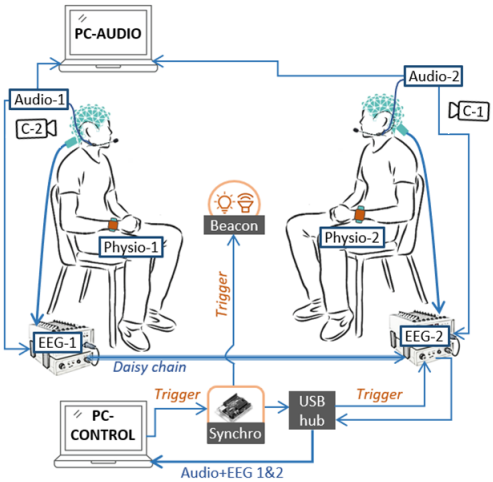

Figure 3: Audio, video, EEG and physiological signal recording and synchronisation. Participant A, on the left, is recorded by Camera-1, Audio-1, Physio-1, EEG-1. Participant B, on the right, is recorded by Camera-2, Audio-2, Physio-2, EEG-2. The EEG of both participant is synchronized thanks to an Arduino prototyping and send to the PC-CONTROL via USB. Audio of both participant is synchronized on the PC-Audio and send back to the cameras.

3.5 Tasks

Task 1. The first task (mean duration = 17,49 min, sd = 8,06 min) consists of a storytelling with two conditions, a control condition and a distracted condition. Fifteen dyads were in control condition and 15 dyads in distracted condition. One participant had the role of the storyteller and the other had the role of the listener. Each participant is randomly assigned a specific discursive role, either storyteller or listener and remains in that role throughout the duration of the first task. The condition for each dyad was randomly defined.

Storyteller instructions. The instructions given to the storytellers were the same in both conditions. The storytelling consists of 3 stories to be told. We chose to ask for 3 stories to avoid too short interactions. The first story consists of telling the video clip of the pear story [18], which lasts 5 minutes 55. We ask them to make the story as interesting as possible. For the second story, we ask the participants to tell the pitch of movie, TV show, book or video game that they feel is not well known enough, to limit the possibility that the other participant already knows it. We ask the narrator to tell the story in a way that makes the other participant want to watch, read or play what has been presented. For the third and

final story, the narrator is asked to tell their favorite vacation. We ask the storyteller to tell the stories one after the other and simply indicate when he/she has finished the last one.

Normal listener instructions. Concerning the role of the listener in the normal condition, we inform him/her that his/her partner will tell 3 stories. We tell the participants to listen carefully and that he/she can freely react, speak, ask questions during the storytelling. We warn them that they have to summarize the stories quickly at the end of the experiment.

Distracted listener instructions. For the distracted condition, we give the participant the same information, but we ask them to count all the words produced by the other participant that begin with a /*t*/, and to press a pressure plate with their foot. Unlike [5], we decided to used the foot rather than the hand in order to preserve hand movements during the conversations. The said plate was actually a trick to encourage them to do the task well. We told them that the other participant should not discover this hidden task.

Task 2. For the second task (30 dyads), we simply ask participants to talk freely for 15 minutes (mean duration = 15,31 min, sd = 3,03 min). We ask them to begin the conversation by a debriefing of the first task. Distracted listeners can reveal their hidden task to the other participant.

Post-experience questionnaires. At the end of Task 1, participants completed an online questionnaire, on FindingFive, from their own phones to self-reported their perception of the experiment. Participants remained seated in the soundproof room, but we asked them to not interact about the experiment until Task 2. Information on age, gender and profession was collected. The storyteller was asked to self-evaluate his/her storytelling, engagement in the conversation and its perceived engagement of the listener. All responses are 5-level Lickert scales. We also ask some additional yes/no questions about the other participant's reactions during the stories.

For the listener questionnaire, yes/no questions about their reactions and 5-point Lickert scale question were asked to assess the quality of the other participants' narration and their engagement in the conversation. They next have to summarize each story in approximately 5 sentences.

For the distracted condition, an additional part is included in the first part of the questionnaire. We asked them how many words they counted, how successful they were at the task and how difficult it was.

