Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5170–5177
Marseille, 20-25 June 2022

The Badalona Corpus
An Audio, Video and Neuro-Physiological Conversational Dataset

Philippe Blache[1], Salomé Antoine[2], Dorina De Jong[3], Lena-Marie Huttner[1], Emilia Kerr[1], Thierry Legou[1], Eliot Maës[4], Clément François[1]
[1]Laboratoire Parole & Langage (CNRS-Aix-Marseille Univ.), [2]Freie Universität Berlin,
[3]Center for Translational Neurophysiology of Speech and Communication (Università di Ferrara),
[4]Laboratoire Informatique & Systèmes (CNRS-Aix-Marseille Univ.)
blache@ilcb.fr, salome.antoine@fu-berlin.de, dorina.dejong@iit.it,
lena-marie.huttner@univ-amu.fr, emilia.kerr@univ-amu.fr,
eliot.maes@etu.univ-amu.fr, thierry.legou@univ-amu.fr, clement.francois@univ-amu.fr

Abstract

We present in this paper the first natural conversational corpus recorded with all modalities and neuro-physiological signals. Five dyads (10 participants, Spanish native speakers) have been recorded three times, during three sessions (about 30 minutes each) with 4 days interval. During each session, audio and video are captured as well as the neural signal (EEG with Emotiv-EPOC) and the electro-physiological one (with Empatica-E4). This resource is original in several respects. Technically, it is the first one gathering all these types of data in a natural conversation situation. Moreover, the recording of the same dyads at different periods opens the door to new longitudinal investigations such as the evolution of interlocutors' alignment over time. The paper situates this new type of resources in the literature, presents the experimental setup and describes different annotations enriching the corpus.

Keywords: Multimodal corpus, natural conversation, convergence, neuro-physiological data, EEG

1. Introduction: studying brain basis of language processing in a natural context

Studying language processing in a natural context means that various sources of information need to be taken into account. Many studies have been devoted to the interaction between the different modalities, verbal and non-verbal (Jewitt, 2013). They have explored in detail how the different sources of information interact in order to encode, transmit and decode information during a natural interaction between interlocutors (Pickering and Garrod, 2021). This new perspective has laid the ground for investigation of the underpinning mechanisms of natural interaction by analyzing the neuro-physiological signals elicited by this phenomenon. The aim of the present project is to study, on the one hand, the physiological signals (electro-dermal reaction, temperature, heart rate, breathing) and, on the other hand, the neural correlates (in particular, neural oscillations at different frequency bands) of language processing. The goal when addressing this type of question is to look for correlates between the neuro-physiological signal and typical phenomena of human interaction such as emotion, engagement in the dialogue, convergence effects, information exchange, etc. Research in this direction has been quite limited so far for various reasons. In particular, at the theoretical level, we still do not have a global view on how the different sources of information interact during a conversation in order to build and exchange meaning. Moreover, we need to explore these aspects both at a local level (i.e., a specific moment of the interaction) and the global one (i.e., the entire interaction). However, at this stage there is no precise definition for the type of neuro-physiological correlates that one should look for when studying natural conversation. Thus, when attempting to obtain data that fit both experimental requirements and a more ecological, naturalistic context, it appears to be tricky to choose what type of data to collect, the context and methods of acquisition and the level of control that would allow researchers to obtain meaningful data yet in a more naturalistic environment. As a consequence, designing an experimental setup becomes a challenge. We need to have a full recording of the conversation, making it possible to apply some level of automatic facial expression recognition, to acquire physiological information and the electro-encephalographic signal. Moreover, these different types of information have to be synchronized. These goals raise important methodological issues, in particular, how to do hyper-scanning in a natural context, what type of signal to acquire and how to gather heterogeneous information.

This paper proposes to address these different questions, and presents a new original resource for human language processing studies. In particular, below we describe in greater detail the methodological background of data acquisition and the data pre-processing techniques applied for corpus annotation and data analysis. The corpus we created at this occasion is the first of this kind, bringing together all these different levels of information.

The originality of this project lies in several aspects.

