Abstract
Research on language use has become increasingly interested in the multimodal and interactional aspects of language – theoretical models of dialogue, such as the Communication Accommodation Theory and the Interactive Alignment Model are examples of this. In addition, researchers have started to give more consideration to the relationship between physiological processes and language use. This article aims to contribute to the advancement in studies of physiological and/or multimodal language use in naturalistic settings. It does so by providing methodological recommendations for such multi-speaker experimental designs. It covers the topics of (a) speaker preparation and logistics, (b) experimental tasks and (c) data synchronisation and post-processing. The types of data that will be considered in further detail include audio and video, electroencephalography, respiratory data and electromagnetic articulography. This overview with recommendations is based on the answers to a questionnaire that was sent amongst the members of the Horizon 2020 research network ‘Conversational Brains’, several researchers in the field and interviews with three additional experts.
1 | INTRODUCTION

Human communication is a multimodal and interactional phenomenon. It takes place not only through speech/sign production and comprehension, but also through multiple auditory and visual cues that contribute to meaning (Rasenberg et al., 2020). Perniss (2018) cites evidence from several fields of study – from neuroscience to language evolution – that corroborates the idea that the phenomenon we categorise as language cannot be strictly separated from other means of communication, such as gestures and ‘non-linguistic’ vocalisations. Besides the multimodal nature of language, theoretical models of conversation, such as the Communication Accommodation Theory (Gallois et al., 2005; Giles et al., 1991) and the Interactive Alignment Model (Pickering & Garrod, 2004, 2006) have contributed to our understanding of language use as an interactional process. That is to say, it cannot be explained by the simple juxtaposition of comprehension and production processes.

Such theories of dialogue exemplify the increasing interest in the study of communication from different angles. For instance, conversation analysts have drawn many insights from real-world conversational data (see Sidnell, 2010). In addition, naturalistic studies of conversation have considered multimodality – for example gestures (e.g., Brassac et al., 2008; Pouw & Dixon, 2019) and visual cues (e.g., Hanna & Brennan, 2007). They have also investigated factors such as the speakers’ gender (Hancock & Rubin, 2015) and the complexity of the conversation’s environment (Hadley et al., 2021).

Furthermore, the idea of utilising interactional paradigms for physiological-linguistic research has been around for several years – for example investigating articulation (Geng et al., 2013; Lee et al., 2018; Tiede et al., 2010; Vatikiotis-Bateson et al., 2014) and brain activation (Babiloni et al., 2006; Duane & Behrendt, 1965; Montague et al., 2002). However, methodological constraints have limited the extent to which researchers can implement such paradigms. Recent advancements in technological and methodological possibilities have facilitated this growing body of multimodal, interactional and physiological language research. For example, although there has been research on turn-taking dynamics considering body movement and gaze behaviour since the 1970s (La France, 1974), more modern technologies allow for more accurate and detailed measurements (e.g. of breathing: Rochet-Capellan & Fuchs, 2014; eye tracking: Li et al., 2020 and lip movement: Krause & Kawamoto, 2021). Indeed, different research groups have been focussing on the relationship between language use and physiological processes (see, e.g. Hardacre, 2020; Stevanovic et al., 2017, 2019, 2021), and even combining different techniques (e.g. eye tracking and electroencephalography (EEG): Hollenstein et al., 2018).

The aim of the present article is to contribute to the advancement of the study of language as a multimodal and interactive phenomenon. It will address some of the methodological issues to be kept in mind when investigating face-to-face verbal interactions between two or more participants in naturalistic settings.

We propose guidelines that researchers should bear in mind when designing studies with this type of setup. We will make recommendations about (a) participant preparation and logistics; (b) experimental tasks; and (c) data synchronisation and post-processing. The next section will include general recommendations applicable to a range of techniques. Then, we will mention a few techniques in more detail: (a) audio and video recording; (b) respiratory recording; (c) EEG; and (d) electromagnetic articulography (EMA). Despite our limited focus on a few techniques, the interested reader is encouraged to look into the work cited herein. In particular, Hardacre’s (2020) book provides a comprehensive overview of the use of psychophysiological measures in language research. Following the trend of current practices in linguistic research, our focus is on the study
of spoken languages, although some of the recommendations may well also apply to investigations of sign languages. With these recommendations, we hope to facilitate research that can further our understanding of language as a multimodal, interactional phenomenon.

The recommendations presented herein stem from our own experience as well as from responses to a questionnaire distributed amongst researchers who had previous experience with empirical studies using dual-speaker designs. The questionnaire included 13 questions about the following topics: what techniques and experimental tasks the respondents had used in dual-speaker designs; what their biggest challenges were; how they solved issues related to participant logistics, data collection and data post-processing; and what types of recommendations they would find useful to read about. There were a total of 29 answers to the questionnaire. We also drew on insights from interviews we conducted with three additional experts in the collection of neurocognitive and physiological data.

2 | GENERAL RECOMMENDATIONS

In this section we provide recommendations concerning participant logistics, experimental tasks, data synchronisation and post-processing. Note that, throughout this article, the recommendations are oriented towards dual-speaker designs. Many of them may, however, still apply to multi-participant (i.e. more than two) settings.