3.6 Procedure

Participants were separated into two different rooms. They were first informed of their rights and signed a consent form. Each participant was informed that the purpose of the study is to better understand spontaneous conversations between strangers. While the equipment was being set up, the participants read instructions on a sheet of paper. One of the experimenters then repeated the instructions orally and answered any questions. The microphone was installed before fitting the EEG cap. The storyteller watched the video of the pear story while the EEG was installed. Once the participants were equipped, both moved into the anechoic room. We make sure they do not get to know each other before the experiment begins. After setting camera framing and the microphone gain, EEG signal was verified. Participants were instructed to begin only after the beacon signal and that the first task will end after the same signal. Then participants took a short break and completed the online questionnaire. The second task starts and ends also after the beacon signal. The total experiment duration was approximately 2 hours.

4 ANNOTATIONS

Several levels of annotations are performed on the corpus. In order to reduce as much as possible the manual annotations, we have used several software programs to perform automatic annotations. So far, all automatic annotations have been performed and manual corrections are in progress (see table 1). The manual annotations and corrections are performed by 3 expert annotators.

4.1 Automatic Annotation

Transcription. We developed a software allowing *(i)* the segmentation of the audio signal into Inter Pausal Units and *(ii)* within which the orthographic transcription was performed. The code is available at `https://github.com/MatthisHoules/ASR-Pipeline`.

i. IPU segmentation is done by computing the Root Mean Square (RMS) value of the audio signal on intervals of 32 ms. Each IPU is bounded by a pause of at least 200 ms. The Speech/Silence threshold was determined from the RMS value of the background noise (mainly due to the other participant's voice).

ii. Transcription is done by using a Wav2Vec2 [3] model trained on 7.6K hours of French audio speech [20] and fine-tuned on 2.2K hours of French audio speech. The output of the model is decoded with a 5-gram Language Model. The model and the language model used are both available on `Hugging Face`.

Aligned annotations. We used SPPAS software [11] to perform several annotations aligned onto the speech signal from the manually corrected transcription, including tokens, syllabification, phonetization, speech activity (speech, silence or laughter), speech activity overlap, self-repetition and other repetition.

Part-of-Speech. The morphosyntactic information is obtained by running the MarsaTag POS-tagger [35] on the manually corrected transcription. For spoken French, the MarsaTag system has been trained on the corpus CID [7] (8-hours of French dyadic conversations) and allows to manage phenomena proper to spontaneous speech such as filled pause, disfluency and truncation [1].

Prosodic Tones. Tones extraction is done automatically thanks to the pitch modeling tool MOMEL-INTSINT [23]. The anchor points are automatically encoded into an alphabet of tonal symbols *T(op), B(ottom), M(id)* referring to absolute values and *H(igher), L(ower), S(ame), U(pstepped), D(ownstepped)* referring to relative values. This encoding provides intonation patterns represented by the key/midpoint and the span of the speaker's pitch range.

Smiles. The smile intensities are automatically annotated using the SMAD tool [34] (downloadable at the open source project url `https://github.com/srauzy/HMAD`). The video is in a first stage treated by the OpenFace software [4] which tracks along time the

face of the participant, measures head pose, landmark positions and gaze direction and detect some of the facial Action Units of the participant. In a second stage SMAD annotates smile intervals following a 5 levels scale. The smile interval labels and boundaries are predicted from the intensities of the facial Action Units outputted by OpenFace. For the present study the SMAD output is finally transformed in a 3 level smiles scale (*NF*, *LI*, *HI*) with *NF* encoding Neutral Face, *LI* Low Intensity smiles (smiles with mouth closed) and *HI* High Intensity smiles (smiles with mouth opened), as proposed in [14].

Gaze and Blink. Gaze and blink are automatically annotated using the BAGMAD tool (Rauzy et al., in preparation) which relies on the gaze direction and AU blink intensity measurements proposed by the OpenFace output. The BAGMAD software proposes two levels of annotation. A blink tier annotates blink events, i.e. rapid, semi-automatic, physiological eyelid closing. A second tier annotates the video in time intervals with a binary label describing whether the participant: 1) look at the other participant, *eye contact* label 2) look away from the other participant or close its eyes, *no eye contact* label. In this scheme, a eyelid closing longer than 400 ms is not considered as a blink but rather as a *no eye contact* label area.

4.2 Manual Corrections

Transcriptions. IPU segmentation and orthographic transcription are manually corrected using Praat software. Following [13], supplementary information were added to the transcription: laughter, laughing pronunciation, repetitions, disfluences, broken words, personal information, specific pronunciation and elision.

