Combined eye tracking and electroencephalography during referential selection in dyadic interaction

Ingmar Brilmayer*, Institute for German Language and Literature I, Linguistics, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany, ingmar.brilmayer@uni-koeln.de

Philip Georgis, Institute for German Language and Literature I, Linguistics, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany

Petra B. Schumacher, Institute for German Language and Literature I, Linguistics, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany

*Corresponding author

Abstract

Understanding language in real-world interaction requires methods that integrate auditory, visual, and neural signals. We present a proof-of-concept study combining mobile electroencephalography (EEG) and eye tracking to investigate referential selection during dyadic communication. Using a simplified version of the director task, pairs of participants engaged in naturalistic object-movement instructions while EEG and gaze data were recorded and synchronized via LabStreamingLayer. This approach allowed us to calculate overlap-corrected, regression-based event-related potentials (rERPs) time-locked to spoken nouns and to participants' fixations, focusing on the N400/P300 (300–500 ms) and later (500–800 ms) time windows. Our results show that gaze behavior and fixation-related potentials provide crucial information for interpreting language-related ERPs: targets competing with occluded referents elicited stronger P300 effects, suggesting higher attentional demands, while fixation timing systematically modulated neural responses. Contrary to predictions, competitors received little visual attention, indicating that participants prioritized direct task-relevant information. These findings highlight the potential of multimodal data integration for understanding attentional and predictive mechanisms in real-world communication. Importantly, we demonstrate that established analytic techniques, such as artifact subspace reconstruction, independent component analysis, and linear deconvolution for ERP calculation, are applicable to noisy, naturalistic data sets. This study thus provides a methodological framework for linking gaze and neural activity during interactive language processing beyond laboratory constraints. All preprocessing and analysis scripts, together with example data sets, are openly available on OSF (https://osf.io/5ds4z; a modified and optimized Python implementation is also available on github: https://github.com/XlinCLab/DGAME).

1 Introduction: Multimodal Language

Language in communication is a complex system of multimodal signals, in which the entirety of a person's body – from speech and breathing to facial expressions, gestures, and other bodily movements – can be a potential source of information (Goodwin, 2003; Mondada, 2016). In recent years, there has been a notable trend in linguistics to acknowledge this complexity by moving towards a multimodal view on human language (Benetti et al., 2023; Hasson et al., 2018; Holler & Levinson, 2019; Levinson & Holler, 2014; Mazzini et al., 2023; Mondada, 2016; Perniss, 2018; Vigliocco et al., 2014). From this perspective, language in face-to-face interaction is anchored in the immediate environment in which it occurs by referring to persons or objects available in a room with which people (want to) interact. Beside speech, listeners and speakers make use of information from the visual domain, for instance to establish reference (e.g., Debreslioska et al., 2013; Graham & Argyle, 1975; Hanna & Brennan, 2007; Kita & Özyürek, 2003; So et al., 2009; Somashekarappa et al., 2020).

Studies of language in real-world interaction have in the majority targeted the time-frequency domain, as for instance studies on entrainment, cross-frequency coupling or inter-brain synchrony (Dikker et al., 2021; Drijvers & Holler, 2022; Pérez et al., 2017). Linguistic event-related potential studies on language in interaction have so far been limited to virtual reality (Huizeling et al., 2023; Tromp et al., 2018; Zappa et al., 2019), probably due to the wide range of difficulties connected with real-world EEG studies, from technical setup, experimental control and bad data quality, to labor-intensive annotation and complex analyses. For instance, it can be very difficult to identify enough occurrences of comparable linguistic events across participants, if they are allowed to speak freely, making an event-related potential study a challenging endeavor.

Still, a growing body of evidence from controlled studies suggests that linguistic (i.e. speech) information and information from the visual channel (e.g. gestures) are simultaneously and immediately processed and integrated by the human brain (e.g. Biau et al., 2016; Hubbard et al., 2009; Özyürek, 2021; Peeters et al., 2017; van Wassenhove et al., 2005; Willems et al., 2008; Wu & Coulson, 2005; Zhang et al., 2021). Using functional magnetic resonance imaging (fMRI), Hubbard et al., (2009) for instance showed that gestures modulate activity in the auditory cortex related to the processing of auditory language, providing evidence for the tight coupling of speech and gesture in the brain. ERP studies have repeatedly shown that it is especially the N400 and P300 which are sensitive to multimodal integration (e.g. Özyürek et al., 2007; Peeters et al., 2015; Willems et al., 2008; Zhang et al., 2021). In two EEG studies, Zhang et al. (2021), for example, were able to show that multimodal cues modulate the N400 event-related potential (ERP) component, usually observable when the meaning of a word does not fit its context in the widest sense (see Kutas & Federmeier, 2011). Zhang and colleagues found interactions in the N400 time window (300-500 ms after word onset) between prosodic accentuation and several types of gestural and facial cues. That is, the probabilistic and predictive integration of incoming linguistic meaning (i.e. word meaning), as reflected in the N400 component (Bornkessel-Schlesewsky & Schlesewsky, 2019), extends beyond the auditory speech signal to include visual cues that are weighted differently depending on their contexts of use (Zhang et al., 2021). Peeters et al. (2015) further demonstrated that the integration of speech and pointing gestures affects the P300 time range, indicating the importance of domain general, attention related mechanisms involved in the integration of visual and auditory cues during language comprehension. There are many more examples in the literature that highlight the crucial role of the N400/P300 time range in the processing and integration of information across different information sources, for instance in

natural reading (e.g. Antúnez et al., 2022; Dimigen et al., 2012; Li et al., 2024; Niefind & Dimigen, 2016), or the the integration of written and spoken language with visual input (e.g. Hirschfeld et al., 2011; Knoeferle et al., 2011; Liu et al., 2011; Sitnikova et al., 2008; Staudte et al., 2021).