2.1 | Speaker recruitment and logistics

In multi-speaker experiments, clearly, at least two participants must be recruited. This warrants some special considerations. For instance, the researcher must coordinate the scheduling of two participants and, if necessary, consider how to match them – in terms, for example, of their gender (e.g. Ye & Palomares, 2013) or certain personality factors (e.g. Broz et al., 2012).

There are a variety of ways in which one can recruit participants for their studies. Most commonly, researchers advertise calls for participants through their university’s participant database, mailing lists, flyers posted on campus and social media groups (e.g. Facebook). It is also common to recruit one’s students, often in exchange for course credit. However, it is essential to bear in mind that the way in which one recruits their participants affects what type of population they will be investigating. It is easy to recruit only participants from WEIRD (Western, Educated, Industrialised, Rich and Democratic) communities, thus biasing one’s findings (e.g. Henrich et al., 2010): this should ideally be reduced to a minimum or, at least, acknowledged as a limitation of the investigation.

Some researchers ask their recruited participants to bring a friend to the data collection session, who will then be their conversation partner. This is one topic which researchers must consider: whether their participants should know each other before the experiment. And, in case they do not know each other, whether they should be allowed to have contact immediately before the data collection section. This will clearly depend on one’s research questions. When investigating how friends interact, one would evidently wish to recruit friends (e.g. Slovák et al., 2014). Conversely, if one wishes to study how strangers behave, for example in a speed-dating setting, one must make sure that the first time they see each other is at the very moment their speech is being recorded (e.g. Fuchs & Rathcke, 2018).
A final theme to bear in mind when preparing for a dual-speaker study is what to do in the event that one of the speakers does not come to the scheduled session. Firstly, one can try to avoid this by contacting the participants the day prior to confirm the appointment. If that still takes place, one could simply compensate the participant present for their time and either cancel or reschedule the appointment. To avoid that, however, it is possible to have another eventual single-participant experiment prepared as a backup in which one can ask the speaker present to participate. This way, one can still obtain data for another study. Alternatively, if the investigation in question also has a single-speaker control condition, the individual can also participate in that. Some more complex experimental setups – for example dual EMA – often involve a considerable amount of technical preparation or the presence of multiple lab members. In these cases, one might be even more inclined to avoid participant absence. It might thus be worth it to invite additional speakers who are willing to take part in the study if needed – if their help is not necessary, one should simply give them a compensation for their availability. For experiments with multiple sessions it is advisable to pay the participants after the last session in order to motivate them to return.

2.2 | Experimental tasks

In order to obtain the desired type of speech data, it is paramount to make use of an appropriate task. Since this article’s focus is on language use in interaction, this section presents a few examples of interactive tasks that have been used to elicit speech. These tasks can be reused as they are or can serve as a starting point for the development of new paradigms. In addition, a table can be found in the Appendix containing the tasks herein described as well as additional tasks and an approximate indication of the balance between interlocutors and how controlled the speech material is.

The most common speech elicitation methods amongst our questionnaire’s respondents are either task-oriented or free conversations. For example, the tangram task gives each of the two participants an identical set of pictures, but arranged in different orders. Without seeing each other’s order of images, the speakers must describe the figures and reorganise them into an identical order. This can be done with (more or less) abstract pictures, such as tangram figures (as in Clark & Wilkes-Gibbs, 1986), or it can include images of concrete objects (as in Brennan & Clark, 1996). Another version of the task (Savino et al., 2016) gives one of the participants – the Director – four images with an arrow pointing at one of the pictures, and one single image to the other speaker – the Matcher. Herein, without seeing each other’s pictures, their job is to decide whether or not the Matcher’s figure is identical to the Director’s highlighted image. Another widely used task is the Diapix, developed by Van Engen et al. (2010) (see also Baker & Hazan, 2011). It provides each participant with one figure, each of which looks almost identical, but contains a few differing details. The interlocutors cannot see the other’s image, and they must talk about the pictures to identify all their differences. Figure 1 displays an example of such a picture pair. This task has been found to be more balanced regarding the amount of contributions of each participant than the older, widely used map task setup (Anderson et al., 1991). Again, the two participants receive two versions of a similar picture – herein, a map. The instruction Giver’s map contains a dotted path, whereas the instruction Follower’s does not. The Follower must trace the path according to the Giver’s instructions using a pencil. The Giver’s and the Follower’s maps have slight differences, but the participants are not informed of this, which makes the task more...
difficult. In a similar task described by a respondent of our questionnaire, one participant sees a drawing and must verbally instruct their conversational partner to reproduce the drawing.

In a more controlled interactive task called the Domino task (Bailly & Lelong, 2010), participants see pairs of words on a computer screen and must say the word whose first syllable matches the last syllable of the word produced previously by their partner. In the example illustrated in Figure 2, participant A sees and produces the word delay. Then, participant B sees two words: the target word – layout – and a random word – here, break – and must choose the word layout, since lay- in layout matches -lay in delay. Participant A must then produce outside, and so
on. The Domino task has been used, for instance, to investigate phonetic convergence (Bailly & Martin, 2014; Mukherjee et al., 2017), including its neural oscillatory markers, using EEG (Mukherjee et al., 2019), and articulatory movements, using EMA (Mukherjee et al., 2018).