Gaze. The gaze correction task include the verification and correction of the time boundaries and of the label (eye contact or no eye contact) using ELAN software. Blink are not corrected. The same annotators team were previously trained on another corpus where they obtained a Fleiss Kappa of 0.92 (almost perfect agreement). Consequently, the annotations have been independently distributed among the annotators.

4.3 Manual Annotations

Narrator Rating. Similarly as [5], we want to evaluate the quality of the storytelling to compare the two conditions. Nonetheless, [5] focuses on the analysis of the quality of the end of the storytelling. Inspired by [5], we define a method for evaluating narrative quality. We have chosen to evaluate the whole story given that the type of story is different and that the ending of our stories is not very salient. The annotators rate storytelling for each story and for each storyteller by giving a score from 0 (not satisfying) to 3 (very satisfying) for 6 criteria:

- Level of detail: 0 (missing detailed or far too detailed), 1 (some unnecessary details or lack of relevant details), 2 (good level of details but lacks a little/few too much), 3 (perfectly measured, relevant details, enough but not too much). κ = 0.27.
- Clearness: 0 (the story is not clear or disjointed), 1 (some elements are understood but most are not), 2 (the story is almost always clear), 3 (everything is perfectly clear). κ = 0.33.
- Story ending: 0 (abrupt end or never-ending end) 1 (end told too fast or too long) 2 (correct ending), 3 (perfect ending). κ = 0.32.
- Rhythm: 0 (much too slow or much too fast), 1 (a little too slow or a little too fast), 2 (overall good pace but some parts with a bad pace), 3 (good pace throughout the story, neither too fast nor too slow). κ = 0.15.
- Interest in the story: 0 (not interesting), 1 (somewhat interesting), 2 (interesting), 3 (very interesting). κ = 0.23.
- Comfort of the speaker: 0 (not at all comfortable), 1 (little comfortable), 2 (comfortable), 3 (very comfortable). κ = 0.60.

Given the high subjectivity of this kind of annotation, the raters annotated all of the 29 storytellers. This annotation was performed with only the audio and video of the storyteller on ELAN software. The condition (normal or distraction) were hidden to the raters. We instructed them to watch the video only once and to attribute a score for each criteria after each story. This is the first annotation task we give them before they get used to the corpus.

Head Movements. Head movements are manually coded by the annotators. The following movements are annotated:

- Nod: simple (single nod) or complex (several repetitions) down and up vertical head movement.
- Shake: simple or complex horizontal left to right movement of the head.
- Tilt: simple or complex head tilt.
- Other: combination of several movements at the same time or that follow each other very quickly.

The annotators rated 5 minutes of 3 storytellers and 3 listeners to compute their agreement. The Fleiss Kappa obtain for all speakers is 0.886 (almost perfect agreement), for the listener the Kappa is 0.938 and for the main speaker 0.807. Given the high agreement, annotators then annotated participants individually. The annotation of a main speaker's head movements is much more ambiguous than for a speaker in listening position, since there are many more movements and their function may be less clear than in listening position.

Feedback. Feedback have been manually annotated by one of the author for 11 listeners for both Task 1 and Task 2 including 6 in distracted condition and 5 in normal condition. Feedback is annotated in generic and specific categories on ELAN software. We consider as a feedback, every reactions produced by one speaker to the production of the other speaker. We do not consider answers to an explicit question as feedback. These reactions can be vocal, verbal or mimo-gestural. Vocal reactions are mostly laughter. Concerning verbal feedback, we do not set a limit on the length of the utterance, but we consider all utterances that react to the previous production of the other speaker. Finally, wince, eyebrows movements, hand movements, head movements, smiles are also annotated as feedback. Most of the time, it is the combination of several elements that composed the feedback. The type tagging is based on the definition between generic and specific feedback of [5]. Generic feedbacks refer to a response not directly related to the content of the narrator's speech. This response is conveyed

Table 1: This table presents the main annotations: IPU, Token, Gaze, No Gaze, Nod, Shake, Tilt, and Other category of head movements. The table includes the number of subjects annotated, the total number of items annotated, the average duration in seconds and the frequency per minute for each category of annotation.

