Studying Common Ground Instantiation Using Audio, Video and Brain Behaviours: The BrainKT Corpus

Eliot Maës
Leonor Becerra-Bonache
Aix-Marseille Université,
CNRS, LIS, Marseille, France
eliot.maes@lis-lab.fr
leonor.becerra@lis-lab.fr

Thierry Legou
Philippe Blache
Aix-Marseille Université,
CNRS, LPL, Marseille, France
thierry.legou@univ-amu.fr
blache@ilcb.fr

Abstract

An increasing amount of multimodal recordings has been paving the way for the development of a more automatic way to study language and conversational interactions. However this data largely comprises of audio and video recordings, leaving aside other modalities that might complement this external view of the conversation but might be more difficult to collect in naturalistic setups, such as participants brain activity. In this context, we present BrainKT, a natural conversational corpus with audio, video and neuro-physiological signals, collected with the aim of studying information exchanges and common ground instantiation in conversation in a new, more in-depth way. We recorded conversations from 28 dyads (56 participants) during 30 minutes experiments where subjects were first tasked to collaborate on a joint information game, then freely drifted to the topic of their choice. During each session, audio and video were captured, along with the participants' neural signal (EEG with Biosemi 64) and their electrophysiological activity (with Empatica-E4). The paper situates this new type of resources in the literature, presents the experimental setup and describes the different kinds of annotations considered for the corpus.

1 Introduction

Language processing in a natural context is inherently multimodal, and many studies have been devoted to better understanding how the interactions between the different channels leads to a better understanding between participants of a conversation. Interaction theories (Pickering and Garrod, 2021) postulate that this understanding is based on an operation of information transfer between participants, leading to the establishment of a common ground of knowledge. These processes happen at different levels, and the encoding and transmitting of information can be manifested through various cues for the different sources. These include feedback from gestures, gaze and facial expressions (Bavelas et al., 2000) and are also manifested at the physiological level with variations in respiratory rate, heart rate, skin temperature, etc. (Włodarczak and Heldner, 2016). Less perceivable to the other speaker but not less interesting for the understanding of their behavior, the brain activity denote of specific rhythmic activity when alignment between speakers occurs in a conversation, with the 10-12Hz (mu) frequency band presenting a specific pattern in the integration of mutual information during an interaction as well as in the coordination between speakers (Mandel et al., 2016; Pérez et al., 2017b; Menenti et al., 2012; Silbert et al., 2014). These new perspectives have laid the ground for investigations of natural conversations using neurophysiological elicited; however enterprises into this domain remain few in number, for various reasons. The main constraint indeed remains the technical difficulty to create a corpus of natural conversations condensing all of the aforementioned information sources, as movement inherent to speech might impede on the quality of the measured brain activity. Furthermore, though models of how the different sources of information interact during a conversation might exist for subsets of the modalities, there is to date no existing global view detailing how audio, video and neurophysiological features in a conversation interact to build and exchange meaning. Furthermore, information can be transmitted and integrated in a local time frame (at the given moment when it appears in conversation) but also with delay, impacting the conversation as a whole. Finally, the question of which experimental design to use to capture the progressive building of the common ground in the conversation needs to be resolved, as conversational tasks might be too constrained to correctly explain conversational

Proceedings of Recent Advances in Natural Language Processing, pages 691–702
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_075

behavior in the wild, and on the other hand free conversation being too reliant on external existing internalised world representations to accurately model and label the different types of information received.

We aim with this paper at addressing some of these questions and presenting a new, original resource for language processing studies. We describe below in greater detail our methodological approach to setting up an adequate experiment for acquiring synchronised multimodal natural conversation recounting of the progressive building of a shared knowledge base, the first steps of preprocessing techniques applied for data cleaning and the results we obtained from the early analyses. The originality of this project lies in two aspects. First, we combined existing conversational tasks to induce a discussion where information transfers could be observed and common ground build on gradually, starting from a very controlled environment and increasingly releasing the constraints on conversational vocabulary and topics. Seconds, we recorded various types of data, namely audio, video, physiological and brain activity, all of which are crucial when studying natural conversation. Compared with recent research which adopted light EEG headset that traded off commodity for recording quality (Park et al., 2020), we aimed at developing a protocol for recording every modality with great quality.

Sections 2 and 3 detail our goals for setting up such an experiment and the context in which it is set. Next, we outline in sections 4 and 5 our experimental protocol. Section 6 describes the processing steps realised on the data to ensure quality and synchronisation between the different modalities recorded, and Section 7 presents the first few analyses we ran on the corpus.

2 Scientific Goals

Unlike language models that learn from massive amounts of data from data sources of various qualities, simulating good language capabilities but failing at delivering a precise description of human language processing, models aiming to better understand language capabilities usually focus on smaller and well-curated datasets. Acquiring data for studying conversation in a natural context remains complex because of the heterogeneous nature of the different sources of information that can be collected and analysed. If audio/video recordings are quite widespread, this is not the case of neuro-physiological recordings which, when they exist, are in limited quantity. A dataset allowing for the extensive study of conversational markers concurrently using audio, video and neurophysiological modalities does not currently exist. With this work, we aim for two goals: first, developing a protocol for acquiring adequate resources for the neuro-physiological study of conversational behaviours in a natural setting; and secondly, designing new resources for the study of information transfer and common ground instantiation in free conversation.