Finally, several researchers also ask participants to have free conversations, without orientation towards a task. Some of those conversations suggest a topic for discussion (e.g. food, holidays and hobbies), whilst others do not. Providing some topics for free conversation is especially helpful for participants who are less talkative and might not know how to start otherwise; it can also be useful in controlling the degree of emotional involvement. On the contrary, certain topics such as religion, politics, and life and death can evoke highly emotional responses, body gestures, and probably also physiological changes (Labov, 1981, 2013).

The choice of speech elicitation method will largely depend on the study’s goals, on the population studied, on the technique used and data artefacts to be avoided. For instance, one of our survey’s respondents noted that the Diapix task elicits more balanced conversations than the Map task (cf. Pardo, 2006). Furthermore, describing concrete objects or abstract tangram figures will involve different types of words, which might have consequences for a study that, for example, investigates lexical or syntactic alignment (e.g. Nenkova et al., 2008). One might also research the effect of different conversation topics, for which they might wish to use free conversations with topic prompts (e.g. Stevanovic et al., 2019). Some less interactive tasks, such as scripted dialogues (McFarland, 2001), joint picture naming (Shockley et al., 2007) or coordinated reading aloud of words, may be warranted, for example, by EEG studies (Mukherjee et al., 2019), which require more control over the speech output. Finally, as one respondent of our questionnaire remarked, if one’s study concerns children, they might find that completely spontaneous conversations are not as productive as more constrained tasks.

2.3 | Data synchronisation and post-processing

When working with multimodal datasets, generated during human interaction, the researcher must make sure that all the data are temporally synchronised. One must distinguish between the synchronisation of the different data streams referent to each participant, and between-participant synchrony. Various data streams stemming from multiple speakers or additionally from multimodal data (e.g. acoustic and breathing data or acoustic and motion capture data) are common in dual-speaker experiments. The simplest and least time consuming way in terms of post-processing might be recording data on one computer using a data acquisition card. The acoustic signal of each speaker as well as other types of data can be stored on separate channels. However, using several computers may lead to synchronisation issues.

2.3.1 | Synchronisation issues using several computers

Whenever two or more computers are used in the recording session, synchronisation might be required because each computer can have a different internal clock determining the sampling frequency and timestamps associated with the data in its memory. These different internal clocks can lead to synchronisation issues because with longer recording times, the two signals drift more and more apart. In rare cases, the sampling rate might not be precisely the one which one defines in the software. Some pilot recordings are advisable.
2.3.2 | Potential solutions

Some manufacturers provide hardware solutions, so that their system can either receive or send data of selected formats from other devices – such as the synchronisation device from OptiTrack. More customised solutions may also be possible. If an experimental setup is used frequently, it makes sense to run some tests to capture the delays amongst the internal clocks of the involved machines. For this purpose a defined signal, such as a sinusoidal waveform, can be recorded in parallel on the different devices and over shorter and longer time periods accounting for the cumulative effect of delays over time. These delays need to be taken into account when pre-processing the data.

Devices from other areas, such as the Arduino board (Smith, 2011), often used in the game industry, have also been adopted for synchronising data streams coming from distinct devices with different delays. The Arduino board is relatively cheap, and both its hardware and software are open source. It is a microcontroller which has an integrated development environment for the user. It runs a single programme at a time, has memory storage and can be connected to analogue or digital technical devices. To use it for synchronisation, one can programme it in such a way that the board waits for the input signals of several computers. Only when all signals have arrived does it allow all data streams to pass. Using the Arduino board in such a way guarantees that the onset of data streams coming from different PCs with different internal clocks is precisely synchronised.

Alternatively, one can also send a synchronisation pulse, such as a prominent rectangular pulse with a steep onglide to start and/or to end the recording to the different computers. This will allow automatic detection of these time stamps and later post-synchronisation of the data by means of cutting files at the time stamps. Both this procedure and the Arduino board can be used to synchronise different signals coming from different participants as well as a single speaker’s own speech data with their other types of data.

Another common procedure is that the audio signal is recorded on two computers. Using cross-correlation of the two audio signals after the recording session is another way to post-synchronise the data, because it provides exact information on how much one audio signal is delayed with respect to the other. However, it also requires sound separation of the two speakers and some computational expertise and may be demanding for an absolute beginner. Finally, if the sampling frequencies are different and adjustable amongst the devices, it is advisable to fix them in a relation such that one is an integer multiple of the other – this makes cross-channel synchronisation considerably easier.

3 | RECOMMENDATIONS FOR SPECIFIC TECHNIQUES

This section contains more detailed recommendations concerning some selected techniques commonly used in the study of language use (see also Hardacre, 2020). In the Appendix, a table can be found that summarises information about different techniques – both those mentioned in this section and some other commonly used ones.
3.1 Audio and video