In the present manuscript, we demonstrate that it is possible to record and analyze multimodal EEG data (visual, auditory) acquired during real-world interaction and offer an approach of how to integrate gaze and EEG data during data analysis and interpretation (all scripts used in this study are available online: https://osf.io/5ds4z/; also see the optimized Python implementation with much shorter running time on github: https://github.com/XlinCLab/DGAME). Using time-resolved regression (Dimigen & Ehinger, 2021; Smith & Kutas, 2015a, 2015b), we calculated language- and fixation-related event-related potentials from 128-channel EEG data recorded with mobile recording devices from freely-moving participants. Focusing on a traditional time window in neurolinguistic research, the N400/P300 time-window (300-500 ms after word onset), and a second, later time window (500-800 ms; see section 2 for motivation), we show how gaze data and fixation-related potentials can be informative about language-related potentials, and how the interpretation of the latter can benefit from the analysis of the former. For this purpose, we report behavioral (eye gaze, fixations) and event-related potential (ERP) results from a real-world study using a simplified version of the *director task* (Keysar et al., 2000) as a proof-of-concept. Before we elaborate on the present study, we want to discuss three main difficulties related with recording and analyzing multimodal data from real-world environments, and how they can be overcome to gain a benefit from multimodal data recordings.

1.1 Multimodal Data Recording in Real World Environments: Challenges

Recording data from multiple modalities, such as EEG, eye tracking, and audio, while offering insights into the interaction of processes from different modalities (Hasson et al., 2018; Holler & Levinson, 2019; Perniss, 2018), introduces significant challenges in data acquisition, synchronization and analysis (Gregori et al., 2023). Here, we want to briefly discuss three of these challenges: data synchronization, noisy EEG data and overlapping event responses.

First, each modality is recorded with its own temporal resolution. For instance, EEG systems often sample at high rates (e.g., 500–1000 Hz) to capture rapid fluctuations in neural activity, eye trackers usually operate from 120 Hz to over 1000 Hz depending on the device, while audio recordings usually use high sampling rates (44.1 kHz or more) to capture the fine structure of acoustic signals. The mismatch in these rates can lead to difficulties in aligning data streams for meaningful analysis. Second, recording EEG data from freely moving participants introduces strong non-neural artifacts related to participants' movements (muscle artifacts and moving recording devices; cf. Gorjan et al., 2022). And, third, freely acting participants perform many tasks beside the processing of experimental stimuli, i.e. they process visual and auditory information, they plan and produce linguistic utterances, fixations, saccades and other movements. This creates a heavy overlap of the event-related responses of all these different events (cf. Dimigen & Ehinger, 2021; Ehinger & Dimigen, 2019; Smith & Kutas, 2015a, 2015b). While in traditional laboratory experiments, this can be effectively controlled, in real-world experiments, it has to be accounted for during analyses. In the following, we discuss our approach to address these three issues.

1.1.1 Data Synchronization: LabStreamingLayer

LabStreamingLayer (LSL, Kothe et al., 2024) is a software framework and protocol designed for real-time data streaming and synchronization in research and experimental environments. LSL facilitates the seamless communication and integration of various data sources and devices, over LAN or WiFi. The LSL framework includes libraries and tools for different programming languages (e.g., C++, Python, MATLAB), allowing researchers to create applications that generate, transmit, or receive data streams. The protocol ensures that data streams are synchronized, which is crucial for accurate analysis and interpretation.

LSL assigns each data sample generated by a source (e.g., a sensor or device) a unique timestamp that reflects the moment of its creation. This timestamp is recorded in the metadata accompanying the data sample. LSL employs a clock synchronization protocol to ensure that the timestamps of data samples are aligned across different devices and sources. This involves maintaining a shared understanding of time across all involved components, i.e. in a network of data sources, one clock is designated as the "master" clock (in our case the recording computer). This clock serves as the reference for synchronization. Other clocks (in our case the clocks of the recording devices, i.e. audio, video, EEG, ET), adjust their timing to match the master clock by periodically requesting synchronization updates from the master clock. This provides the necessary adjustments to bring the clocks into alignment. These adjustments are then applied to the timestamps of data samples generated by recording devices, allowing the precise analysis of simultaneously recorded data streams in time.

1.1.2 Noisy EEG data: ASR and ICA

In naturalistic research settings, EEG recordings are subject to a variety of artifacts from muscle activity, movement, and environmental noise, making it challenging to extract clean

neural signals (Gorjan et al., 2022). Our approach uses a combination of Independent Component Analysis (ICA) and Artifact Subspace Reconstruction (ASR) (Delorme & Makeig, 2004; Kothe & Jung, 2016; Mullen et al., 2015; Kothe et al., 2019). ICA decomposes the complex EEG signal into a set of statistically independent components, allowing the isolation and removal of those components that represent stationary, non-neural artifacts (Delorme & Makeig, 2004). ASR, by contrast, is well-suited to detect and correct for transient, high-amplitude disturbances in the data. It works by reconstructing the underlying brain activity from a subspace defined by the cleaner segments of the recording (see Kothe & Jung, 2016). Together, ICA and ASR form a powerful, complementary framework that can effectively clean noisy EEG data, even when collected in uncontrolled, real-world environments (see Gramann, 2024 for a very recent review on mobile EEG data).

1.1.3 Event overlap

Accounting for the overlap of event-related responses in EEG data (and other time series data) can be achieved using (non-)linear deconvolution (e.g. Dimigen & Ehinger, 2021; Smith & Kutas, 2015a, 2015b). In Smith and Kutas (2015ab), the deconvolution technique within the rERP framework is introduced as a solution to the problem of overlapping event-related potentials (ERPs) in continuous EEG recordings. Traditional ERP methods average epochs time-locked to events, assuming either that responses do not overlap, or they control for the overlap via stimulus or experiment design. Yet, in real-world experimental paradigms, events occur in rapid succession, and their corresponding neural responses mix together, very likely distorting the observed ERP waveform of interest. The deconvolution approach treats the continuous EEG signal as a (linear) combination of overlapping responses. By formulating a general linear model (GLM), each event is represented by a predictor that spans a defined time window, effectively “tagging” the EEG data with the timing and expected

shape of the ERP. The design matrix constructed from these predictors allows for the simultaneous estimation of regression coefficients for each event type. These coefficients can then be directly interpreted or be used to reconstruct the original, overlap-corrected ERPs.

The regression-based deconvolution method (rERP) thus disentangles the overlapping neural responses, enabling more accurate estimation of the true underlying ERP waveform for an event of interest. It also permits the inclusion of additional covariates and interactions, enhancing the capacity to model complex, real-world data where events are not neatly separated into categories. The technique has been implemented in several statistical software environments (e.g. Matlab and Julia: Ehinger & Dimigen, 2019; Python: Sassenhagen, 2019), providing researchers with tools to analyze continuous EEG data under more naturalistic conditions where event overlap is unavoidable.