First, we modified more traditional experimental tasks into a real-time and active interaction experience between participants. This direction of research has been still very limited mostly due to methodological issues like speech artifacts in the obtained signal: a lot of research that investigates human-human (verbal) interaction uses experimental designs where participants are not engaged in a natural communicative act (Park et al., 2020). Second, we recorded various types of data, namely, audio, video, physiological and EEG, all of which are crucial when studying natural conversation. It is well-known that when people interact with one another, the nature of this interaction is multimodal that includes language use, gestures and gaze, physiological responses like heart rate and skin temperature and, of course, their neural correlates (Pickering and Garrod, 2021). Thus, looking at only some part of these data might lead to an incomplete picture. That is the reason why we attempted to develop a paradigm that can allow to explore the nature of human communication at its fullest, that is, where multimodal sources of information are recorded during an experimental session. And the third important point is the longitudinal design that was implemented in this project. Thanks to the unique environment offered by the EPSN (see section 5), we recorded the same dyads over the course of two weeks in three different sessions. This gave us the opportunity to see the progression and changes in how our participants aligned over time. To our knowledge, this is the first study of its kind, and our goal is to promote this paradigm for future projects.

2. Scientific Goals

Acquiring and studying conversation in a natural context remains complex in particular because of the heterogeneous nature of the different sources of information to be gathered and analysed. Trying to explore on top of that the neural and physiological correlates of interactional mechanisms is a real challenge. One research direction is to look for relationship between specific phenomena during the time course of a conversation and the neuro-physiological signal. For example, several studies have explored the notion of speakers' engagement, defined on the basis of prosodic, lexical and more generally audio-visual features (Huang et al., 2016; Yu, 2015). What is interesting is that this phenomenon has been correlated with excitement and arousal (Voigt et al., 2014), that can directly be observed at the brain level through modulations of frequency bands (Balconi and Pozzoli, 2009). On their side, different physiological features have also been identified to be associated with emotion and arousal (Londhe and Borse, 2018; Naeem et al., 2012; Monster et al., 2016). We propose that other types of high level interactional phenomena have to be studied in the same perspective, namely, by analyzing features from different modalities and correlating them with brain and physiological signals. Among them, *convergence* between interlocutors' behavior occupies a central place. Interactional theories (Pickering and Garrod, 2021) underline the importance of such alignment mechanisms (also known as linguistic entertainment phenomena (Levitan et al., 2015)) that have been observed at the verbal and non-verbal modalities but also more recently at the brain level (Pérez et al., 2017), showing how neural oscillations progressively get similar during a conversation. Moreover, an important issue lies in the study of the evolution of convergence phenomena not only during a conversation, but also at a larger temporal scale, showing how mutual knowledge influences this mechanism. Concretely, such question requires longitudinal studies, where progression of the mutual knowledge between participants can be observed.

It is then necessary to acquire conversational datasets enriched with annotations (from which feature models can be built) and bearing synchronized electrophysiological information. The difficulty in doing that in mainly technical. The process of acquiring multimodal corpora is nowadays well known, and adding physiological data is not an issue, taking into account the robustness and the simplicity of the equipment. The situation, however, is not the same for the EEG signal. The so-called *second-person neuroscience* approach (Schilbach et al., 2013) aims at elaborating setups that make it possible to design experiments in a natural environment where participants are equipped with EEG headsets. However, this remains a challenge taking into account the sensitivity of the EEG signal to different sources of noise such as gestures and speech. It is then necessary to find a good trade-off between signal quality and the degree of freedom that participants can have during an experiment.

Our goal is then to build an adequate resource for studying the neurophysiological underpinnings of convergence. As explained in section 5, in this project we successfully implemented a multimodal setup that allowed us to acquire rich information of various sources in a longitudinal way: each dyad was recorded three times in three different days. The participants did not know each other before the experiment and spent two weeks in an intensive collaborative project. Their mutual knowledge progressively increased over the 3 experimental sessions, offering a way to study the evolution of the type and level of convergence depending on the participants' proximity level.