According to our questionnaire’s responses, the most common dual-speaker setup involves making only audio or, in addition, video recordings of the interlocutors. Although the equipment involved is not complex to use, some issues still need to be borne in mind. For instance, the researcher must decide how to record the data. Stereophonic (a.k.a. stereo) recordings are recommended over monophonic (a.k.a. mono) ones. The former have two channels, which allows for two microphones to be recorded synchronously, facilitating the postprocessing of the data. In addition, to minimize the extent to which one speaker’s microphone captures their partner’s speech, one can make use of microphones with cardioid directionality. These capture signals coming from the front of the microphone well, but is not sensitive to sounds coming from the back (see Švec & Granqvist, 2010). For further technical recommendations concerning types of microphones, see Švec and Granqvist (2010), and for considerations about microphone calibration, see Švec and Granqvist (2018). Finally, one must consider the microphone’s positioning (e.g. head-mounted or stand-mounted) if one wishes to investigate voice intensity. Švec and Granqvist (2018) point out that head-mounted microphones are more likely to stay at a fixed distance from the mouth, thus being more accurate than stand-mounted equipment. One must make sure, however, that the head-mounted microphone is stable and will not move in relation to the participant’s head, because the distance to the mouth is so small that even a 0.5 cm change can impact the audio signal, whilst the impact is smaller for a stand-mounted microphone.

When it comes to video recordings, a few points also have to be considered. Firstly, one must find a position in which all participants are part of the frame. Having the participants sit next to each other partly facing the camera seems to produce adequate video data that allows for the analysis of their gestures and facial expressions – see the setup in Hardacre (2015), illustrated in Figure 3; see also Fujiwara and Yokomitsu (2021) or, for a different setup, Bailly and Lelong (2010). Secondly, in most cases, videos are recorded at a much lower sampling rate than is acoustic data. This may not be a problem when studying manual gestures, but we do not recommend using

![Figure 3](An example of a multi-speaker video study in which all participants are caught on a single frame (Hardacre, 2015, figure reproduced with permission of the author))
video to study faster (e.g. lip) movements – in these cases other techniques might be more appropriate, such as EMA (see subsection 3.4) and motion capture systems. A standard video format often has a frame rate of 24 Hz and records audio data at 48 kHz (note that 48,000 is an integer multiple of 24). If the researcher is only recording audio data, a sampling rate of 22.05 kHz is the minimum to accurately capture spectral characteristics of fricatives (e.g. Fox & Nissen, 2005).

Finally, there are a number of tools available that can detect body movement from video, allowing for quantitative analysis of gestures in conversation (e.g. Fujiwara & Daibo, 2016; Tschacher et al., 2014). As Pouw et al. (2020) explain, these tools work either with the detection of pixel changes or with deep learning. Pixel-based techniques calculate the speed and direction of movements by tracking the change of brightness in the pixels in a frame-by-frame procedure. Deep learning methods detect the position of individual body parts in the scene and are able to assign each body part to its corresponding individual. This might make deep learning techniques more suitable for dual-speaker videos, since it does not matter if the participants’ images overlap. If they do not overlap, though, pixel-based movement detection may also be sufficient. Fujiwara and Yokomitsu (2021) have compared the two techniques and found that both provide robust and comparable body movement data. Thus, the choice of which to use depends on factors such as the quality and nature of the video recordings, which is dealt with differently by each method (for a discussion of the advantages and limitations of each technique, see Fujiwara & Yokomitsu, 2021; Pouw et al., 2020).

3.2 Respiratory inductance plethysmography

Measuring respiration involved in human interaction is relatively easy to carry out in comparison to other physiological techniques that require intensive preparation. There are more and more studies using this technique in dual- or multiple-speaker designs (Bailly et al., 2013; Fuchs & Rathcke, 2018; Ishii et al., 2014; McFarland, 2001; McFarland et al., 2020; Müller & Lindenberg, 2011; Rochet-Capellan & Fuchs, 2014; Torreira et al., 2015; Warner et al., 1983; Włodarczak et al., 2015; Włodarczak & Heldner, 2020).

In inductance plethysmography, two flexible respiratory belts are wrapped around the participants’ upper trunk, one around the rib cage and the other around the abdomen (for an overview on breathing and speech, see Fuchs & Rochet-Capellan, 2020). Preferably, participants wear thin, tight clothes and for the participant’s and experimenter’s comfort, the belts are put on top of the clothes. Some systems allow to adjust the size of the belts to the respective size of the participant (e.g. Columbi Computers’ RespTrack) whilst for others (e.g. Ambulatory Monitoring’s Inductotrace System) flexible belts of different sizes are available, from small children to adult participants with different Body Mass Indices. There have also been studies on multi-party conversations where only one respiratory belt was used for each speaker (Ishii et al., 2014). Whilst one belt does not reflect changes in overall lung volume, it may still provide valuable information about temporal aspects of breathing in conversation. According to our experience, speakers would always use the rib cage for speech breathing, but show different degrees of abdominal breathing (Mitchell et al., 1996); we would thus recommend the use of the rib cage belt if only one belt can be used.

Whilst breathing is relatively easy to record and it occurs during all activities, there are also certain issues to consider. Firstly, breathing can be controlled voluntarily by the participant who is at any time aware that their breathing is recorded. Hence, engagement in the experimental task is important and can reduce other potential sources of influence. Breathing is also affected by
participants’ attention, emotion (e.g. attraction to the interlocutor), cognitive load (Grassmann et al., 2016; Homma & Masaoka, 2008; Shea et al., 1987) or even their circadian rhythm (Spengler et al., 2000). Some of these influences might be reduced by familiarity with the laboratory, the experimenters, the equipment or recording at a specific time of the day.