1.1.4 Benefits: Gaze and fixation-related potentials as cues to language processing

The present study should be understood as a proof-of-concept. We want to demonstrate that considering the aspects described above allows us to tap into the neural integration of spoken language with visual information during real-world interaction. For instance, when participants listen to an utterance, such as “Move the candle to the left, please” while viewing a real-world visual scene, eye-tracking data can inform us whether their gaze is directed toward the target object (candle) or not at the moment the word ‘candle’ is uttered. This allows us to investigate how visual attention or, more generally, visual information modulates neural responses to language in real time. Furthermore, we can relate fixation-related potentials to language-related potentials within and across experimental variables. For instance, fixation-related potentials show a reliable positive component for gazes at target objects, as compared to distractor objects (e.g. Brouwer et al., 2013; Kamienkowski et al.,

2012). In case of referential ambiguities, as in the present study (see section 2), fixation-related potentials can thus be informative cues as to whether participants treated an object as a target or 'just' as a referential, but task-irrelevant competitor. Moreover, studies on natural reading demonstrated that semantically meaningful parafoveal preview can reduce the N400 component when fixating a word during reading (Antúnez et al., 2022; Degno & Liversedge, 2020; Dimigen et al., 2012; Li et al., 2024; Niefind & Dimigen, 2016), revealing that tracking the eyes can also provide important information about the availability of predictive cues during language comprehension. Likewise, studies that use a combination of visual stimuli and spoken or written language, such as the visual world paradigm or sentence-picture verification tasks (Hirschfeld et al., 2011; Knoeferle et al., 2011; Liu et al., 2011; Sitnikova et al., 2008; Staudte et al., 2021), suggest that the degree of similarity between visual and linguistic input is strongly reflected in the N400 time range. Interestingly, not only visual cues influence linguistic processing: research has also demonstrated that linguistic cues influence visual processing (Hirschfeld et al, 2011; Henderson & Hayes, 2017). In the present study, in which participants engage with complex visual scenes while processing spoken language, combined EEG and eye tracking recordings can thus inform us on how the brain's electrophysiological responses in the N400 time range are shaped by the (un-)availability of visual and auditory linguistic cues that can be used for predicting upcoming sensory input. In the following section, we will describe the study design that we used for our proof-of-concept experiment.

2 The present study

2.1 The Task

In the present study, we designed a simplified real-world version of the *director task* (Keysar et al., 2000) to demonstrate the validity of our approach. In this task, two participants are

seated on opposite sides of a vertical array (shelf) in which objects are placed, for instance two boxes and two candles. One participant (the director) gives instructions to the other participant (the matcher) to move objects in the shelf. However, some of the objects are hidden from the view of the director. For instance, when one of the two candles is hidden from the director's view (i.e., it is in the participant's privileged ground, PG) and the director asks the participant to move the candle, a referential conflict arises: from the perspective of the director, the candle is uniquely identifiable, but from the perspective of the matcher, the referential expression is ambiguous, unless the matcher considers the perspective of the director.

We chose this task for several reasons. First, it is well suited to examine how gaze behavior and language-related event-related potentials are integrated during referent selection in a naturalistic experimental setting, since each instruction must be followed by the visual identification of the target object, before it can be moved. Second, the director task can be designed to be repetitive enough to be used in an event-related potential study, which usually requires a large number of comparable trials to produce reliable results. Finally, there is a rich literature on gaze behavior in the director's task (e.g. Barr, 2008; Epley, Keysar, et al., 2004; Epley, Morewedge, et al., 2004; Ferguson & Breheny, 2012; Hanna et al., 2003; Heller et al., 2008; J. J. Wang et al., 2019), and there also exist two event-related potential studies (Richter et al., 2020; Sikos et al., 2019) that used computerized versions of the director task to which we can compare our results.

In the years following the first eye tracking study of the director task (Keysar et al., 2000) the literature was dominated by a debate about the time point in processing at which participants consider the perspective of others, early (Heller et al., 2008; Hanna et al., 2003; Ferguson & Breheny, 2012; Epley et al., 2004) or late (Keysar et al., 2000; Barr, 2008; Wang et al., 2019;

Barr, 2016). Today, it is general consensus that whether and at which time in processing participants consider the perspective of others depends on a variety of factors, including (linguistic) context (e.g., Heller et al., 2008; Brown-Schmidt et al., 2008), task (Ferguson et al., 2015; Ryskin et al., 2020), culture (Wang et al., 2019; Wu et al., 2013; Wu & Keysar, 2007) and the availability of cognitive resources (Cane et al., 2017; Epley et al., 2004). Hanna et al. (2003), for instance, demonstrated in two experiments that addressees are unable to ignore visually salient objects in privileged ground if they match a speaker's description and assign reference as early as possible once a potential referent is identifiable. Recently, a probabilistic model of reference resolution has been developed assuming that the knowledge of the self and the (assumed) knowledge of the other in dyadic interaction are assigned a weight and are then combined into a probabilistic communicative model by both speaker and addressee (Heller & Brown-Schmidt, 2023; Mozuraitis et al., 2016, 2018; Ryskin et al., 2020). This approach highlights that the other's knowledge can only be *inferred* based on weighted cues from different available sources. In the director task, one of the rather salient and informative cues for the director's knowledge and intended reference is visual: whether a potential referent of a referential expression is occluded or not. Gazes at privileged ground competitors thus occur because they are informative when it comes to inferring the knowledge of the director – participants are collecting evidence for or against a referential choice.