Item	Subjects	Total	Mean dur. (s)	Frequency
IPU	54	26753	1.69 ± 0.47	15.21 ± 4.08
Token	54	195583	0.23 ± 0.03	111.34 ± 50.59
Gaze	56	16562	8.22 ± 10.38	8.94 ± 4.88
No Gaze	56	16483	1.82 ± 0.79	8.90 ± 4.89
Nod	32	7438	0.93 ± 0.21	7.39 ± 3.03
Shake	32	2316	1.03 ± 0.22	2.23 ± 1.43
Tilt	32	1021	0.67 ± 0.15	1 ± 0.67
Other	32	306	1.12 ± 0.32	0.27 ± 0.36

through short vocalizations (*"mh mh", "ok", etc.*) and/or by nodding. The main function of such feedbacks is to help the main speaker in monitoring the interlocutor's comprehension. In contrast, specific feedbacks help the speaker to tell a story by displaying a range of behaviors (happiness, sadness, surprise, etc.). These responses can include verbal/vocal and gestural content (wince, smiling, laughing, hand movements, head movement, etc).

5 PRELIMINARY RESULTS

To assess the reliability of our corpus and replicate the findings of [5], we conducted a preliminary analysis of narrator scores for all storytellers and feedback frequency and duration for a subset of 11 listeners, comprising 6 distracted and 5 attentive individuals. Additionally, we analysed the self-report questionnaires of the storyteller regarding their perception of the experiment.

Explicit marks of the common ground elaboration and promoters of alignment, feedbacks are thus one of the major phenomenon to better understand how participants elaborate meaning together and understand each other. We hypothesize that distracted listeners will provide altered feedback, with changes in both the type (generic and specific) and frequency of feedback. Despite the distraction, feedback may still be triggered by low-level cues from the main speaker and produced almost automatically. Our main hypotheses are that 1/ the narrator quality score will be lower in the distraction condition, and 2/ there will be a decrease in the number of specific feedbacks provided by distracted listeners. Table 1 presents general information about the annotations conducted.

5.1 Narrator Score

As described in section 4.3, the annotators evaluated the quality of storyteller's three stories. Six criteria are considered: clearness, details, story ending, rhythm, comfort of each speakers and the personal interest of the raters in listening to the story. We average the score coded by the annotators for each criterion. Then we compute an overall score for each speaker by adding up each average score. This gives us the overall quality of the narration for all 29 speakers and for each story, which can range between 0 and 18. The average score obtained for the 3 stories combined is 12.97 (sd = 3.10, min = 6, max = 18). The t-test showed no differences between Story 1/Story 2 ; Story 1/Story 3 nor Story 2/Story 3. For this reason, we decided to keep the general score obtained for the 3 stories. Finally, we performed a two-sample t-test to compare the speaker's scores in normal condition (M = 14.21) and distracted condition (M = 11.99). The results revealed a significant effect (p = 0.0005) with a medium effect size, Cohen's d = -0.77.

5.2 Feedback

Feedback frequency. We examine the frequency per minute of generic and specific feedback (see 4.3) in normal and distracted conditions, i.e. $f = n/(tmax * 60)$

with n the number of feedback and $tmax$ the task duration in seconds.

In order to correct the feedback production frequency from listener effect (i.e. the high variability between listeners on feedback frequency), we make use of the listener feedback frequency observed in the free conversation task. For each listener, we compute Δ the logarithm of the ratio of feedback frequency between the storytelling task and the free conversation task.

$$\Delta = \log(F_{T1_i}) - \log(F_{T2_i}) \quad (1)$$

In this equation the frequency for the free conversation task is corrected from the main speaker role effect (i.e. when the listener is actually the main speaker and consequently does not produce feedback). The $tmax$ of free conversations is indeed obtained by subtracting from the total $tmax$ all the video frames in which the listener has produced more speech in the previous 5 seconds than the main speaker, except in the case of feedback production.

We next performed a two sample t-test to compare the frequency of generic and specific feedback by condition. The results show no significant differences for generic feedback (p-value = 0.076), but a significant effect for specific feedback (p-value = 0.037) with a medium effect size, Cohen's d = -0.61. There is 40% less specific feedback in distracted condition than in normal condition.