Whilst breathing has a large flexibility and often differs in specific parameters between participants, within-individual consistency has also been observed for a given task (Benchetrit et al., 1989; Shea & Guz, 1992; Serré et al., 2021) over several days or even years. For this reason it might be advisable to favour a within-subject design – for example speaking in one condition to a familiar interlocutor and in another condition to an unfamiliar one. However, the experimental design is crucially dependent on the research question.

It must also be kept in mind that interlocutors may gesture during their conversation, which can in some cases cause perturbations of the respiratory belts, in particular the rib cage belt. Restricting the participants’ arm motion, for example by asking them to leave their hands on a table or chair, might be one solution, but may not be optimal, because gesturing is part of the conversation. Body sways or head turning can also lead to perturbations depending on the magnitude and direction of the motion. It may affect the offset of the respiratory signal. What worked well in our own recordings was telling the interlocutors they could use gestures, but should avoid large movements, if possible. Additionally, we recommend recording the whole setup with a video camera that protocols when gestures with particularly large amplitudes are made. Joint laughter may be a common feature in dialogue. It looks like a perturbation at first because lung volume may drop to an absolute minimum – see Figure 4. However, laughter is not a perturbation, but often goes hand in hand with a rapid and exhaustive exhalation down to the residual volume (Filippelli et al., 2001).

If, in one’s experiment, joint motion is one of the tasks besides communication, it might be useful to pilot the task, investigate potential perturbations and make sure that the wires connected to the respiratory belts do not hinder motion at any time. In some tasks one belt may be more affected than the other. For instance, pedalling on a stationary bike (e.g. Fuchs et al., 2015) may show motion oscillations on the abdominal belt due to motions of the hips, but may be less perturbing on the rib cage belt. Individual differences can also occur; thus, it is advisable to pilot the experiment with several participants.

Depending on the duration and the task of the experiment and the age of the participants, either a sitting or standing position may be preferred. A sitting position may be more comfortable especially for older participants, but can constrain inhalation more than a standing position. Sitting in a relaxed position for a while may lead to positional changes in the abdominal belt. Standing has the advantage that participants can fully inhale (Katz et al., 2018), but has the drawback that participants produce frequent body sways, which can sometimes produce artefacts.

Inductance plethysmography can be combined with a variety of other techniques, such as EMA (Rasskazova et al., 2019), motion capture (Fuchs & Rathcke, 2018), eye tracking or electroencephalography (Perl et al., 2019), but most of the work carried out so far has only been used for single speakers.

### 3.3 Electroencephalography

One of the most widely used neuroimaging techniques is EEG. Electrodes placed outside the head on the scalp record weak electrical activity as produced by the brain. Simply put, it allows researchers to investigate, at a high temporal resolution, neurological processes in humans (for
FIGURE 4  Example of the respiratory kinematics of two interlocutors during speed dating (data are taken from the Fuchs & Rathcke (2018) study). Left: photo by Andres Ayrton via Pexels. Right: periods of speech (exhalation) are marked in blue and bouts of laughter (exhalation and subsequent inhalation) are marked in brown. The line’s upward movement indicates inhalation, and its downward movement indicates exhalation. Data of the two interlocutors are given in top and bottom panel.
an introduction to EEG, see Dickter & Kieffaber, 2013; for an introduction to event-related potentials, ERPs, see Luck, 2014).

A delicate issue for EEG studies is their experimental task. Facial and eye movements related to speaking greatly influence the electrical activity on the skin of the head, creating motion artefacts in the EEG signal that can overshadow the data of interest. One must thus take that into account when designing tasks (see Section 2.2) for EEG studies. Furthermore, since EEG data collection involves a considerable amount of time and effort, it is advisable to pilot the experiment thoroughly to ensure that it elicits the desired type of data before one recruits participants.

It is time-efficient to have two lab members prepare each of the two participants at the same time (i.e. two experimenters per participant) whilst a fifth researcher explains the procedure and experiment to them. In addition, it is advisable to have the same lab members perform the same tasks (e.g. preparing electrodes or explaining the procedure to the participants) in all experimental sessions. With the increased fluency in the task gained by practice, not only do the sessions run more smoothly, but the participants are reassured by the experimenters’ confidence.

Dual-EEG research can take one of two approaches. It can analyse the two EEG signals in isolation, investigating how each participant responds to the interaction (e.g. Ménoret et al., 2014), or it can investigate neural coupling. Indeed, several hyperscanning studies reported that communication enhances synchronisation between the interactants’ brain activity (for reviews, see Liu et al., 2018; Nam et al., 2020), although interpretation of the data must be taken with caution (Burgess, 2013; Hamilton, 2021). For instance, one must make sure that any potential neural coupling observed is due to actual entrainment, and not because of artefacts, such as similar brain activity in response to stimuli present in the environment. There is as of yet no consensus on best practices for EEG hyperscanning analysis, although Ayrolles et al. (2021) have initiated efforts in that direction. Furthermore, Barraza et al. (2019) provide guidelines on the implementation of EEG hyperscanning setups. When it comes to EEG data analysis more broadly, some of the most commonly used tools are MNE-Python (Gramfort et al., 2013, 2014) and the MATLAB toolboxes FieldTrip (Oostenveld et al., 2011) and EEGLAB (Brunner et al., 2013).