3. Related works

Many studies have been done in the perspective of emotion analysis based on corpora recording a set of modalities comparable to ours (audio, video, physiological and neural signals). In the last decade, among others, four such annotated affective databases have been proposed: DEAP (Koelstra et al., 2011), MAHNOB-HCI (Soleymani et al., 2011), DREAMER (Katsigiannis and Ramzan, 2018) and AMIGOS (Miranda-Correa et al., 2021). These resources have been used for train-

ing predictive model of emotion, and more specifically arousal and valence. They offer a set of multimodal features (in particular, from the neuro-physiological sources) in order to classify different types of emotions, confirming experimentally and thanks to machine learning techniques the correlation between multimodal signals and emotions. Note that in a recent study, the same method has been applied to the evaluation of text difficulty levels in a the context of Intelligent Tutoring Systems (Alqahtani et al., 2019).

The protocols proposed in these resources are comparable: they record participants' signals elicited by different stimuli (usually videos) with more or less canonical emotions.Although these studies use a similar instrumental apparatus for recording their data (e.g., two of them (DREAMER and AMIGOS) use "low-cost off-the-shelf devices" similar to ours, namely, the Emotiv-EPOC for recording the EEG signal), they do not involve any interaction between participants or any production at the verbal level, and this is a major difference with our dataset.

In most of these works, the features used in the models rely almost exclusively on EEG/ECG, showing in particular the correlation between valence and frequency bands, negative correlation between arousal and the theta, alpha and gamma bands, and between heart rate and heart rate variations depending on the type of emotion.

Because of the type of the stimuli (no interaction) as well as the focus of the experiments, no features are extracted from the verbal/gestural modalities. For the same reason, the question of the synchronization between the different sources is not specifically described. The database MAHNOB-HCI constitutes an exception by also involving facial expressions and eye-gaze and provides a precise synchronization of the different modalities (eye gaze data, video, audio, neural and physiological signals) obtained thanks to a specific synchronization setup, providing exact temporal relations between events in the different channels.

More recently, a new dataset called K-Emocon (Park et al., 2020) has been made available for emotion detection. It proposes the same type of experimental apparatus as discussed above, but in contrast to the other datasets that rely on passive observation, K-Emocon records participants involved in an active interaction (a debate on the question of the refugee crisis). All different modalities that we mentioned above were recorded for both participants. The apparatus also relies on light devices (Empatica wrist and Neurosky MindWave headset). No specific information is given as for the synchronization of the different sources of information. One main interest of this type of resources as for the neural signal is that they address the question of hyper-scanning (i.e. recording simultaneously the brain activity of two interacting participants) during a natural interaction. Note that natural conversation comes under interest in hyperscanning studies (Nam et al., 2020), but despite the increasing focus on the ecological validity of experiments, such setups remain difficult to analyse (no triggers) thus they are sparsely used.

4. Setup, Instruments

4.1. Material and methods

The interaction between participants is recorded on multiple levels, from acoustic data to EEG data, in order to be able to detect possible convergences between the participants. Convergence or alignment could be observed by analysing the dialogue content itself, but also the behaviour (facial expression, gesture and posture), physiological parameters and cerebral activity of each participant. To record audio and video, both participants were equipped with head microphones (Sennheiser AKG C520), each recorded at 44.1kHz/24 bits by a Zoom H4. Each participant was filmed from the front by a camera located behind and above their interlocutor.

Concerning physiological parameters, participants were equipped with the Empatica E4 wristband, that synchronously records several physiological signals, namely, the blood volume pulse (BVP), the electrodermal response (EDA), the inter bit interval (IBI), the heart rate (HR), the temperature (TEMP), and also behavioural information thanks to a 3 axis accelerometer (ACC).

Beyond the recorded parameters, EDA provides information on sympathetic activation, also called sympathetic arousal. This activation is known to be modulated by emotion. More specifically, the sympathetic activation is said to increase in case of excitement but stress may also increase this activation. Whereas EDA measures the conductance of the skin innervated by the sympathetic nervous system, HR and IBI are modulated by both sympathetic and parasympathetic activations. Factors influencing sympathetic activation are emotion, or stress, whereas parasympathetic activation can be influenced by several factors like, for example, a relaxing situation, or deep breathing as well as the digestion of a big meal.