With these research questions, an important feature for designing experimental protocols is balancing the conversation environment. Constrained experimental tasks such as the MapTask (Anderson et al., 1991) are indeed great at generating conversational attempts, measuring task successes and failures and linguistic alignment; information transfers are clearly identifiable and conversation evolution can be parameterized. They are however restrictive and not representative of most conversational behaviors, which can cover a wide range of topics and usually rely on knowledge far from the experimental context, conversational schemes and experience specific to a speaker. For these reasons, information transfers are more difficult to study in natural conversation, as they can take a larger range of shapes and intensities. With this in mind, we recorded participants through a several tasks experiment, designed so as to progressively release the constraints on conversational topics and gradually allow for the introduction of new vocabulary, concepts and knowledge to the conversation. Each 30 minutes experiment starts with a 15min collaborative video game where one player possesses all information relative to solving the game and must instruct the other player who operates the game. Once this controlled task completed the experiment then moves on to the discussion of personal views, with a moral dilemma that participants have to discuss and agree on, before finally moving on to the topics of their choice and a freer conversation. Participants familiarity and mutual knowledge progressively increase throughout the course of these experiments (dyads were not acquainted before the experiment), offering a way into the study of their progressive alignment. The combination of these very different tasks also allows for the comparison of communication strategies and efficiencies be-

tween very specific contexts and completely free conversation.

We collected 28 such interactions (∼14 hours) between French speakers, complete with the recordings of their verbal, behavioral, physiological and neurological activities and later enriched with various annotations and descriptors for the different modalities (transcription and morpho-syntactic labeling, facial landmarks and movement annotations, moments corresponding to information exchanges...). When collecting such corpora, a specific attention must be paid to the technical difficulties that arise, namely the synchronisation between all modalities and how behaviors in one modality might affect the collected quality of another. It is for instance necessary to find a good tradeoff between EEG signal quality, which can be very affected by sources of noise such as gestures and speech, and the degree of freedom given to participants for the experiment be considered naturalistic. The corpus will then be used to study conversational patterns across all collected modalities, as the progressive alignment of speakers in conversation can be observed in their verbal (reuse of lexical terms, prosodic similarity), behavioral, physiological (respiratory, heart rate etc) but also neurological activity (Pérez et al., 2017a). Physiological and neurological correlates for information transfers, speakers alignment, parameters for the success of an interaction will be investigated, both at local and larger scales.

Despite the focus of the experimental design on generating information transfer between participants, the inclusion of a free conversation task will allow for the wider reusability of the rare corpora for other research questions which might benefit from any kind of multimodal setups. Finally, increasing our understanding of human linguistic behaviors might find applications for the improvement of Human-Machine interfaces.

3 Related works

3.1 Multimodal datasets

Several datasets have been acquired targeting a set of modalities similar to ours (audio, video, physiological and neural signals). Most of them have been designed in perspective of the study and prediction of emotions, more specifically arousal and valence. Among the most renowned, we can mention DEAP (Koelstra et al., 2011), MAHNOB-HCI (Soleymani et al., 2011), DREAMER (Katsigiannis and Ramzan, 2018) and AMIGOS (Miranda-Correa et al., 2021).

Recently, the push for naturalistic experimentation seems to have stimulated the interest in this topic. Despite known hurdles, several datasets pertaining to multimodal conversation and including neurological data have been collected, such as K-Emocon (Park et al., 2020) or the Badalona corpus (Blache et al., 2022). These acquisitions however remained limited, both in the duration of interactions recorded as well as in the quality of neurological data acquired, as only light headsets were used.

3.2 Video games as an experimental paradigm

In addition to free conversation, we include in our paradigm a more controlled conversational task, a game setup fostering information exchanges. Rather than using the MapTask (Anderson et al., 1991) - which is a common design for eliciting information exchanges and conversation - we turned to video games for a more immersive design.

The use of video games in experimental paradigms has soared over the past few decade (Washburn, 2003; Lim and Holt, 2011) as games provide both incentive for the recruitment of participants, and by their design ensure the continuous engagement of participants in the task. Games have also been found to be appropriate tools to elicit and study human interactions and spontaneous natural conversations (Duran and Lewandowski, 2020; Ward and Abu, 2016). Despite the large number of existing games than can be tuned to the research questions, it is however often necessary to adapt the setup, either to allow for the exact control of stimuli, or to monitor participants actions during a task.

We propose a setup using the game Keep Talking and Nobody Explodes, a collaborative game between two (or more) players which has been used previously to study communication in virtual settings (Baker, 2018). Similarly to the MapTask, this games requires the two participants to share the information they have in order to succeed with the task.