3.4 Electromagnetic articulography

One problem with dual or multi-speaker audio recordings is the source separation, since speakers overlap and talk at the same time. One way to extract speech data from a single speaker in a multi-speaker environment is to record their articulatory movements. Apart from this, it also provides crucial information about speakers’ actions during phases of silence and how and when they prepare for taking the turn whilst their interlocutor is still speaking. One of the most popular techniques for recording speech movements within the mouth is EMA (for recent overviews on different systems and methods, see Cho & Mücke, 2020; Rebernik et al., 2021). This method is based on the physical principle of electromagnetic induction whereby an electromagnetic field is generated by several transmitters. If a coil is positioned within this electromagnetic field, a current is induced. The strength of the coil’s induced voltage is inversely proportional to the distance from the transmitters. Since the transmitters are running on different frequencies, the x, y and z coordinates of the coil can be estimated, together with two orientation angles. Small sensors, containing a coil, can be attached to moving articulators within the mouth, for example the tongue, the lips and the jaw. Until recently, two manufacturers offered EMA systems for speech research: Carstens Medizinelektronik, in Germany and Northern Digital Inc. (short NDI), in Canada. The latter discontinued the production of this branch in early 2020.
In the last four decades a huge amount of research on read speech of single participants provided the field of linguistics with crucial insights in the kinematic and dynamic mechanisms underlying speech production (for an extensive overview on research and applications of this method as well as sensor handling, see Rebernik et al., 2021). Herein, we will focus on dual EMA experiments for which research with three different setups have been published. The first research group to undertake a dual EMA experiment in Edinburgh used two identical 3D EMA machines of the type AG500, Carstens Electronics (Geng et al., 2013). Because of the interference between the two machines, they had to be 6 m apart in two different rooms allowing for no visual contact between the two speakers. Later experiments either used machines from different manufacturers, such as an AG500 (Carstens Electronics) and an NDI Wave system (see Tiede et al., 2010; Tiede & Mooshammer, 2013; Vatikiotis-Bateson et al., 2014) or the newer and much improved system AG501 and NDI Wave (Mukherjee et al., 2018). Newer setups take advantage of the possibility of running two identical machines on different frequencies, provided by the manufacturers. For example in the laboratory at USC, Los Angeles, two NDI wave systems are used for dual speaker experiments (see Lee et al., 2018), and in the laboratory at the Humboldt-Universität zu Berlin, with two AG501 systems. Even though for both setups the interference is minimised, for example by using different frequencies, a minimal distance of 2 m between the two machines is recommended. Informal tests showed that below this distance the measurement error increases.

Apart from the problem of interference between two EMA machines there is also synchronisation between the two machines to consider. The dual NDI system provided a hardware solution whereas other combinations needed individual solutions with external synchronisation impulses and post-processing with custom-made scripts in MATLAB.

Before the experiment, all sensors need to be prepared, that is, sterilised, often covered with latex and in some cases glued to silk (for a recent comparison of common practices and recommendation for sensor preparation, see Rebernik et al., 2021). Once the sensors are dry, they can be attached to the articulators of the participants. Herein, it has to be kept in mind that for single-subject EMA experiments two experimenters are needed. Depending on how experienced that team is at working together, three persons might be sufficient for handling the adhesive, the tweezers and taping the wires for protection in dual-speaker experiments. Otherwise, four people are strongly recommended. During the experiment one experimenter is needed for controlling the computers and one should be present in case sensors come off and have to be re-glued. As for single-subject EMA experiments, the sensors have to be cleaned after the experiment and the data have to be post-processed: calculation of the coordinates, subtraction of head movements and orientation to the bite-plane. Additionally, the data have to be synchronised and, depending on the analysis software, the acoustic and movement data are aggregated into a single file.

4 | CONCLUSION

This article has summarised some considerations to be borne in mind when carrying out multi-speaker experiments, mentioning different approaches, general recommendations and specific techniques. Crucially, there are no golden standards or ultimate solutions for most of the issues raised herein, although researchers have come up with some creative solutions. We hope our recommendations will encourage interested researchers to study language in interactive settings. In general terms, it is essential to take the time to pilot one’s experiment well and test all the equipment involved. It is also advisable to invest energy into learning how to customise scripts and use
command-line tools to process data. Ultimately, there needs to be an interdisciplinary effort to create solutions that take advantage of the expertise of researchers from different fields. This will hopefully produce significant advancements in our understanding of language as a multimodal and interactional phenomenon.

ACKNOWLEDGEMENTS
The authors would like to thank Alessandro D’Ausilio, Mark Tiede and Phil Hoole for providing expert insights and tips, and Noel Nguyen and Alessandro D’Ausilio for their helpful comments on the manuscript. This project has received funding from the European Union’s Framework Programme for Research and Innovation Horizon 2020 (2014–2020) under the Marie Skłodowska-Curie Grant Agreement No. 859588.

Open Access funding enabled and organized by Projekt DEAL.

ORCID
Tom Offrede https://orcid.org/0000-0002-9481-8473
Susanne Fuchs https://orcid.org/0000-0001-6751-9286
Christine Mooshammer https://orcid.org/0000-0002-7836-1741

REFERENCES


### AUTHOR BIOGRAPHIES

**Tom Offrede** is a PhD candidate at the Institut für deutsche Sprache und Linguistik at the Humboldt-Universität zu Berlin. He researches the interaction between human physiology and conversational dynamics.