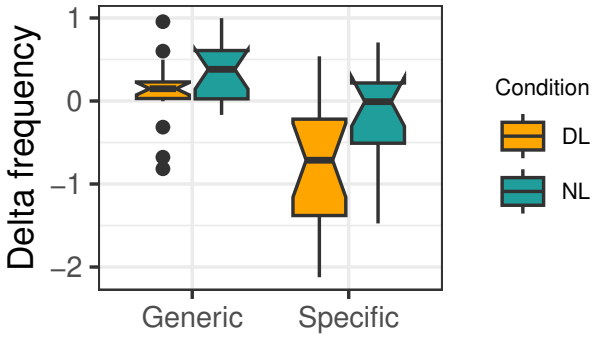

Figure 4: The mean corrected frequency of generic and specific feedback by condition: Distracted Listener (DL), Normal Listener (NL). The corrected frequency is calculated based on the delta of feedback frequency between the free conversation (task 2) and during the storytelling (task 1), as describe by the 1 formula.

Table 2 presents the generic and specific frequencies per minute obtain by task performed.

Table 2: Frequency of feedback per minute for generic and specific feedback type in normal condition (NC) and distracted condition (DC) for the storytelling task and the free conversation task (FC).

Feedback Type	NC Frequency	DC Frequency	FC Frequency
Generic	8.18 ± 2.56	8.17 ± 4.22	6.84 ± 3.52
Specific	4 ± 1.76	2.45 ± 1.90	4.02 ± 1.20

Nonetheless, we found a weak Pearson correlation between feedback frequency and storytelling quality, $r = 0.17$ for generic feedback and $r = 0.25$ for specific feedback.

Feedback duration. A total of 2892 feedback were annotated, of which 1933 were generic and 959 specific. The average duration of generic feedback is 0.93 ms (sd = 0.51, min = 0.098, max = 4.96). The average duration of specific feedback is 1.53 ms (sd = 1.18, min = 0.19, max = 12.34). When we compared the average duration of feedback produced during the first task between the normal and the distracted condition, we found that generic feedback is 18% shorter in distracted condition than in normal condition (t-test shows a p-value < 0.001). We did not find this effect for specific feedback which have a similar average duration between the two conditions (1.75 ms in normal condition and 1.81 ms in distracted condition).

5.3 Self-evaluation results

We analysed the questionnaires completed by the 60 participants at the end of the storytelling task. As described in section 3.5, storytellers evaluate their own production and the listener's behavior. We found no significant differences between the normal and distracted condition concerning their self-evaluation of their storytelling nor about their engagement in the interaction (the effect was searched for the overall task and for each story). Finally, concerning their evaluation of the listener, we also find no significant differences between the two conditions concerning their perception of listener engagement. These results suggest that none of the storytellers when faced with a distracted listener noticed anything suspicious. In both conditions, 100% of the storytellers found their partner to be appropriately responsive and attentive during the task. We also asked the storytellers if they thought their partner helped them during the storytelling, 80 % of storytellers in normal condition answered *yes* for only 66.7 % in the distracted condition. We know that the production of feedback is affected by the distraction task, but what the [5] experiment did not answer is the congruence of these feedbacks. Here, these results give us information about the reliability of the feedback produced, which appears to be congruent with the speech of the main speaker.

6 CONCLUSION AND DISCUSSION

In this paper, we present the SMYLE corpus, in which 60 participants engaged in spontaneous conversations while being audio-video and neurophysiologicaly recorded. Each interaction consists of a guided task and a free-conversation task. The storytelling has two conditions, where half of the listeners were given a hidden t-counting task whose purpose was to interfere in the global access to the meaning of the discourse. Several visual, vocal and multimodal annotations were performed.

A preliminary analysis was conducted. The results confirm those of the original study by [5], namely that the quality of the storytelling is negatively affected during the distraction task as well as feedback production for the 11 listeners analyzed (5 attentive and 6 distracted). Indeed, the production of specific feedback decrease by 40 %. In contrast to [5], we did not find differences in the frequency of generic feedback, but we did find that generic feedback are 20% shorter when the listener was distracted. [5] found that interaction quality is affected by the decrease in feedback produced while listening, especially specific feedback. Nonetheless, our preliminary analysis found a weak correlation between the storytelling quality and the frequency of feedback. We believe that the distraction task affects feedback production not only in terms of quantity but also in terms of function and forms. We found that generic feedback produced in the distracted condition is shorter than in normal condition. These results suggest that the generic feedback produced despite the distraction is less elaborate than in normal condition (i.e. shorter, unimodal, lower intensity, etc.).