Even if the recordings of BVP, IBI, HR, EDA, TEMP and ACC are done synchronously, the sampling rate differs depending of the recorded parameter. The photoplethysmography sensor is sampled at 64 Hz, IBI is not strictly sampled but provided at 1/64 sec resolution. EDA, TEMP are sampled at 4 Hz while the 3 axis +/- 2g is sampled at 32Hz. From the three accelerometer signals along X, Y and Z, and given the fact that these three signals are independent, a resulting "activity signal" is calculated with the quadratic sum of the three X, Y, Z signals. Concerning cerebral activity recording, each participants was equipped with an EpocX 14 channels wireless EEG headset, sampled at 128Hz. EEG recordings were synchronized with the presentation script (see section 5) that sent triggers to both EEG headsets.

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
Emotiv EpocX	EEG	128Hz

Table 1: Mobile devices used and data recorded.

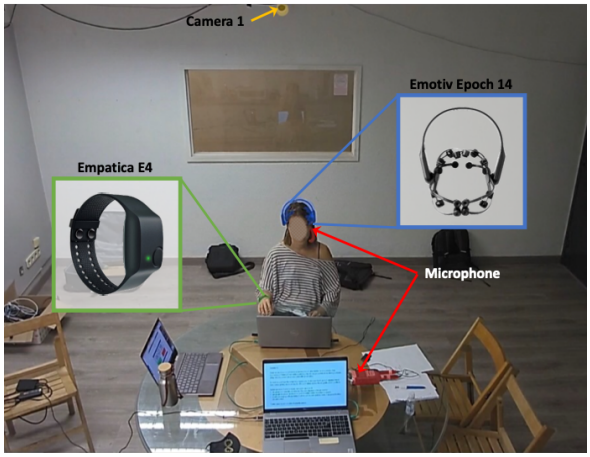

Figure 1: Setup with participant in gear - scene recorded by one of the cameras during the experiment

5. The Experiment

5.1. Participants

The data was collected over a 12 day period in the context of a preparatory workshop for the European Performing Science night (EPSN) in collaboration with the Fundació Épica Fura dels Baus in Badalona, Spain. Over the course of two weeks, 28 performers were to create an art performance inspired by science. The performers had never met one another before the start of the workshop. For two weeks, they rarely interacted with anyone outside of the group, making the workshop a unique setting to study the joint development of multimodal communicative behaviour. For the purpose of the workshop, the performers were divided into 5 groups. Two performers from each of these groups were selected as participants in the data collection. Thus, the 10 participants (6 females) were grouped into 5 dyads, based on their workshop groups. The participants' native language was Spanish, all were right-handed, and the dyads were matched for gender.

5.2. Tasks

The experimental session lasted for approximately 45 minutes and consisted of three tasks - two controlled divergent thinking tasks, the Alternative Uses Test and the Name Invention Task (Guilford, 1967; Agnoli et al., 2016; Fink et al., 2009; Fink et al., 2007) and one free conversation task (Koskinen et al., 2021). Each task is described in greater detail below.

5.2.1. The Alternative Uses (AU) Test

The AU test is one of the tasks used to measure creativity. Participants were shown pictures of ordinary objects and are asked to come up with as many unusual uses as possible. The stimuli were chosen from the MultiPic database (Duñabeitia et al., 2019) that contains 750 lexical items already normalized across six European languages. Of the 750 words, 'number' were chosen to be used in our task (cf. appendix). The order of stimuli presentation was fully randomized.

5.2.2. The Name Invention (NI) Task

The NI task was also used as a measurement of creativity, where participants are presented fictional abbreviations, for example, "K. M.", and are asked to invent an original name consisting of two words that should start with the letters from the prompt, e.g., "Kissing Manual". The stimuli were created with regards to the frequency of syllable onsets in Spanish (Sandoval et al., 2008) and were loosely grouped into more and less frequent letter combinations. A total of 67 abbreviations were included in the task. The order of stimuli presentation was fully randomized.

5.2.3. 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. The discussion was to last for around 10 minutes. A total of three different dilemmas was chosen, one for each recording day.

5.2.4. Procedure

We controlled the timing of the stimulus presentation during the AU and NI task, because we wanted to separate the period of idea generation and the period where the participant would articulate their ideas. The separation of the two actions was of special importance for our EEG measurements, as the signal is very noisy when someone speaks. During one trial, participants first were presented with a fixation cross for five seconds which served as a reference for the brain activity. Participants were then given fifteen seconds to think, and finally eight seconds to articulate the ideas they thought of. The whole task included two practice trials and thirty experimental trials. To create an interactive environment during the controlled tasks, participants took turns responding in the trials, so that participant B listened to participant A responding in one trial, and participant A listened to participant B in the next. In the free conversation task, participants were free to talk in a non-constrained manner.