To the best of our knowledge, there exist only two event-related potential studies using versions of the director task. The first study by Sikos et al. (2019) utilized temporary referential ambiguities and the Nref ERP component, a sustained negativity interpreted as a marker of referential ambiguity (Van Berkum et al., 2007). On a computer screen, participants saw displays of cards showing animals with accessories, e.g. a brontosaurus with

boots and a brontosaurus with a purse, such that when participants were asked to “move the brontosaurus with boots”, “brontosaurus” was referentially ambiguous until the word “boots”. They found a graded Nref effect that has the highest amplitude when the referential competitor is in common ground, followed by trials with no competitor and trials with a competitor in privileged ground. They argue that the reduction of the Nref in trials with privileged ground competitors shows that participants do not consider privileged ground competitors as potential referents in a way as they do for common ground competitors. In other words, they consider the perspective of the other while determining referential candidates. In the ERP study by Richter et al. (2020), participants were presented with 4x4 grids of objects on a computer screen. Referential ambiguities were created using size contrasted object triplets (e.g. small, medium, large candle) of which either the large or small object was in privileged ground (hidden from the director). When the director now asks to “move the small candle” while the small candle is in privileged ground of the addressee, a referential ambiguity arises: the small candle in privileged ground and the medium candle in common ground are potential referents of the utterance from the speaker’s perspective. In two separate studies, Richter et al. (2020) recorded eye tracking and EEG data using the same experiment, creating comparable data sets of two modalities. While the behavioral results and the results of the eye tracking study replicate previous results (longer reaction times and an increased number of gazes at privileged ground competitors in *pair* trials), the ERP results are different from Sikos et al. (2019). Across conditions, Richter et al. (2020) found a broadly distributed, positive ERP component with posterior maximum, that was more pronounced (more positive) for trials with privileged ground competitor as compared to trials without, which is the opposite of the negative effect in Sikos et al. (2019). Richter et al. (2020)

interpret this late positive effect as reflecting difficulties associated with the integration of privileged ground information.

In the present study, we used a version of the director task modified in way to make it very simple for participants and to reduce the amount of variance in our data. We used a 4x4 shelf on a table as visual display and placed six objects in it (Figure 1). While there were two singleton, i.e. unique objects (e.g. one cream jar, one tube), four objects were pairs of identical objects (e.g. two candles, two roses), with one of each pair hidden from the director's view. We chose this very simple design, because the matcher would only need one piece of information to resolve the referential ambiguity: Is an object occluded from the director's view, or not? By this, we wanted to minimize eye movements related to the comparison of objects for decision making: In Richter et al.'s (2020) study, for instance, participants had to compare sizes of objects to resolve the referential ambiguity, in Sikos et al.'s (2019) study, participants had to compare accessories (e.g. a Brontosaurus with boots and one with a purse), thus possibly fostering gazes at the competitor object, in order to collect evidence for or against a referential choice. Here, no comparison was necessary. Participants could either gaze at the competitor, see that it is hidden and then switch to the target, or they could gaze at the target, see that it is not hidden, and move it. Effectively, this makes the target the most informative object in the present study, because it provides the necessary visual information (hidden or not) *and* because it is the actual target object. Accordingly, we expected participants to pay little attention to the hidden competitors, signaling no difficulties in taking the perspective of the other into consideration.

Since we recorded EEG and gaze data of the matcher, we are able to track where participants' gaze was directed, while they heard a target word and by considering it as a predictor in our rERP calculation. Likewise, we could also include the progress of linguistic information in

the calculation of the fixation-related potentials, in order to examine differences between trials, in which participants were hearing or had already heard the target word while fixating the target object, and such trials, in which they were not. Here, we focus on the N400/P300 time range (300-500 ms) and included an additional later time window (500-800 ms) to account for later effects that we expected based on the very long effect latencies in Richter et al. (2020) and Sikos et al. (2019).

We assume that if participants collected sufficient visual information about the target object (e.g. its position in the shelf) before they have linguistic evidence for the identity of the upcoming target (target fixations before noun), this possibly facilitates target selection during and after linguistic input (e.g. noun onset) in a top-down manner. We assume that this (predictive) information is reflected in a reduction in N400 amplitude and/ or an increase in P300 amplitude for the event-related potentials time locked to noun onset. In trials in which participants made the majority of target gazes during the noun, visual target information and linguistic target information processing occur simultaneously, with visual and auditory linguistic input providing bottom up information at the same time. The relative absence of *a priori* visual information about the target object (e.g. its position) probably leads to relative difficulties in target detection, as compared to the cases, in which participants already have information about the target object before noun onset. This should be reflected in a more negative going ERP between 300-500 ms, e.g. an increase of the N400, and a decrease of the P300 in the ERP time locked to noun onset. Finally, in trials in which target gazes occur later than the offset of the critical noun, participants already have complete linguistic evidence regarding the identity of the target stimulus, before they start actively searching for the object in the shelf. That is, it should mirror the cases, where visual information is available before linguistic information. Again, we expect this to be reflected in a further decrease in P300

amplitude (difficulties in target selection and task execution) and an increase of the N400 amplitude (integration difficulties since no target has been visually identified). For the fixation-related potentials, we expect the effects to be similar, but reversed: fixations to the target object before noun onset occur without the presence of linguistic information, thus they are uninformed fixations with regard to the task-related status of an object. That is, it is still unclear if it will become the target or not, while fixations to the target object after the noun has been fully uttered can be considered informed fixations, with the task-related knowledge about the objects already provided. Thus, N400 and P300 amplitude should be reversed as compared to language-relate event-related potentials. Fixation-related potentials of fixations that occur during the noun, on the other hand, should behave similar to language-related ERPs, since in these cases, the relative information status of both linguistic and visual input is similar in both cases.

3. Experiment

3.1 Materials and Methods

3.1.1 Participants

Thirty-three native speakers of German (17 female, mean age: 25.48, range: 20-32) participated in the current study after written informed consent. The EEG recordings of 7 participants had to be excluded from analysis due to technical problems with the recording equipment or human error (7 participants; 4 female). Consequently, we also excluded these participants from the analysis of the eye tracking data (final sample: 26 subjects, 14 female; mean age: 24.94, range: 20-32). Participants either received course credits or monetary compensation for their participation.

3.1.2 Experimental Stimuli and Procedure

In the experiment, a participant and a confederate were seated face-to-face on opposite sides of a table. The participant's (the matcher) task was to move objects around a 4x4 shelf placed in the middle of the table according to the instructions of the confederate (the director). Four of the sixteen compartments of the shelf were occluded from the view of the confederate. In the shelf, six objects were placed, two were singleton items (*singleton)*, the other four were two pairs of identical objects (*pair)*. In two of the hidden compartments, one object from each pair of identical objects was placed. The other two hidden compartments remained empty. In German, gender is marked on the determiner preceding a noun and can be used as a cue for referential predictions, when a set of potential referents with different grammatical gender exists. In order to avoid these predictions, all objects used in the current experiment had two-syllable, feminine names (*singleton objects*: *die Pflanze*, 'plant'; *die Spritze*, 'syringe'; *die Flasche*, 'bottle'; *die Dose*, 'can/tin'; *object pairs*: *die Kerze*, 'candle'; *die Blume*, 'flower'; *die Vase*, 'vase'; *die Tasse*, 'cup'). The objects were grouped into two sets of two critical pairs and two singleton objects. The grouping of objects into sets was counterbalanced across participants. In the experiment, all participants saw all objects across four experimental blocks. In Block 1 and Block 2, the objects from set A were used, in Block 3 and Block 4 the objects from set B. In the experimental pauses after Block 1 and Block 3, the objects in the shelf and the hidden compartments were randomly scrambled by our lab assistants. The directors were instructed to move every object in the shelf ten times within one experimental block, to write down the respective target position in the shelf, not to move the same object twice in a row, and to use 'left' and 'right' from the matcher's perspective. Matchers were instructed to move objects only one compartment at a time into any direction and to remember that they can see more than the director.