4 Data Collection Setup

4.1 Materials and Methods

When humans interact, various modalities are used to transmit a message across. Visual clues such as facial expressions and gestures complement the

linguistic content uttered; prosody might enhance understanding or give away a speaker's state of mind. Conversational phenomenons such as convergence and alignment between participants can be observed in those channels, but also in neurophysiological data, which are affected by mental states and emotions. Considering the various modalities are correlated and complementary, we record the interaction between participants at various levels, using audio, video, and neurophysiological devices.

Both participants were equipped with head microphones (AKG C520) and filmed from the front by a camera (Canon XF105) located behind the other participant and hidden by a green sheet. The microphones recorded the audio at 48kHz/16 bits and were connected to a RME Audio Inferface for sound quality and gain control. The sound was then sent both to a computer for recording (Audacity) and to the cameras for synchronisation with the video.

Participants brain activity was recorded using the BioSemi ActiveTwo system with headcaps with 64 electrodes.

Finally, Empatica E4 wristbands were used to log participants physiological parameters during the interaction. Those include blood volume pulse (BVP), electro-dermal response (EDA), inter bit interval (IBI), heart rate (HR), skin temperature (TEMP), and also behavioural information using a 3 axis accelerometer (ACC). Despite being monitored by the same device, physiological parameters are recorded with different frequencies (see Table 1 for details).

Auditory, visual and numerical (EEG) triggers were included across all modalities so as be able to reconstruct the multimodal signal (see Section 6.1).

All data collection sessions were conducted in a sound-proof room with controlled temperature and illumination. The two participants sat across a table facing each other with a distance in between for a comfortable communication (see Figure 1).

Devices	**Collected data**	**Sampling rate**
Empatica E4 Wristband	3-axis acceleration	32Hz
	BVP	64Hz
	IBI	n/a
	Heart Rate	1Hz
	EDA	4Hz
	Body Temperature	4Hz
BioSemi 64	EEG	2048Hz
Canon XF105	video	25fps
AKG C52	audio	48kHz

Table 1: Mobile devices used and data recorded.

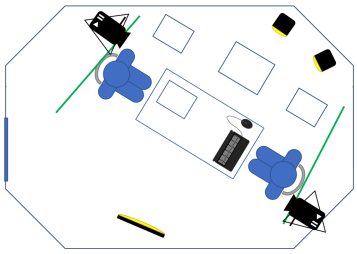

Figure 1: Diagram of the setup: both participants are installed facing each other, separated by a table (about 1.4m wide) and material used during the tasks. Cameras were positioned opposite to the participant they were filming, above the other participant's left shoulder.

Figure 2: Video montage of the feed captured by the two cameras during the experiment, with the participants in gear.

4.2 Post-Experiment Questionnaire

Participants were asked to answer several questionnaires after the completion of their tasks, both a record of their subjective analysis of the experiment and a log of their personality.

In line with existing research (Baker, 2018), we included a shortened version (9 questions) of the trust measure developed by (Couch et al., 1996). The communication questionnaires included a 5-item team effectiveness measure (Gibson et al., 2003) to gauge their assessment of their performances during the first task (game), as well as an evaluation of the fluency of their transmissions on the Communication Quality Scale (González-Romá and Hernández, 2014) (both tasks). Finally, as involvement in a conversation is a key feature of communication success, we included questions targeting their perception of both participants engagement throughout the experiment.

5 The Experiment

5.1 Participants

56 participants (age: 22.6 $\pm$ 3.6 yo; 44 females for 12 males) were recruited between November and December 2022 based on postings on lab's the social network accounts and in nearby universities. Participants were French natives who were required to have normal to corrected vision with no color-blindness, no history of neurological disorder nor photosensitive epilepsy. We checked that the two participants of a dyad were not acquainted with one another, so that the experiment would not be biaised by pre-existing shared communication schemes.

5.2 Data Collection Procedure

Data collection sessions were conducted in five stages: 1) Onboarding 2) Installation 3) Material check and instructions 4) 2-task Experiment 5) Post-experiment questionnaire. Two to three experimenters administered each session.

Onboarding Upon arrival, participants were each provided with two consent forms to sign. Upon agreeing with participating with the research, they were given an additional document containing the instructions for both tasks (see Section 5.3). They were then asked to decide among themselves which role they would have in the experiment.

Installation Participants were prepped in separate rooms, so that any chitchat during the installation of the recording equipment would not affect the tasks, which required the participants not to have any knowledge of one another. Measures were taken of the participants heads so as to chose the best fit for the EEG caps. Participants were then setup with the equipment in the following order: first, three EEG electrodes used for references were placed, one under the left eye and two under each mastoid. Secondly, the head microphone was placed. Finally, the EEG cap was placed. Electrolyte gel was applied on the subjects heads before connecting the electrodes, bridging the gap between the scalp and the measurement probes. Electrodes were positioned following the International 10-20 system. Participants were then moved to the experiment room and the Empatica E4 wristband was placed on their arms.