5.3. Data Collection

Each dyad was tested three times – on days 3, 6 and 9 of the EPSN event. Due to scheduling issues, dyad 5 was tested on Day 10 instead of Day 9. The experiment took place in a room isolated from outside

noise. Participants were seated at a table facing each other. Two laptops were placed back to back on the table, one in front of each participant (see Fig.1). Stimuli were visually displayed on the laptops' screen using an in-house Python script. Two cameras were placed on two walls opposite each other, each one facing one participant. Once the participants were comfortably seated, wearable-devices were placed around each participant's wrist (Empatica E4 wristband, MIT, Cambrdige), recording participants' physiological responses. Then, in order to record participants' brain activity, the portable 14 channel EEG system EMOTIV EPOC (San Francisco, U.S.A.) was placed on their heads. Finally, a head-mounted microphone was placed on them, taking care to not impede the EEG signal, to record their speech during the experiment.

Before the experiment started, instructions were presented to participants on the laptops' screens. Participants were informed that they would start with a creativity-related task (AU or NI) and do the free conversation task after a break. Three dyads completed the AU task on the first and third session while they did the NI task during the second session. The remaining two dyads did the opposite and started with the NI task on the first session. Participants were presented with different items during the second time they did a particular task. The tasks were randomized to avoid order effects. Participants were also asked to minimize unnecessary movements to avoid recording artifacts.

A practice trial was implemented before the creativity-related task while the experimenters stayed in the room, to ensure that the participants understood the task. The real experiment started once participants confirmed they understood the instructions and the experimenters left the room. At the end of the creativity-related task, the experimenters came back to the room and explained the free-speech task to participants in greater detail. The scenario of the dilemma was presented to the participants and the experimenters left the room once participants confirmed they understood the scenario. Participants had ten minutes to discuss and agree on a common answer to the dilemma.

6. Data Pre-processing

The corpus is a record of 5 dyads in interaction over three sessions that represents a total of 12:30 hours of multimodal data.

6.1. Annotations

To lay the ground for future analyses, we generated common audio-video features for our recordings. Conversations were first automatically transcribed and aligned using the BAS Web Services from Ludwig-Maximilians University Munich[1]. The pipeline "ASR G2P CHUNKER MAUS" (Schiel, 1999; Schiel, 2015; Reichel, 2012; Poerner and Schiel, 2018; Kisler et al., 2017) was used with default parameters except for the ASR which was done using Google Speech Cloud ASR[2]. Generated TextGrid files contain one tier for automatic transcription with loose temporal boundaries and four tiers with transcription alignment on the audio signal: two tiers with word-level alignment, one tier with words grapheme-to-phoneme analysis, and one tier with phoneme-level alignment (see Figure 2).

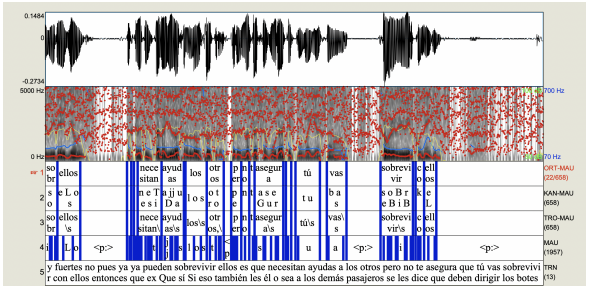

Figure 2: *Automatic transcription and annotations from BASWebServices* (Kisler et al., 2017)

Transcription are to be checked and corrected manually, then realigned using SPPAS (Bigi, 2012) to obtain the corrected word and phoneme levels of alignment. Part of Speech Tagging will finally be applied.

Video analysis pipelines such as FeatureExtraction from OpenFace (Baltrusaitis et al., 2018) is also used to make the most of our multimodal design and generate features from head movements and gaze (see Figure 3). 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.