In future work, we aim to extend this analysis by taking all the data into account and by examining how the feedback is produced, i.e., the verbal, vocal and gestural elements used to realized the feedback and how they are combined. Finally, storytellers in faced of a distracted listener do not judge them as less engaged in the interaction than in the normal condition. This result may be due to several factors. First, the type of story used in the first task are less emotional than the story used in the study of [5], where the storytellers was asked to recount a close call or a near-miss incident. Second, the distraction task requires the listener to focus on both the discourse and the spoken word, which implies that the listener has been concentrating and watching the storyteller a lot, giving the impression of being engaged. Third, the feedback generated during distraction may be crucial or "can't miss" feedback, that listeners managed to produce despite the distraction task.

In further work, we plan to use the SMYLE corpus to compute an interpretable multimodal model of feedback and to compare a distracted model and an attentive model. This approach will help us in understanding the feedback produced despite distraction and on which main speaker features they are based to better understand the context of feedback production. Finally, we want to exploit the brain signal to study the attentional and predictive mechanisms of feedback. This dataset is a valuable resource for studying the natural conversation and developing new models of human communication.

7 ACKNOWLEDGEMENTS

This work, carried out within the Institute of Convergence ILCB, was supported by grants from France 2030 (ANR-16-CONV-0002) and the CNRS MITI, the Cognition Institute and the ANR (Project COPAINS—ANR-18-CE33-0012). This dataset has been recorded in the soundproof room of the CEP experimental platform (LPL, AMU-CNRS, Aix-en-Provence, France). We would like to thank the team of annotators. AB warmly thanks Louis-Philippe Morency for his hospitality at Multicomp Lab.

REFERENCES

[1] Mary Amoyal, Roxane Bertrand, Brigitte Bigi, Auriane Boudin, Christine Meunier, Berthille Pallaud, Béatrice Priego-Valverde, S. Rauzy, and Marion Tellier. 2022. Principes et outils pour l'annotation des corpus. *Travaux Interdisciplinaires sur la Parole et le Langage* 38 (Dec. 2022). https://doi.org/10.4000/tipa.5424

[2] Mary Amoyal, Béatrice Priego-Valverde, and Stéphane Rauzy. 2020. PACO: A corpus to analyze the impact of common ground in spontaneous face-to-face interaction. In *Language Resources and Evaluation Conference*.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477 [cs.CL]

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[5] Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology* 79, 6 (2000), 941.

[6] Roxane Bertrand. 2021. *Linguistique Interactionnelle: du Corpus à l'Expérimentation*. Ph. D. Dissertation. Aix Marseille Université.

[7] Roxane Bertrand, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Revue TAL* 49, 3 (2008), pp.105–134. https://hal.science/hal-00349893

[8] Roxane Bertrand and Robert Espesser. 2017. Co-narration in French conversation storytelling: A quantitative insight. *Journal of Pragmatics* 111 (2017), 33–53.

[9] Roxane Bertrand and Beatrice Priego-Valverde. 2017. Listing practice in French conversation: From collaborative achievement to interactional convergence. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* 20 (2017).

[10] Dana Bevilacqua, Ido Davidesco, Lu Wan, Kim Chaloner, Jess Rowland, Mingzhou Ding, David Poeppel, and Suzanne Dikker. 2019. Brain-to-brain synchrony and learning outcomes vary by student–teacher dynamics: Evidence from a real-world classroom electroencephalography study. *Journal of cognitive neuroscience* 31, 3 (2019), 401–411.

[11] Brigitte Bigi. 2012. SPPAS: a tool for the phonetic segmentations of Speech. In *The eighth international conference on Language Resources and Evaluation*. 1748–1755.

[12] Philippe Blache, Salomé Antoine, Dorina De Jong, Lena-Marie Huttner, Emilia Kerr, Thierry Legou, Eliot Maës, and Clément François. 2022. The Badalona Corpus An Audio, Video and Neuro-Physiological Conversational Dataset. In *Language Resources and Evaluation Conference*.