Material check and Experiment instructions Participants were placed in the experiment room following the diagram in Figure 1. Participant 1 (P1) was given the computer and two tutorials to complete, so as to learn how to interact with the game for the experiment. Participant 2 (P2) was given the game manual for the bomb defusal and a few minutes to browse through it; they were instructed not to try too hard to understand it (which can be difficult with no knowledge of the game) but rather prioritise understanding of the manual structure and how to lookup information during the task. Concurrently, EEG signal quality and electrodes impedances were checked; gain for both microphones was adjusted. Once both participants were ready, final instructions were given and recording equipment was started: cameras first, then E4 wristbands, audio and eeg recording. Experimenters left the box.

Experiment Three audiovisual triggers informed the participants of moments to start the experiment, switch tasks, and finish. As conversational progress was favored over exact task duration, they were told to ignore the stopwatch appearing on the computer. Both tasks were to last for about 15 minutes, with one experimenter keeping track of the conversation so as to trigger the task end in adequate moments.

Post-experiment questionnaires Upon tasks completion, participants were quickly unequipped and given the link to the post-experiment questionnaire, hosted on FindingFive[1] (see Appendix C). They were to fill the questionnaire without exchanging with the other participant on their impressions, but an experimenter remained with them to answer possible questions. Completing the questionnaire would unlock payment through the platform.

5.3 Tasks

The experimental session consisted of two tasks: a controlled conversational task and a free conversation task, amounting in total to about 30 minutes. Each task is described in more detail below.

5.3.1 Keep Talking and Nobody Explodes

Keep Talking and Nobody Explodes[2] is a collaborative game for two or more players, freely available to the public on the game platform Steam. The developers encourage the use of the game for non-commercial educative or company events as long as a licence has been purchased for every computer it runs on.

[1] https://findingfive.com
[2] https://keeptalkinggame.com

Figure 3: Screenshots (front, side, back) of the bomb the participants team had to defuse, as it appeared for P1. There are 7 modules to defuse on the bomb. A timer and an error counter are included but not for the defusal in our case.

Upon arrival, participants were introduced to the general concept of the game and the two possible roles they could have. They had to collaborate to defuse the bomb in a video game. They could either play as the *bomb defuser* (P1), interacting directly with the game interface, or the *expert* (P2), holding the bomb manual and being the knowledge reference for the bomb defusal.

Manual The bomb manual participants used was almost identical to the game version. The biggest edit consisted in the removal of pages that were irrelevant to the experiment and a few addendum meant to help new players grasp the concepts of the game quickly and locate information. One of the module pages was also edited to match the setup customisation.

Game configuration In order to ensure customisation to our needs as well as identical reproduction of the bomb design accross all experiments (which is not present by default in the game), several mods are used in the experiment. Mods are player-coded adds-on to the game allowing for customisation, from adding levels and modules to the bombs, to creating controlled experiments. In our setup, the *Dynamic Mission Generator*[3] (DMG) was used to configure the bomb. The DMG relies on the *Mod-Selector*[4] to be installed to run. Considering our interest was more on discussion mecanics rather than performance, we chose a configuration of the bomb (see Figure 3) such that most new player teams would either not manage to defuse the bomb in time, or manage but with very little time left.

[3] https://github.com/red031000/ktane-DynamicMissionGenerator

[4] https://steamcommunity.com/sharedfiles/filedetails/?id=801400247

5.3.2 Free Conversation Task

The participants were given a moral dilemma to discuss during the Free Conversation task. The participants' goal was to discuss the possible outcomes of the dilemma and to eventually agree on a solution. When they had agreed on a solution, they were enjoined to learn about each other. The discussion was to last for around 15 minutes; the document listing the instructions was left in the experiment room and could be consulted by the participants at any time.

The moral dilemma used is known as the "hot-air balloon" dilemma and is commonly used in research to elicit natural conversations (Koskinen et al., 2021):

> A hot-air balloon is losing altitude and is about to crash. The only way for any of the three passengers of the balloon to survive is that one of them jumps to a certain death. The three passengers are: a cancer scientist, a pregnant primary school teacher, and the husband of the teacher, who is also the pilot of the balloon. Who should be sacrificed?

Conversation excerpts and details about the game configuration are available in Appendix A.

6 Data Pre-processing

6.1 Synchronization

Synchronisation is primordial for the optimal use of the corpus. However, since the modalities were recorded through separate means, several strategies were used to ensure that the data could correctly be synchronised properly:

- Audio-Video: a clapperboard was used to create an audio-visual trigger at the beginning of

the experiment. Furthermore, separate recordings were made of the audio signal (cameras and computer using the RME software)

- Audio-EEG: tasks in the EEG recording were delineated by triggers, which were accompanied by an audio-visual signal.
- Video-Empatica wristband: at the start of each experiment, the pressing a button on the watch flashed a led, which is captured by the camera and recorded as a timestamp in the device memory.

Alignment check between the different modalities was realised mostly automatically using Python, with human verification and correction for a few files.