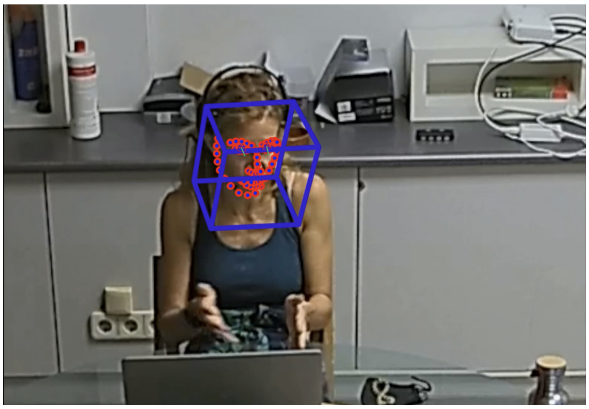

Figure 3: *Facial landmarks and gaze tracking using OpenFace* (Baltrusaitis et al., 2018)

6.2. Synchronization

Synchronization represents an important issue for such a rich multimodal signal. Different simple techniques can be used, based on claps or synchronization signals.

[1] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

[2] https://cloud.google.com/speech-to-text

As described above, in order to capture the best audio quality, separate recordings for audio and videos have been done. The first step consists then in post-synchronizing these signals. An an audio-video clap determines the onsets, which are technically aligned using ELAN (Brugman and Russel, 2004).

The video signal can then be used to synchronize the physiological signal recorded with the EMPATICA wristband. Starting a recording with this material is indicated by a specific LED signal. At the beginning of each session, the device is switched on facing the camera, generating a visual signal (playing the role of a clap) recorded by the cameras. In this case, the video is the common reference.

The EEG signal synchronization concerns two levels: synchronization of the two brain signals from the 2 participants (a classical problem in hyperscanning) and synchronization with the audio-video signal. The difficulty when recording natural conversation is that there is no specific trigger (e.g., a key press) associated to the signal. We solved this problem by generating triggers associated with EEG using the Lab Streaming Layer (LSL) (Kothe, 2014). LSL allows the exchange of time series between devices, programs and computers. It is based on clock offset measurements to handle event information and timing. Synchronization can then be done between devices capable of delivering a data stream output. We automatically generated the triggers at regular intervals by means of a Python script, such triggers being integrated with the EEG signal with LSL.

Finally, the synchronisation of this various data (audio, video, physiological and cerebral activities) is done achieved via the video: as described above, a led of the Empatica wristband lights up when the start button is pressed, the video records the screen activity of each participant (in this way, changes in the activity can be dated), which also permits to synchronize EEG records thanks to a trigger sent in the EEG data via the Python scripted that sequenced the experiment.

Figure 4: *Audio/Video synchronization*

7. Conclusion

New investigations addressing the brain and physiological basis of language processing now require that the natural context of language be taken into account: conversations. In this paper, we have presented the first dataset offering a rich set of sources of information (audio, video, physiological and neural signals), recorded in a natural environment. This first multimodal conversational corpus including neuro-physiological data for spontaneous speech has been enriched in different ways (transcription, alignment) also on the way to be completed (e.g., facial expressions, nods, morpho-syntactic annotations, prosody and phonological annotations). It is being made available through the Ortolang repository (https://www.ortolang.fr/workspaces/badalona-epsn).

8. Bibliographical References

Agnoli, S., Corazza, G., and Runco, M. (2016). Estimating creativity with a multiple-measurement approach within scientific and artistic domains. *Creativity Research Journal*, 28(2):171–176.

Alqahtani, F., Katsigiannis, S., and Ramzan, N. (2019). On the use of ecg and emg signals for question difficulty level prediction in the context of intelligent tutoring systems. In *Proceedings of 19th International Conference on Bioinformatics and Bioengineering*.

Balconi, M. and Pozzoli, U. (2009). Arousal effect on emotional face comprehension: Frequency band changes in different time intervals. *Physiology and Behavior*, 97(3-4):455–462.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66.

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

Brugman, H. and Russel, A. (2004). Annotating multimedia/ multi-modal resources with elan. In *Fourth International Conference on Language Resources and Evaluation*.

Duñabeitia, J., Crepaldi, D., Meyer, A., New, B., Pliatsikas, C., Smolka, E., and Brysbaert, M. (2019). Multipic: A standardized set of 750 drawings with norms for six european languages. *Quarterly Journal of Experimental Psychology*.