The experimental procedure was very simple. After setup or redecoration, the lab assistants left the room and the director started with the instructions. After each object was moved ten times, a block ended. The experiment lasted between 35-60 minutes. Including setup, an experimental session lasted between 2-3 hours.

3.1.3 Data Recording

All data were recorded and synchronized with the other data streams via labstreaminglayer (LSL; https://github.com/sccn/labstreaminglayer) using the LabRecorder app for LSL (https://github.com/labstreaminglayer/App-LabRecorder).

Audio Data

We recorded the voice of the confederate using a headset microphone (AKG C 520). The recording was digitized with a sampling rate of 22050 Hz (Mono) using an external audio interface (Focusrite Scarlett 2i2 3rd Gen).

Eye tracking data

We recorded participants' gaze using pupil core eye tracking glasses (Pupil Labs) with a monocular setup. The pupil core headset records pupil data at 120 Hz nominal sampling rate and a second ("world") camera records the scene from a first person perspective with a sampling rate of 30 Hz. We used single marker calibration as implemented in the recording software PupilCapture. Angular accuracy and precision were kept below 1.5° and 0.15°, respectively. The data were recorded to hard drive and synchronized with the other data streams via LSL. For tracking the positions of the objects in the shelf throughout the recording, we attached four AprilTags (Krogius et al., 2019; Olson, 2011; Wang & Olson, 2016), with a working principle similar to QR-codes, to the shelf, one in each corner. Using

the surface tracking plugin provided by Pupil Labs, we defined surfaces (areas of interest) offline (here the 16 compartments of the 4x4 shelf). An algorithm then calculates the distances between the tags and the surfaces for each video frame, making it possible to correct for changes in viewpoint (e.g., by head or body movements) and to identify the surfaces' location throughout a recording. Thus, we are able to track at which of the 16 surfaces a participant is gazing at a given point in time.

EEG Data

EEG Data were recorded using a CGX-mobile-128 System (Cognionics, San Diego, CA, USA) with 128 AgCl active electrodes relative to a left mastoid reference at a sampling rate of 500 Hz. The right mastoid served as ground. Four of the 128 channels were never recorded ('FTT10h', 'FTT9h', 'FFT10h', 'FFT9h'), because they had to be removed in order to fit the eye tracking glasses. Channel impedance was kept below 20 kΩ at the beginning of the experiment.

The EEG was digitized and synchronized with the other streams via LSL using the LSL-relay implemented in the CGX-Recording software. Each experimental block was recorded to a separate file (2 patterns and 2 object sets = 4 files), since in the time between blocks, participants were moving freely and did not perform any task-related actions. In addition, we thought of this as a security measure, so that in case of data loss (e.g. due to connectivity issues), we would not lose the data of an entire experimental session.

3.1.4 Data Preprocessing

Audio Data and Annotations

The recorded audio data was imported into Matlab using the xdf import functions, where it was normalized and exported to .wav. The resulting audio files were then subjected to automatic speech recognition and automatic segmentation (Schiel, 1999, 2015) via the BAS web interface (Kisler et al., 2017). Afterwards, we manually controlled the output and applied corrections where necessary. Manual corrections mostly involved misplaced fricative onsets and unclear pronunciations. We then manually annotated the audio recordings for experimental conditions. During the experiment, we also asked the confederates to write down the current position of the objects in the shelf. This information was then manually transferred into lists and combined with the annotations using custom scripts. This enabled us to automatically match the gaze data with the current object positions.

Although we did not systematically analyze the linguistic constructions used by the director, we would like to briefly note that instructions for all participants always included a full noun phrase consisting of a determiner (*die,* 'the.FEM') and a noun (e.g. *Und dann die Tasse nach oben, bitte!*, 'And then the cup up, please.').

Eye Tracking Data

The preprocessing of the eye tracking data was carried out in R using the library eyetrackingR (Forbes et al., 2025). Any data with a pupil-detection confidence <.60 were excluded from analysis and non-AOI gazes were not treated as missing. As there is no software or package available that can deal with the kind of data we recorded (e.g., our data is not trial based but continuous gaze data), large parts of the scripts we used for preprocessing

are custom scripts needed to prepare the data for analysis (e.g., combine the gaze data with the linguistic annotations, create trials).

EEG Data

The EEG recordings of each block were imported into EEGLAB (Delorme & Makeig, 2004) using the import functions for .xdf files of the MobiLab plugin (Ojeda et al., 2014). Before any preprocessing steps, we created an event structure for each block containing all words spoken by the director and all fixations of a participant. The data was resampled to 250 Hz. Afterwards. the data were concatenated and preprocessed as follows. We first filtered the data with a 2Hz FIR filter (pop_firfiltnew, cutoff frequency (-6dB): 1 Hz) to approach stationarity for later ICA decomposition. Afterwards, we rejected noisy channels in a semi-automatic fashion. First, we rejected channels selected by visual inspection of the raw EEG data. Then we used the pop_clean_rawdata() function (Kothe et al., 2019) to automatically reject noisy channels, followed by automatic rejection using the kurtosis method of the pop_rejectchan() function with a threshold of 2. Afterwards, we applied a 100 Hz low-pass filter to the data and then used the function cleanLineNoise to remove line noise artifacts at 50 Hz. Then, we used artifact subspace reconstruction (ASR; Mullen et al., 2015; Kothe & Jung, 2016) as implemented in the clean_rawdata plugin for EEGLAB (Kothe et al., 2019) to identify and correct noisy segments in the recording.