For all experiments, Camera 1 audio-video signal was used as a reference. Camera 2 is aligned at the video frame level during montage, so that the experiment start clap happens simultaneously in both videos. Refining is then done for the audio using Python: each channel of the RME signal is separately aligned to the corresponding channel in the camera signal, then the difference between the two RME channels is used to realign both audios in the camera signal. The RME signal was not kept (despite a better audio quality) as in some files the audio seemed to skip short (0.2s in average) parts of the conversation, desynchronising from the video.

The video signal is then synchronised to the video signal from the Empatica wristbands.

There was no issue concerning the synchronisation of the EEG brain signal from the 2 participants as both participants were recorded simultaneously by BioSemi ActiveTwo. The synchronisation of EEG to the other modalities relied on the detection of the simultaneous audio-EEG trigger in the audio signal. The frequency used for the trigger was very distinctive (2793.82Hz, F7 on a keyboard), which could be localized accurately during silence moment that preceded the start of the experiments.

EEG and Empatica files were trimmed / padded to match the start and duration of audio and video files.

6.2 Data Quality

As this kind of audio-video setup has been realised before (Blache et al., 2022; Amoyal et al., 2020), our main concern was the brain signal quality. We used MNE-Python (Gramfort et al., 2013) to preprocess the EEG data, splitting speakers signals into separate files, applying first preprocessing steps. Filters were applied to remove activity outside of the 1Hz-70Hz band, bad channels and channels with correlated activity were located and interpolated channels correlated activity. Finally, the extended infomax ICA algorithm (Lee et al., 1999) was run to identify bad components in the signal. Automatic labelling of ICA components was used to facilitate component annotation and run using `ICLabel` (Li et al., 2022).

Two files were automatically rejected during preprocessing because of noisy signal and a high number of bridged electrodes.

6.3 Annotations

A two steps procedure is used to generate automatic transcriptions of the corpus: first, units of continuous speech (IPU) without pauses longer than 200ms (IPU) are identified in the speech signal; each IPU is transcribed using Wave2Vec2.0[5] (Baevski et al., 2020). The transcripts are then manually checked and corrected. Finally, word and phonemes alignment to the audio signal, and Part of Speech tagging are realised using SPPAS (Bigi, 2012). Additional high level annotations such as the different themes of the conversation are added using ChatGPT[6] (Ouyang et al., 2022). Regarding the video modality, video analysis pipelines such as OpenFace's (Baltrusaitis et al., 2018) FeatureExtraction are used to compute head movements and gaze. The generated coordinates for facial landmarks and actions units are then fed into the HMAD (Rauzy and Goujon, 2018) R library for extraction of nods and smile annotations.

Additional annotations will be added in the future to support the investigations into information transfers in conversation and other research questions that may arise.

6.4 Dataset Organisation

The BrainKT dataset is available upon request on Ortolang[7].

Each file is tagged by collection date (`<date>`), dyad initials (`<dyad>`), participant identifier (`p<X>` or participant initials `<ipart>`) and

[5] A fine-tuned model for french was used `bofenghuang/asr-wav2vec2-ctc-french` https://huggingface.co/bofenghuang/asr-wav2vec2-ctc-french

[6] https://openai.com/blog/chatgpt

[7] https://hdl.handle.net/11403/brainkt

General	Number of dyads	28
	Participants average age	22.6 ± 3.6
	Participants gender	44F - 12M
	Total corpus duration (hours)	14
	Number of words (KTaNE game)	≈60k
	Number of words (free conversation)	≈75k
Task1	Average number of cleared modules	5.3 ± 1.5
	Median number of cleared modules	6
	Average number of errors	13.8 ± 16.3
	Median number of errors	8
	Max number of errors	70
	Shortest defusal	13min
	Number of groups defusing the bomb	6
Task2	Average duration of the dilemma topic in conversation	6min ± 4min
	Shortest time spent on the dilemma	35s
	Character sacrificed most times	pilot
	Average number of themes in conversation (automatic annotation)	12.7 ± 3.7

Table 2: General analysis of the corpus

task identifier (`t1` or `t2`) depending on the requirements of the modality. Therefore each file is named based on the pattern: `bkt-<date>-<dyad>(-p<X>)(-t<i>)`

metadata this folder contains csv files for EEG data quality, experiment results, temporal markers of events in the experiment, and anonymised participants answers (`.csv`) to the post-experiment questionnaire.

video for each experiment, the video `.mp4` montage of the two camera recordings of the participants, and the view on the computer screen during the first task

audio for each experiment, a `.wav` file with two channels (first channel being P1, and second channel P2)

e4 for each participant, a JSON file containing the physiological signals recorded by the wristband (heart beat, movement...)

eeg-raw for each participant, a `.fif` file (MNE-Python format) of the aligned signal

eeg-task for each task, a `.edf` file containing the preprocessed EEG data, from task start trigger to task end trigger

transcript for each experiment, a `.eaf` file with the transcripted utterances for each participant (`-<ipart>`)

Audio, physiological and neurological (`eeg-raw`) data are aligned to the video signal (start / end), as can be seen in Appendix B.