[13] Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prevot, and Stéphane Rauzy. 2017. The Corpus of Interactional Data: a Large Multimodal Annotated Resource. In *Handbook of Linguistic Annotation*, N.Ide & J.Pustejovsky (Ed.). Springer, 323–1356. https://doi.org/10.1007/978-94-024-0881-2_51

[14] Auriane Boudin, Roxane Bertrand, Magalie Ochs, Philippe Blache, and Stéphane Rauzy. 2022. Are you Smiling When I am Speaking?. In *Proceedings of the Smiling and Laughter across Contexts and the Life-span Workshop @LREC2022*. Marseille, France. https://hal.science/hal-03713867

[15] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Thierry Legou, Magalie Ochs, and Philippe Blache. 2023. SMYLE. https://hdl.handle.net/11403/smyle ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[16] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A Multimodal Model for Predicting Conversational Feedbacks. In *International Conference on Text, Speech, and Dialogue*. Springer, 537–549.

[17] Pablo Brusco, Jazmín Vidal, Štefan Beňuš, and Agustín Gravano. 2020. A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. *Speech Communication* 125 (2020), 24–40.

[18] Wallace L. Chafe. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex.

[19] Herbert H Clark. 1996. *Using language*. Cambridge university press.

[20] Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, N. Tomashenko, Marco Dinarelli, Titouan Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier. 2021. LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. *ArXiv* abs/2104.11462 (2021).

[21] Greta Gandolfi, Martin J Pickering, and Simon Garrod. 2023. Mechanisms of alignment: shared control, social cognition and metacognition. *Philosophical Transactions of the Royal Society B* 378, 1870 (2023), 20210362.

[22] Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* 27 (1987), 181–218.

[23] Daniel Hirst. 2022. A multi-level, multilingual approach to the annotation of speech prosody. In *Prosodic Theory and Practice*. MIT Press, 117–149.

[24] William S Horton. 2017. Theories and approaches to the study of conversation and interactive discourse. In *The Routledge handbook of discourse processes*. Routledge, 22–68.

[25] Yi Hu, Yafeng Pan, Xinwei Shi, Qing Cai, Xianchun Li, and Xiaojun Cheng. 2018. Inter-brain synchrony and cooperation context in interactive decision making. *Biological psychology* 133 (2018), 54–62.

[26] Brent A Kelsen, Alexander Sumich, Nikola Kasabov, Sophie HY Liang, and Grace Y Wang. 2022. What has social neuroscience learned from hyperscanning studies of spoken communication? A systematic review. *Neuroscience & Biobehavioral Reviews* 132 (2022), 1249–1262.

[27] Sivan Kinreich, Amir Djalovski, Lior Kraus, Yoram Louzoun, and Ruth Feldman. 2017. Brain-to-brain synchrony during naturalistic social interactions. *Scientific reports* 7, 1 (2017), 17060.

[28] Iwan de Kok and Dirk Heylen. 2012. A survey on evaluation metrics for backchannel prediction models. In *Feedback behaviors in dialog*.

[29] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems* 20, 1 (2010), 70–84.

[30] Alejandro Pérez, Manuel Carreiras, and Jon Andoni Duñabeitia. 2017. Brain-to-brain entrainment: EEG interbrain synchronization while speaking and listening. *Scientific reports* 7, 1 (2017), 1–12.

[31] Alejandro Pérez, Guillaume Dumas, Melek Karadag, and Jon Andoni Duñabeitia. 2019. Differential brain-to-brain entrainment while speaking and listening in native and foreign languages. *Cortex* 111 (2019), 303–315.

[32] Martin Pickering and Simon Garrod. 2021. *Understanding Dialogue*. Cambridge University Press.

[33] Béatrice Priego-Valverde, Brigitte Bigi, and Mary Amoyal. 2020. "Cheese!": a Corpus of Face-to-face French Interactions. A Case Study for Analyzing Smiling and Conversational Humor. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 467–475.

[34] Stéphane Rauzy and Mary Amoyal. 2020. SMAD: a tool for automatically annotating the smile intensity along a video record. In *HRC2020, 10th Humour Research Conference*.

[35] Stéphane Rauzy, Grégoire Montcheuil, and Philippe Blache. 2014. MarsaTag, a tagger for French written texts and speech transcriptions. In *Second Asian Pacific Corpus linguistics Conference*. 220–220.

[36] Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk* 71 (1982), 71–93.

[37] Tanya Stivers. 2008. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on language and social interaction* 41, 1 (2008), 31–57.

[38] Victor H. Yngve. 1970. On Getting a Word in Edgewise. In *Papers from the Sixth Regional Meeting*, Mary A. Campbell (Ed.). Chicago Linguistics Society, Department of Linguistics, University of Chicago, Chicago, 567–578.