After ASR, we interpolated missing channels (mean = 26, range: 17-46) and re-referenced the data to average reference under reconstruction of the original reference. After re-referencing, the original reference and interpolated channels were removed again in order to avoid rank deficiency. Then, we used the AMICA algorithm to decompose the data into independent components (1 model, 2000 iterations, 10 times rejection of unlikely data). The resulting

weights and sphere of the model were then transferred to a 0.6 Hz low-pass and 20 Hz high-pass filtered version (pop_firfiltnew, cutoff frequency (-6dB): 0.3 Hz, 44.3 Hz) of the data with 250 Hz sampling rate. We then used the ICLabel plugin for EEGLAB to classify components and subsequently rejected all components not labeled “brain” (mean: 43, range: 14-62).

3.1.5 Data analysis and Results

Gaze data

We examined three measures of the gaze data: to investigate the time-course of target selection, (i) proportion of gazes at the target object and the competing object in a pair (the former being in common ground, the latter in privileged ground); (ii) gaze proportions at competitor objects relative to all gazes at other non-target objects; and (iii) the mean time point at which participants gazed at the target object for each trial, which we used as a predictor for the ERP analysis. This measure summarizes all gazes to the target, collapsing the first gaze at the target, the last gaze at the target and gazes in between into a single value. While this has the disadvantage of being rather imprecise, as for instance compared to the first gaze at a target object, it provides a measure for the processing load associated with target identification throughout an entire trial and beyond the single word level. This measure is thus comparable to the *total reading time* in eye tracking research on reading (e.g. Kliegl et al., 2004, 2006; Kuperman et al., 2018).

We contrasted the target of a pair (which conflicts with an object in privileged ground; hence “*pair condition*”) with the target of a singleton (“*singleton condition*”). We used a cluster-based permutation analysis ((Maris & Oostenveld, 2007)) to test for significant differences in gaze proportions between the *pair* and the *singleton* condition for the target AOI (cf. Richter

et al., 2020). We calculated a paired, two-sided t-test for 100 ms time bins from 0 to 3500 milliseconds after critical noun onset. For all time bins with a significance threshold of 0.05 the procedure was then repeated 4000 times with randomly shuffled (not sequentially ordered), i.e. permuted versions of the binned data. This produces a distribution that is expected at chance level. These results are then compared to the original model using the sequentially ordered time series. Through this procedure, we can obtain a p-value that tells us how likely we would find a given effect when the distribution was random, i.e., if there was no effect of our experimental variable.

Since in the *singleton* condition, the privileged ground competitor was absent, we could not compare gazes at the privileged ground competitors between conditions in the same way. Instead, we compared the gaze proportions at the competitor object in the *pair* condition with the gaze proportions averaged over all other, non-target objects: the two singleton objects and the two objects from the other pair (one in privileged ground, one in common ground).

Results

[Figure 2 here]

Figure 2 shows the gaze proportions of the different objects in the shelf by condition. The figure reveals a higher proportion of gazes at the target object in the *pair* (solid) than in the *singleton* (dotted) condition. The cluster based permutation analysis of the target-AOI revealed a significant effect in a cluster between 1100-1900 ms after critical noun onset ($p = 0.009$). In a post-hoc analysis of the goal-AOI, although numerically present, the difference between conditions did not reach statistical significance ($p > 0.1$). Since there were no privileged ground competitors in the *singleton* condition, we compared the gaze proportions of privileged ground competitors in the *pair* condition with the mean gaze proportions of all

other, non-target objects (Figure 3): the two singleton objects and the two objects from the other pair (one in privileged ground, one in common ground). As Figure 3 demonstrates, the gaze proportions associated with the hidden competitor is numerically lower than the mean gaze proportions associated with all other non-target objects in the shelf. According to the results of the permutation analysis, this difference is even statistically significant in clusters between 1400-3400 ms ($p = .001$) after stimulus onset.

[Figure 3 here]

EEG data

First-order statistical analysis: rERP calculation

In order to account for the overlap between brain responses to the stimuli of interest (critical nouns) and other stimuli (determiners, words following or preceding the noun, fixations), we calculated a general linear regression model with a linear time basis function from -500 to 1500 milliseconds after event onset for 4 different event types: determiner preceding the noun, the critical noun, words preceding the determiner (max. 2), intervening between determiner and noun, and following the noun (max. 2), as well as fixations within -1500 to 3500 ms after noun onset. We were interested in the event-related potentials following critical nouns and fixations, the other events were only included to account for the overlap. For the determiners, we included the experimental condition (*pair, singleton*) as a categorical predictor and added *trial* as a covariate without interaction. For critical nouns, we additionally added a continuous predictor *mean target gaze time* (see section *Gaze Data*), for which we assumed an interaction with condition. For all other words before the determiner, or before or after the noun, we only calculated an intercept model to account for the overlap.

The following list contains the formulae used to compute the overlap corrected ERPs for all four event types described above.

1 determiner preceding critical noun: µV~condition+trial

2 critical noun: µV ~ condition*mean_target_gaze_time+trial,

3 preceding, intervening and following words: µV ~ 1

4 fixations within -1.5 to 3.5 seconds around critical nouns: µV ~ condition*fixatedObject*trial_time+spl(saccadicAmplitude)

We used the resulting beta estimates to reconstruct the ERPs of critical nouns and fixations including the conditional means of the covariates (nouns: trial, fixations: saccadic amplitude).

Results: Language-related ERPs

[Figure 4 here]

Figure 4 shows the overlap corrected, reconstructed ERPs by condition and ROI. These can be interpreted just as traditional grand-averages. The most pronounced patterns include a positivity for the *pair* condition (~250-500 ms) spanning left and central prefrontal, frontal and central ROIs, followed by a later positivity for the *pair* condition with posterior maximum peaking at ~500 ms after stimulus onset. The frontal negativity clearly visible for the *pair* condition at ~500 ms over central- and right-prefrontal ROIs that extends to the frontal ROI over right-hemispheric channels is very similar in timing and morphology as compared to the posterior-occipital positivity and probably reflects a common underlying dipole. Moreover, a negativity for the *pair* condition is visible between 0 and ~ 400 ms at left-, right- and central-occipital channels. Over right-hemispheric channels, it is also visible

at the posterior and central ROI. Finally, there is a later positivity at left-frontal and centro-frontal ROIs between ~600-850 ms for the *pair* condition. All other differences between conditions seem rather small and will not be considered.