7 Dataset Analysis

A first analysis of the corpus can be done based on experimental videos, transcripts and questionnaire answers (see Table 2). Overall, most players had a very sparse gaming activity and had either never heard of the game, or heard of it and never played (knowledge on average: 0.32 / 3). They rated their engagement during the experiment as rather high (4.5±0.6/5 overall). During the first task, most groups did not manage to defuse the whole bomb (only 6 did so) but still came close to finishing (5 modules solved on average). The module that was solved the most times is the Wires module placed on the front of the bomb. The module solved the least amount of times was the Simon, also placed on the front face. A detailed account of game statistics is given in Appendix D. The free conversation (Task 2) has about 25% more words than the game (Task 1), as participants would have had needed to take the time to try and understand how the game worked and mostly did that by muttering to themselves or reading the instructions in their minds. In Task 2 however, the conversation flowed more naturally.

8 Conclusion and Perspectives

In this paper, we presented a procedure for collecting new types of naturalistic corpora including a larger number of sources of information (audio, video but also physiological and neural signals) and the dataset collected as a result. The perspectives of use of this data are numerous: as a language resource, this dataset can be used in the study of convergence and alignment between participants in a conversation, through its tasks gradually releasing

the constraints on conversation. The neurological part of the data can be used to further the research into natural conversation procedures and how to deal with noise and movement when running such experiments. But most interestingly, this new kind of corpora opens the way to the possibility of multimodal models complementing audio-video analysis with neurophysiological cues. Future works will focus on enhancing the dataset with additional annotations and a more in-depth analysis of the corpus. The dataset is being made available through the Ortolang repository.

Acknowledgments

This work was carried out within the Institut Convergence ILCB (ANR-16-CONV-0002) and as such has benefited from support from the French government, managed by the French National Agency for Research (ANR). The experiments were conducted with the help of the ILCB Center of Experimental Resources (CREx) and with the backing of Auriane Boudin (LPL) and her team. The work of Leonor Becerra-Bonache (LIS) has been performed during her teaching leave granted by the CNRS (French National Center for Scientific Research) at the Laboratoire Parole et Langage of Aix-Marseille University.

References

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 *LREC procs*. pages 628–633.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech* 34(4):351–366.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33:12449–12460.

Anthony Lee Baker. 2018. *Communication and trust in virtual and face-to-face teams*. Embry-Riddle Aeronautical University.

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)*. pages 59–66.

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

Brigitte Bigi. 2012. Sppas: a tool for the phonetic segmentations of speech. In *The eighth international conference on Language Resources and Evaluation*. pages 1748–1755.

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 neurophysiological conversational dataset. In *Language Resources and Evaluation Conference*.

Lauri L Couch, Jeffrey M Adams, and Warren H Jones. 1996. The assessment of trust orientation. *Journal of personality assessment* 67(2):305–323.

Daniel Duran and Natalie Lewandowski. 2020. Demonstration of a serious game for spoken language experiments-gdx. In *Workshop on Games and Natural Language Processing*. pages 68–78.

Cristina B Gibson, Mary E Zellmer-Bruhn, and Donald P Schwab. 2003. Team effectiveness in multinational organizations: Evaluation across contexts. *Group & Organization Management* 28(4):444–474.

Vicente González-Romá and Ana Hernández. 2014. Climate uniformity: Its influence on team communication quality, task conflict, and team performance. *Journal of Applied Psychology* 99(6):1042.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. 2013. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience* page 267.

Stamos Katsigiannis and Naeem Ramzan. 2018. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical And Health Informatics* 22(1).

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3(1):18–31.

E. Koskinen, S. Tuhkanen, M. Järvensivu, E. Savander, T. Valkeapää, K. Valkia, E. Weiste, and M. Stevanovic. 2021. The psychophysiological experience of solving moral dilemmas together: An interdisciplinary comparison between participants with and without depression. *Frontiers in Communication* 6.

Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. 1999. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation* 11(2):417–441.

Adam Li, Jacob Feitelberg, Anand Prakash Saini, Richard Höchenberger, and Mathieu Scheltienne. 2022. Mne-icalabel: Automatically annotating ica components with iclabel in python. *Journal of Open Source Software* 7(76):4484. https://doi.org/10.21105/joss.04484.

Sung-joo Lim and Lori L Holt. 2011. Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive science* 35(7):1390–1405.

Anne Mandel, Mathieu Bourguignon, Lauri Parkkonen, and Riitta Hari. 2016. Sensorimotor activation related to speaker vs. listener role during natural conversation. *Neuroscience letters* 614:99–104.

Laura Menenti, Martin J Pickering, and Simon C Garrod. 2012. Toward a neural basis of interactive alignment in conversation. *Frontiers in human neuroscience* 6:185.

Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2021. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* 12(2):479–493.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35:27730–27744.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7(1):293. https://doi.org/10.1038/s41597-020-00630-y.

A. Pérez, M. Carreiras, and J. Duñabeitia. 2017a. Brain-to-brain entrainment: Eeg interbrain synchronization while speaking and listening. *Scientific Reports* 7(4190).