**Susanne Fuchs** is a senior researcher at Leibniz-Zentrum Allgemeine Sprachwissenschaft in Berlin. She researches the multimodal production of human interaction with a focus on physiological processes.

**Christine Mooshammer** is a professor of Phonetics at the Institut für deutsche Sprache und Linguistik at the Humboldt-Universität zu Berlin. Her research interests are speech production, speech errors, speaking styles and prosody.

---

## APPENDIX

**TABLE A1** Summary of tasks that have been used in dual-speaker studies

<table>
<thead>
<tr>
<th>Task</th>
<th>References</th>
<th>Brief description</th>
<th>Freedom in speech output</th>
<th>Between-speaker balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint character description</td>
<td>Stevanovic et al. (2017) and Stevanovic et al. (2021)</td>
<td>The participants are told to imagine they are writing a children’s book to teach the alphabet. The book features characters, either from real life or fictional. The participants must jointly choose eight adjectives – each starting with a different letter – for each character.</td>
<td>Free</td>
<td>Yes</td>
</tr>
<tr>
<td>Free conversation</td>
<td>Belz et al. (2021)</td>
<td>The participants are instructed to have a free conversation, either about any topic they would like or about prompt topics given.</td>
<td>Free</td>
<td>Yes</td>
</tr>
<tr>
<td>Diapix</td>
<td>Baker and Hazan (2011) and Van Engen et al. (2010)</td>
<td>Each participant sees one figure, each of which looks almost identical, but contains a few differing details. The interlocutors cannot see the other’s image, and they must talk about the pictures to identify all their differences.</td>
<td>Fairly free</td>
<td>Yes</td>
</tr>
<tr>
<td>Tangram task</td>
<td>Brennan and Clark (1996) and Clark and Wilkes-Gibbs (1986)</td>
<td>Each participant sees an identical set of pictures, but arranged in different orders. Without seeing each other’s order of images, the speakers must describe the figures and reorganise them into an identical order.</td>
<td>Fairly free, with lexical constraints</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Note: A brief description of their procedure, how controlled the speech output is and how balanced the interaction is. This table contains the tasks mentioned in Section 2.2 as well as additional tasks.*
### TABLE A2  Summary of tasks that have been used in dual-speaker studies

<table>
<thead>
<tr>
<th>Task</th>
<th>References</th>
<th>Brief description</th>
<th>Freedom in speech output</th>
<th>Between-speaker balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree decoration</td>
<td>Ito and Speer (2006)</td>
<td>The Director sees instructions on where to put each ornament on a tree. They tell the Decorator how to decorate the tree. The Director and Decorator are separated by a wall, but there is a window that lets the Director see the Decorator's hands and the tree.</td>
<td>Fairly free, with lexical constraints</td>
<td>No</td>
</tr>
<tr>
<td>Map task</td>
<td>Anderson et al. (1991)</td>
<td>Each participant sees a different version of a map. The Instruction Giver's map contains a dotted path, whereas the Instruction Follower's does not. The Follower must follow the Giver's instructions to trace the path with a pencil.</td>
<td>Fairly free, with lexical constraints</td>
<td>No</td>
</tr>
<tr>
<td>Matching task</td>
<td>Savino et al. (2016)</td>
<td>The Director sees four images with an arrow pointing at one of them, and the Matcher has one image. They must talk to decide whether or not the Matcher's figure is identical to the Director's highlighted picture.</td>
<td>Fairly free, with lexical constraints</td>
<td>No</td>
</tr>
</tbody>
</table>

### TABLE A3  Summary of tasks that have been used in dual-speaker studies

<table>
<thead>
<tr>
<th>Task</th>
<th>Reference</th>
<th>Brief description</th>
<th>Freedom in speech output</th>
<th>Between-speaker balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word repetition</td>
<td>Tiede et al. (2010)</td>
<td>Participants are asked to produce a two-word sequence (e.g. 'cop top') repeatedly over the course of 30 s. The participants are instructed not to start at the same time and to produce different sentences (e.g. 'cop top' vs. 'top cop').</td>
<td>Not free</td>
<td>Yes</td>
</tr>
</tbody>
</table>

(Continues)
<table>
<thead>
<tr>
<th>Task</th>
<th>Reference</th>
<th>Brief description</th>
<th>Freedom in speech output</th>
<th>Between-speaker balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domino task</td>
<td>Bailly and Martin (2014)</td>
<td>Participants see pairs of words and must say the word whose first syllable matches the last syllable of the word their partner has just pronounced.</td>
<td>Not free</td>
<td>Yes</td>
</tr>
<tr>
<td>Maze task</td>
<td>Lee et al. (2018)</td>
<td>The participants see a maze with a marked path, but each participant only sees part of the path. The participants must go through the maze describing its landmarks with set frame sentences, taking turns to describe the parts of the path that each one can see.</td>
<td>Not free</td>
<td>Yes</td>
</tr>
<tr>
<td>Music game</td>
<td>Wieling et al. (2020)</td>
<td>The participants must play melodies together by producing the syllables /ka/, /ki/, and /ku/. They can also communicate only with these syllables. The director sees the melodies to be played and must instruct the follower. They do not see each other, and they switch roles every turn.</td>
<td>Not free</td>
<td>Yes</td>
</tr>
<tr>
<td>Board game directions</td>
<td>Schafer et al. (2000)</td>
<td>Each participant sees a chess-like board with pieces on it. The Driver knows the goals for the pieces to get to and the Slider knows the bonuses and dangers on the way. They use a fixed set of sentence frames and game pieces to give instructions, requests and acknowledgements.</td>
<td>Not free</td>
<td>No</td>
</tr>
</tbody>
</table>
## Summary of techniques