Fink, A., Benedek, M., Grabner, R., Staudt, B., and Neubauer, A. (2007). meets neuroscience: Experimental tasks for the neuroscientific study of creative thinking. *Methods*, 42(1):68–76.

Fink, A., Grabner, R. H., Benedek, M., Reishofer, G., Hauswirth, V., Fally, M., Neuper, C., Ebner, F., and Neubauer, A. (2009). The creative brain: investigation of brain activity during creative problem solving by means of eeg and fmri. *Human Brain Mapping*, 30:734–748.

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw Hill.

Huang, Y., Gilmartin, E., and Campbell, N. (2016). Conversational engagement recognition using auditory and visual cues. In *Proceedings of Interspeech*.

Jewitt, C. (2013). *The Routledge Handbook of Multimodal Analysis*. Routledge.

Katsigiannis, S. and Ramzan, N. (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).

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, September.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31.

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

Kothe, C. (2014). Lab streaming layer (lsl). Technical report, https://github.com/sccn/labstreaminglayer.

Levitan, R., Benus, S., Gravano, A., and Hirschberg, J. (2015). Entrainment and turn-taking in human-human dialogue. In *Proceedings of AAAI Spring Symposium*.

Londhe, S. and Borse, R. (2018). Emotion recognition based on various physiological signals - a review. *ICTACT Journal on Communication Technology*, 9(3).

Miranda-Correa, J. A., Abadi, M. K., Sebe, N., and Patras, I. (2021). Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493.

Monster, D., Hakonsson, D., Eskildsen, J. K., and Wallot, S. (2016). Physiological evidence of interpersonal dynamics in a cooperative production task. *Physiology and Behavior*, 156.

Naeem, M., Prasad, G., Watson, D., and Kelso, J. S. (2012). Electrophysiological signatures of intentional social coordination in the 10–12 hz range. *NeuroImage*, 59(2).

Nam, C. S., Choo, S., Huang, J., and Park, J. (2020). Brain-to-Brain Neural Synchrony During Social Interactions: A Systematic Review on Hyperscanning Studies. *Applied Sciences*, 10(19):6669.

Park, C. Y., Cha, N., Kang, S., Kim, A., Khandoker, A. H., Hadjileontiadis, L., Oh, A., Jeong, Y., and Lee, U. (2020). K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293.

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

Pickering, M. and Garrod, S. (2021). *Understanding Dialogue*. Cambridge University Press.

Poerner, N. and Schiel, F. (2018). A web service for pre-segmenting very long transcribed speech recordings. In *Proceedings of LREC*, Miyazaki (Japan).

Rauzy, S. and Goujon, A. (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*, pages 7–pages.

Reichel, U. (2012). PermA and Balloon: Tools for string alignment and text processing. In *Proc. Interspeech*, page 4 pages.

Sandoval, A. M., Toledano, D. T., de la Torre, R., Garrote, M., and Guirao, J. M. (2008). Developing a phonemic and syllabic frequency inventory for spontaneous spoken castilian spanish and their comparison to text-based inventories. In *Language Resource and Evaluation Conference*, pages 1097–1100.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. of the ICPhS*, pages 607–610, San Francisco, August.

Schiel, F. (2015). A statistical model for predicting pronunciation. In *Proc. of the ICPhS*.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., and Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral And Brain Sciences*, 36:393–462.

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

Voigt, R., Podesva, R. J., and Jurafsky, D. (2014). Speaker movement correlates with prosodic indicators of engagement. In *Proceedings of Speech Prosody*.

Yu, Z. (2015). Attention and engagement aware multimodal conversational systems. In *Proceedings of ICMI*.

A. Stimuli

Spanish	English
peine	*comb* guitarra
guitar	
pipa	*pipe*
caña	*fishing rod*
regla	*ruler*
escoba	*broom*
bufanda	*scarf*
linterna	*torch*
embudo	*funnel*
cinturón	*belt*
saxofón	*saxophone*
moneda	*coin*
guante	*gloves*

Number	Letters
1	L.Z.
2	R.Q.
3	G.Z.
4	Q.P.
5	R.Z.
6	F.N.
7	F.J.
8	B.Q.
9	Q.C.
10	C.Q.
11	F.D.
12	P.Q.