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

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

Stéphane Rauzy and Aurélie Goujon. 2018. Automatic annotation of facial actions from a video record: The case of eyebrows raising and frowning. In *Workshop on" Affects, Compagnons Artificiels et Interactions", WACAI 2018*. page 7.

Lauren J Silbert, Christopher J Honey, Erez Simony, David Poeppel, and Uri Hasson. 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences* 111(43):E4687–E4696.

Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3(1):42–55.

Nigel G Ward and Saiful Abu. 2016. Action-coordinating prosody. In *Speech Prosody*. pages 629–633.

David A Washburn. 2003. The games psychologists play (and the data they provide). *Behavior Research Methods, Instruments, & Computers* 35(2):185–193.

Marcin Włodarczak and Mattias Heldner. 2016. Respiratory turn-taking cues. In *Interspeech 2016, San Francisco, USA, September 8–12, 2016*. The International Speech Communication Association (ISCA), pages 1275–1279.

A Tasks details

A.1 Conversation excerpt

Excerpts for each task can found in Table 3.

A.2 Game resources

The first task relied on existing resources: the *Keep Talking and Nobody Explodes* game with its computer version and online manual, and adds-on developed by the gaming community. Minimal adjustments were made to the player manual to adapt it to our configuration (no time nor error limit, reduced variety in modules) and so that new players could grasp the context faster. Figure 4 shows two pages taken from the adapted manual.

For the bomb, variations on module combinations and game seeds were tested until we obtained a satisfying configuration. The original game features 12 types of modules: *Wires, Button, Keypads, Simon Says, Who's on First, Memory, Morse Code, Complicated Wires, Wire Sequences, Mazes, Passwords*, and *needy modules*. We only kept the modules we deemed easiest to understand, though some were still more difficult than others. Two modules were duplicated with slightly different versions so as to make possible the study of the evolution of communication strategies once extra information and knowledge was added to the common ground. The final configuration of the game included: 2 Wires modules, 2 Keypad modules, 1 Maze, 1 Simon Says and 1 Password module.

B Synchronisation

Despite the experiment not being as controlled as is usually the case for protocols involving EEG, with (for instance) triggers sent to the signal for each stimulus presentation, the various triggers left in the different modalities still allow for the synchronisation and precise analysis of each signal. Figure 5 shows how such a synchronisation can be observed: annotations of dialogue spoken and heard can be added to the brain signal, and interest locations can be targeted for analysis.

C Questionnaires

Post experiment, in order to unlock payment, participants had to fill several questionnaires quizzing their experience during both tasks. The questionnaire were hosted on FindingFive[8] (see Figure 6 for a screenshot of the interface). Besides participants demographics and game knowledge, we included several questions probing participants attitude toward new people (dyads weren't acquainted pre-experiment), their verbal behavior and engagement during the tasks. Indeed, personality features and involvement in the conversation might be of interest when investigating interaction success. A complete list of questions asked can be found in Table 6.

D Statistics

A brief analysis of team performances and choices in each tasks can be found in Tables 4 (game) and 5 (dilemma).

During the game, most participants started defusing the modules on the front face of the bomb, with Wires being the top-left most module often being the first one attempted. However exploring the bomb and acquiring new information lead to other modules being finished first. Keypad and Wires modules were completed the fastest, with the second instance of the module being completed in half the time. The most difficult module to complete was the Simon, as the number of errors could suddenly affect the behavior of the module.

Two options were favored in the dilemma, either sacrificing the pilot or the researcher. 8 out of 28 groups either did not manage to agree on a solution or agreed on other strategies despite the instruction. Several groups went on to discuss other dilemmas as part of the free conversation.

[8] https://findingfive.com

speaker	text
EM	ok. après j'ai
	quatre boutons
	rouge bleu jaune et vert
	dans un module
TR	ok. c'est peut être le simon. ouais c'est ça
	il y en a un des quatre qui s'allume
EM	hm...non
	ah si le rouge
TR	le rouge
EM	oui il clignote de temps en temps
TR	ok
EM	je pense que je tuerais le scientifique
	parce que déjà le mec qui conduit la montgolfière à quel moment il va accepter de
	balancer sa femme par dessus bord
TR	oui c'est vrai en fait
EM	et oui
	je pense parce que de toute façon euh
	c'est malheureux mais ça fait deux contre un donc euh
TR	à moins qu'il y ait des problèmes de couple
	tu sais pas
EM	c'est pas faux
TR	mais le scientifique c'est vrai que la première chose que je m'étais dit bah s'il a des recherches contre un cancer
	il serait
	possiblement
	important entre guillemets
	en même temps si ses recherches
	si on sait que ses recherches
	pourraient guérir un cancer. c'est-à-dire si elles sont assez avancées et qu'un autre chercheur
	pourra les reprendre
EM	bah j'ai eu la même
	au début là quand j'ai lu le truc c'était en vrai le scientifique il peut être utile à l'humanité donc
	il faudrait le sauver
	et en même temps est-ce que qu'on met une hiérarchie sur les vies en fonction de
TR	de la profession

Table 3: Conversation excerpts for the game (top) and dilemma (bottom) conversations. Speakers are referenced to by their initials. Different lines correspond to utterances separated by pauses longer than 200ms.