### TABLE A4  Summary of information concerning acoustic, video and respiration recordings

<table>
<thead>
<tr>
<th></th>
<th>Acoustics</th>
<th>Video</th>
<th>Respiration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to prepare equipment</td>
<td>5–10 min</td>
<td>5–10 min</td>
<td>5 min</td>
</tr>
<tr>
<td>Time to prepare participants</td>
<td>5–10 min</td>
<td>5–10 min</td>
<td>10 min</td>
</tr>
<tr>
<td>Lab members to prepare participants</td>
<td>1–2</td>
<td>1–2</td>
<td>1</td>
</tr>
<tr>
<td>Lab members to record data</td>
<td>1</td>
<td>1</td>
<td>1–2</td>
</tr>
<tr>
<td>Computers to record data</td>
<td>1</td>
<td>1</td>
<td>1–2</td>
</tr>
<tr>
<td>Postprocessing software</td>
<td>Audacity; SoX; MATLAB; Python; Praat</td>
<td>FFmpeg; R; MATLAB</td>
<td>Python; R</td>
</tr>
<tr>
<td>Annotation software</td>
<td>Praat; ELAN; MATLAB; Excel</td>
<td>ELAN; FFmpeg; Praat; Excel</td>
<td>Praat; ELAN</td>
</tr>
</tbody>
</table>

**Note:** Time needed and number of lab members needed to prepare the equipment and participants; number of lab members and computers needed to record data during the experimental session; and software used for data postprocessing and annotation. The time frames for preparation indicated are based on estimates by the respondents of our questionnaire and correspond to the time needed by experienced researchers. Furthermore, the time indicated for equipment preparation corresponds to the approximate time needed before each recording session. It assumes that the devices have already been installed and calibrated as necessary per each given technique. Finally, the software mentioned does not include all the tools available for working with the data; again, they represent the answers given by our questionnaire’s respondents.

### TABLE A5  Summary of information concerning eye-tracking, motion capture and EMA recordings

<table>
<thead>
<tr>
<th></th>
<th>Eye tracking</th>
<th>Motion capture</th>
<th>EMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to prepare equipment</td>
<td>5–10 min</td>
<td>10 min</td>
<td>1 hr</td>
</tr>
<tr>
<td>Time to prepare participants</td>
<td>5–10 min</td>
<td>5 to 10 min</td>
<td>30–45 min per participant</td>
</tr>
<tr>
<td>Lab members to prepare participants</td>
<td>1–2</td>
<td>1–2 per participant</td>
<td>2 per participant</td>
</tr>
<tr>
<td>Lab members to record data</td>
<td>1</td>
<td>1</td>
<td>2–3</td>
</tr>
<tr>
<td>Computers to record data</td>
<td>2</td>
<td>1</td>
<td>2–3</td>
</tr>
<tr>
<td>Postprocessing software</td>
<td>R; MATLAB</td>
<td>Python; MATLAB</td>
<td>Python; MATLAB</td>
</tr>
<tr>
<td>Annotation software</td>
<td>ELAN; MATLAB; Excel</td>
<td>Praat; ELAN</td>
<td>Python; MATLAB (e.g. MView); VisArtico</td>
</tr>
</tbody>
</table>
### TABLE A6  Summary of information concerning EEG, skin conductance and fNIRS recordings

<table>
<thead>
<tr>
<th></th>
<th>EEG</th>
<th>Skin conductance</th>
<th>fNIRS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time to prepare equipment</strong></td>
<td>20 min</td>
<td>5–10 min</td>
<td>30 min</td>
</tr>
<tr>
<td><strong>Time to prepare participants</strong></td>
<td>ca. 30 min per participant</td>
<td>5–10 min</td>
<td>20 min</td>
</tr>
<tr>
<td><strong>Lab members to prepare participants</strong></td>
<td>2 per participant</td>
<td>1–2</td>
<td>2 per participant</td>
</tr>
<tr>
<td><strong>Lab members to record data</strong></td>
<td>1–2</td>
<td>1–2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Computers to record data</strong></td>
<td>2–3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td><strong>Combine data</strong></td>
<td>Python; MATLAB</td>
<td>Python; MATLAB</td>
<td>MATLAB</td>
</tr>
<tr>
<td><strong>Postprocessing software</strong></td>
<td>Python; MATLAB</td>
<td>Python; MATLAB</td>
<td>MATLAB</td>
</tr>
<tr>
<td><strong>Annotation software</strong></td>
<td>MATLAB</td>
<td>Python; MATLAB; ACKNOWledge</td>
<td>MATLAB; Equipment’s software</td>
</tr>
</tbody>
</table>