Figure 5 shows the ERPs to critical nouns in the *pair* (left) and *singleton* (right) condition, grouped by ROI (saggital axis only) and the mean time point relative to noun onset at which participants fixated the target object (grouped for plotting purposes only; *before noun*: -1000ms to 0ms; during noun 0ms to 471 ms; after_noun: > 471 ms). Overall, there are considerable differences between conditions. While the *pair* condition (red, left) shows the strongest effects at occipital channels before mean noun offset (more negative for later mean target fixation time), the *singleton* condition shows more sustained effects from ~125 ms to ~625 ms with reversed polarity (more positive for later mean target fixation times), but similar topographic distribution (occipital ROI). In addition, in the *singleton* condition shows a clear negative effect over central ROIs (more negative with later mean target fixation times) that is basically absent for the *pair* condition. The posterior and occipital positivity and (pre-)frontal negativity found for the *pair* condition at ~500 ms is also affected by the mean target fixation time: with later fixations, the positivity is stronger, but it does not allow a differentiation between fixation times *during* and *after* the critical nouns.

To summarize the results, we found that the pair condition elicited a distinct pattern of ERP effects, including a sustained frontal and posterior positivity peaking around 500 ms, a frontal negativity over right-hemispheric sites, and early occipital negativities, all largely absent or reduced in the singleton condition. Additionally, fixation timing modulated these effects differently across conditions: in the pair condition, later fixations amplified a posterior positivity, whereas in the singleton condition, later fixations led to a more sustained occipital positivity and central negativity.

Results: Fixation-related ERPs

Figure 6 shows the event-related potentials relative to *target* fixation onset by condition. The large positive deflection at 0 ms is probably the lambda response (Ries et al., 2018). It usually occurs 80-120 ms after fixation onset with stationary eye tracking equipment. As of today, we have not identified the source of this latency shift, but we assume that the source is the dispersion-based fixation detection algorithm assuming a dispersion of less than 1.2 degrees visual angle for at least 80 ms, the exact time of the latency shift. In combination with the limited accuracy of the mobile eye tracker (~1.8 degrees) this probably induced this shift. We are currently working on improving our pipeline to also account for head movements, in order to be able to move to a saccade-based detection algorithm as proposed by (Engbert & Kliegl, 2003). Except for a small positivity (~400-650 ms) for the *pair* conditions, there is almost no difference between conditions. However, if we turn to Figure 7, which shows the fixation-related ERPs by fixation time at selected channels, we can see clear differences between the two conditions. First, the effect of fixation time starts very early for the *pair* condition at channel Fpz (0-250 ms), with a more pronounced positivity for earlier fixations (before > during > after noun). It is followed by and partly overlaps with two positive effects (~200-700 ms and ~740-1000 ms) which are strongest over Cz (with the same direction (before > during > after noun). At channel Oz, there is a series of negative effects visible (before > during > after) which is less clear than the effects at channels Fz and Cz. For the *singleton* condition, on the other, the effects are different. At channel Fz there is only a rather late positivity for early fixations starting at ~500-1000 ms after fixation onset (before > during > after). It is preceded by a positivity over channel Cz between ~250 and 500 ms (before > during > after), which is similar to the effect found for the pair condition, but

shorter (~250-500 ms). This positivity is followed by a later positive effect from ~800-1000 ms after fixation onset (after > during > before noun). At channel Oz, an early positivity is visible for the *singleton* condition (~100-375 ms) which is stronger for earlier fixations (before > during > after noun). It is followed by a later negativity (~500-800 ms) with a similar gradient (before > during > after noun).

Overall, the fixation-related potentials differed markedly by condition, with the pair condition showing earlier and more sustained effects of fixation timing across frontal and central sites, while the singleton condition exhibited later and more transient effects with distinct temporal and topographic profiles.

Second-order statistical analysis: permutation analysis

Results: Language-related ERPs

For the second-order statistical analysis of the language-related ERPs, we summarized the ERP data into time bins of 100 ms, in order to account for auto-correlation. We then calculated a linear regression model for each of the resulting time bins using the following formula data~laterality*saggitality*condition*mean_target_fixation*baseline, were baseline (mean predicted microvolt -250 ms to 0 ms relative to noun onset) was added to the model formula to apply robust baseline correction (Alday, 2019), and saggitality and laterality were categorical topographic predictors (laterality: left, central, right; saggitality: prefrontal, frontal, central, posterior, occipital). Afterwards we used a time bin-based permutation approach to test for significant effects. For this purpose, we shuffled the predicted microvolt values within each time bin to decouple the dependent and independent variables, thereby simulating the null hypothesis, under which the predictors of our model and the dependent variable (predicted microvolt) are not related. Then we calculated linear models for 2000

permuted versions of each time bin and compared their absolute t values $|t|$ to the original, unshuffled models' $|t|$ values. From this comparison, we created a probability distribution for each time bin, making it possible to compute how likely $|t|$ is larger under the assumption that the null hypothesis is true, with smaller *p* indicating lower probability that the null hypothesis is true and higher probability that the effect of the unshuffled data is a true effect. *Figure 8* shows a selection of significant effects in the time bins within the N400/P300 (black rectangle) and late positivity (blue rectangle) time range (see supplementary materials at https://osf.io/5ds4z/ for a full graphical and text summary of the permutation results and the results of the linear models per time bin). We see that the effect of *mean target fixation* significantly interacts with condition and with the topographical factors in both time windows, but with a focus on the earlier P300/N400 time window.

[FIGURE 9 AND FIGURE 10 HERE]

The effect plots of the time-range from 0-800 ms following noun onset (Figures 9 and 10) make the time and topographic distribution of the significant effects a bit clearer. It becomes obvious from the effect plots in Figure 9, that the effect of *mean target fixation* for the *pair* condition is strongest within the time-range of the N400/P300 (300-500 ms), visible as a positivity over central occipital channels. It is reversed over frontal and prefrontal channels and followed by a negativitiy strongest over right-central, frontal and prefrontal channels. Although the earlier onset and more sustained nature of the negativity may suggest two different underlying effects, we cannot decide on basis of the present data whether the negativity and positivity are different effects or share an underlying dipole. Therefor, we will not discuss them separately. Note, that the very early effects directly after noun onset are very much attenuated as compared to the ERPs, because the *baseline* predictor canceled them out. For the *singleton* condition (Figure 10), the picture is very similar, although the occipital

positivity starts earlier (200-300 ms) and is less pronounced (hence the main effect of *condition* visible in the ERPs). The same is true for the centro-frontal negativity, which starts very early for the *singleton* condition (0-100 ms) and is rather left-lateralized and stronger over the central ROI, as compared to the *pair* condition. In a similar way as the positivity, it is less pronounced than for the *pair* condition, explaining the main effect of *condition* visible over these ROIs and time-windows in the ERPs.