Table 2: *The left table presents examples of words used in the Alternative Use task. The name of the image and their English translation are those of the MultiPic database (https://www.bcbl.eu/databases/multipic/). The rightmost table presents example of pairs of letters used in the Name Invention task. Letters are present in more or less frequency*

B. Raw data

B.1. Physiological data

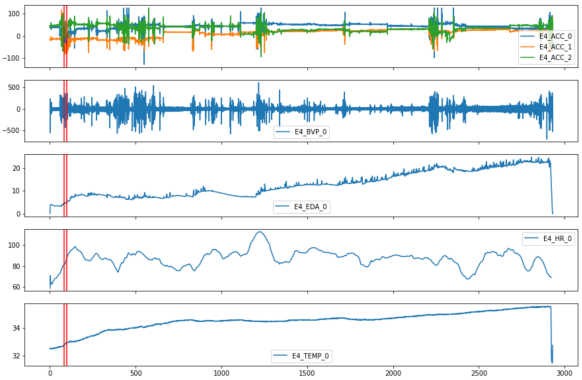

Figure 5: *Visualisation of the data collected by the Empatica E4 Wristband placed on one participant during the first session. From top to bottom: 3-axis acceleration, BVP, EDA, Heart Rate and body temperature. x-axis is in seconds since the start of the watch. Red horizontal lines represent times at which the button is pressed on the watch.*

B.2. Transcription

Excerpt of the automatic transcription of a free conversation:

⟨0⟩ Vale pues
⟨0⟩ yo te diría que en primer lugar pensaría en la profesora de primaria embarazada
⟨0⟩ porque como no saben tampoco cómo van a llegar al ⟨0⟩ o sea
⟨0⟩ Aunque no salte aunque ella se quedase quizás se estrella más el globo
⟨0⟩ y puede ser que esté dañado y dañar a su a sombríon con lo cual puede ser que les esté salvando pero que al final no lo esté salvando porque quizás se hace daño y pierde el crio entonces
⟨0⟩ teniendo un científico que pueda aportar un tratamiento revolucionario o sea que puede salvar muchas vidas
⟨0⟩ Pues quizás
⟨0⟩ bueno estoy entre la profesora y el marido
⟨1⟩ pero el marido es el piloto
⟨0⟩ si el marido es el piloto
⟨1⟩ Entiendo que no puede saltar porque si salta el piloto y nadie sabe pilotar se mueren todos
⟨1⟩ Eso es mi entendimiento
⟨1⟩ sabes? que luego hubiera el típico truco de no hay piloto automático, entonces pues woaw
⟨1⟩ sabes lo que haría?
⟨0⟩ si
⟨1⟩ Bueno, sacaría el piloto por cuestión de supervivencia porque si muere el piloto mueren todos y preguntaría a ellos quien quiere morirse
⟨1⟩ porque yo no sé
⟨1⟩ ser científico con la investigación revolucionaria
⟨1⟩ tiene 75 años por ejemplo y la investigación nunca es una persona sola
⟨1⟩ sea ese puede morir pero su equipo de trabajo va a seguir y su legado va a estar
⟨1⟩ claro pero esto
⟨1⟩ está muy bien pensado

B.3. EEG

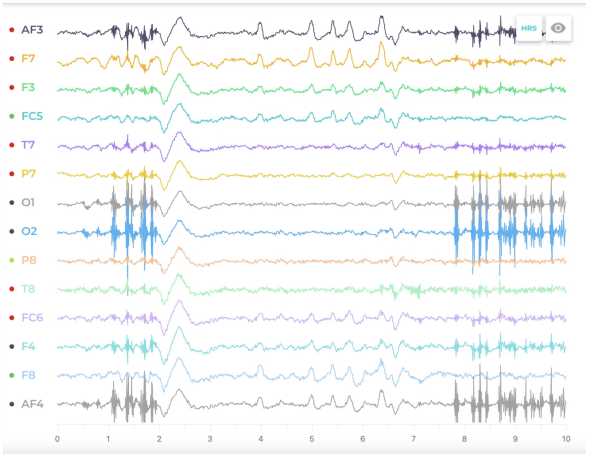

Figure 6: *Visualisation of brain activity using the EmotivPRO app - raw EEG data tab. Left, the color indicates the quality of the signal for each electrode.*