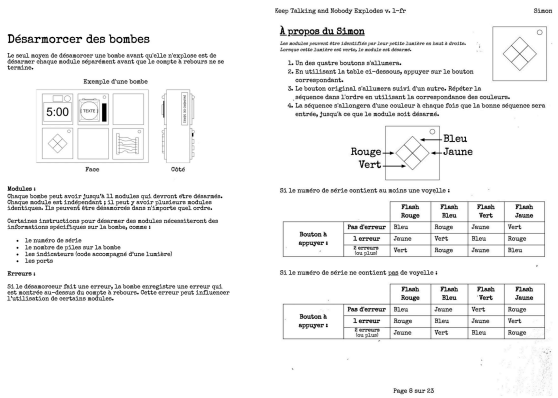

Figure 4: Extract of the game bomb manual: left, the instruction page explaining how to disarm the bomb; right, the instructions for one of the modules

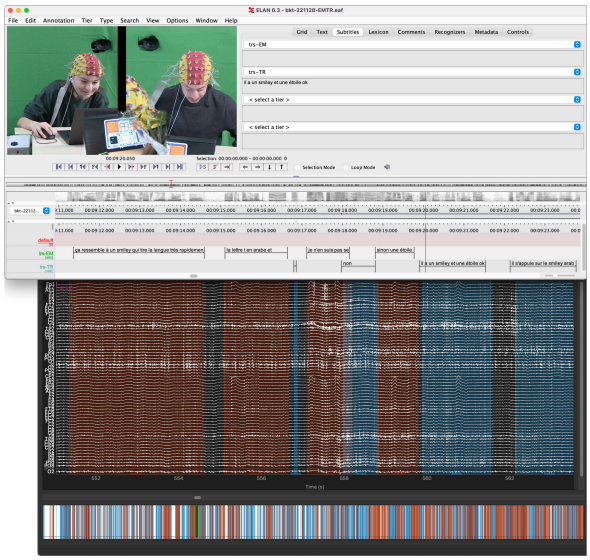

Figure 5: Parallel view of the same moment in the experiment, with video / transcription in ELAN and EEG in the MNE browser. Red (respectively blue) annotations on the EEG signal correspond to spoken (respectively heard) by the participant. The synchronization procedure allows for the parallel annotation and analysis of all modalities.

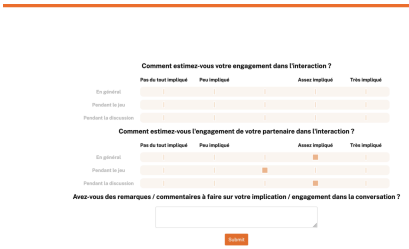

Figure 6: Screenshot from the FindingFive website, where the questionnaire was hosted

	Completion Rate	Average Duration	First Attempted	First Validated
Keypad (Top)	20	127.6s	0	0
Keypad (Bottom)	20	63.1s	0	0
Wires (Front)	26	127.8s	20	16
Wires (Back)	23	59.4s	1	5
Maze	21	204.92s	0	0
Password	24	210.5s	1	6
Simon	15	246.6s	6	1

Table 4: Detailed analysis of the KTaNE task results

	Times sacrificed	Most recurrent reason
Teacher	5	cannot pilot nor potentially save lives
Researcher	7	cannot split the couple, team research
Pilot	8	failure at piloting, life with least value
Other option	5	lightening the balloon, killing ever
No consensus	3	Ran out of time, refused to agree

Table 5: Dilemma agreement results

Questionnaire	**Target**	**Questions**	**Answer range**
General		Gaming Activity	
		Previous knowledge of the KTane game	(None, Heard of, Played a few times, Expert level)
Generalised Trust Scale		I tend to be accepting of others	Completely disagree (1) → Completely agree (7)
		My relationships with others are characterized by trust and acceptance	
		I make friends easily	
		I find it better to accept others for what they say and what they appear to be	
		Experience has taught me to be doubtful of others until know they can be trusted	
		I tend to think that things will work out in the end	
		I tend to take others at their word	
		I feel I can depend on most people I know	
		It is better to be suspicious of people you have just met. until you know them better	
Team Effectiveness Scale	KTaNE	This team has a low error rate	Completely disagree (1) → Completely agree (5)
		This team does high quality work	
		This team consistenty provides high-quality output	
		This team is consistenty effor-free	
		This team needs to improve its quality of work	
Communication Quality Scale	Separate questions for KTaNE + Discussion	*Was the communication between you and the other participant:* clear ? effective ? complete ? fluent ? on time ?	Completely disagree (1) → Completely agree (5)
Engagement in the experiment	Separate questions for Self + Partner Assessment	*How involved were you...* In general, throughout the experiment / During the game / During the discussion	Not at all (1) → Very involved (5)

Table 6: List of questions in the questionnaire, by order of apparition