Results: Fixation-related ERPs

[FIGURE 11 HERE]

The statistical procedure for the fixation-related ERPs was identical to the one described for the nouns above. The model formula included the same topographic factors, a factor for *condition* (pair/singleton), a continuous predictor for the *trial_time* relative to noun onset at which a fixation occurred, and a continuous predictor *baseline* (-250 to 0 ms relative to fixation onset). For the second order analysis, we only focused on the reconstructed rERPs for target objects, dropping all other fixation-related potentials. Figure 11 summarizes the significant effects in the N400/P300 (300-500 ms) and late positivity time window (500-800 ms). Other than for the nouns, we can see that *trial time* significantly interacts with *condition* and the topographical factors in all time bins, without a clear focus on the earlier (N400/P300) or later (late positivity) time window. Figure 12 and Figure 13 summarize the corresponding effects by ROI.

[FIGURE 12 AND FIGURE 13 HERE]

3.1.6 Summary

Gaze Data

The analysis of the gaze proportion revealed a significant increase in gazes at the target object in the presence of a referential competitor in privileged ground (*pair* condition). Competitors, on the other hand, received significantly less gazes than all the other non-target objects in the visual display. The numeric increase in gazes at the goal compartment for the *pair* condition did not reach significance. We assume that the absence of gazes to the competitor objects can be explained via attentional mechanisms: As we discussed in the introduction, the target objects are the most informative object with regard to solving the experimental task. First of all, they fit the meaning of the linguistic utterance (e.g., being a vase), second, they provide all information necessary for resolving the referential ambiguity (hidden or not), and, most importantly, they are the target object that is supposed to be moved. Given that the hidden competitors remained in place for an entire experimental block (10-15 minutes), while the target objects changed position every few trials, participants probably acquired some knowledge about where the target objects are *not located*, thus paying little attention to the hidden objects. However, why the target objects in the *pair* condition received more gazes than the target objects in the *singleton* condition is hard to explain without additional data. Fortunately, the fixation-related potentials can provide clarification.

Fixation-related potentials

In the present study, the amplitude of the fixation-related potentials time-locked to target fixations was not directly influenced by condition in the time windows of interest (300-500, 500-800 ms). Both, targets in the *singleton* and *pair* condition elicited very similar ERPs with

only minor differences. Still, we have reason to assume that participants treated the "*pair* targets" different, as compared to the "*singleton* targets", since differences in the fixation related potentials become apparent, when we consider the time in a trial at which a fixation occurred: when more linguistic evidence was available that the object currently gazed at is the actual target object (fixations during or after the noun), fixations to targets in the *pair* condition elicited an early occipital positivity (200-300 ms), followed by a sustained positivity with a similar gradient response. This early positivity was absent in singleton trials, which only showed a sustained positivity at occipital channels. Both effects, the early positivity for *pair* trials and the sustained positivity for both conditions are mirrored over central, frontal and prefrontal electrodes. We assume, that the amount of available linguistic information is responsible for the graded effects of fixation time. Fixations that occurred during (partial information available) or after the noun (all information available) elicited more positive going fixation-related potentials over posterior and occipital channels, and vice versa over frontal and prefrontal channels, because the visual information provided by the target object matches the linguistic information provided by the word. The fact that the effect is stronger in the *pair*, along with the *pair* specific early positivity, strengthens our assumption that targets in the *pair* attract attention and are, in a certain sense, treated as "targets among targets".

Language-related potentials

The language-related potentials time-locked to noun onset revealed a positivity for the *pair* condition in the posterior and occipital ROIs. In central, frontal and prefrontal ROIs the amplitude of this effect reversed, while its temporal and spatial distribution remained rather constant. The statistical analysis confirmed that this effect is significant in the N400/P300

(300-500 ms) and in the later time window (500-800 ms) on which we focused our analysis. We assume that this positive deflection is a P300 related to attentional reorientation following the detection of a linguistic target. It is reduced for targets in the *singleton* condition, since they do not have referential competitors and are thus easier to identify by virtue of our experimental design. Moreover, target words for which the majority of fixations to the target object occurred before or during hearing a target noun, i.e. trials in which participants either seemed to have made rather accurate predictions of the upcoming object (most gazes before noun) or have collected enough (partial) linguistic information to make a prediction (most gazes during noun), show a decreased positivity in both conditions, although this effect is less strong and starts earlier in the singleton condition. Overall, this likely demonstrates the P300 in this case reflects the degree of attentional reorientation that is necessary to solve the experimental task: to identify the target object. When the target object's location is already known (before noun), the P300 is maximally reduced, whereas in trials in which participants still have to actively search for the object after or during hearing the noun, target detection is more difficult and requires more attentional resources, because participants have to initiate a visual search, hence shift from a perceptive to an action-based attentional state.

4 General Discussion

We presented results from an interactive, real-world study using the director task paradigm that used mobile EEG and eye tracking equipment to simultaneously record eye tracking and EEG data of participants engaged in the director task. With as little instructions for the directors as needed, we tried to create an experimental situation that is as natural as possible given the experimental task and its repetitive nature. We discussed and demonstrated how the difficulties connected with event-related potentials in real-world EEG studies can be

overcome using analysis tools suited for this kind of data, and how multimodal data recordings (EEG, ET) can be reciprocally informative. We showed that the temporal distribution and availability of visual (gaze) and speech information have an effect on both fixation-related *and* language-related event-related potentials. Together with the behavioral gaze data, the multimodal data recording provides us with a relatively complete picture of the state of participants' visual and auditory attention, allowing us to draw conclusions for one modality which are informed by the other. Since in real-world interaction, visual and auditory information are virtually always present simultaneously, our results highlight the importance of analyzing them together when dealing with data acquired from real-world, interactive experiments. In the following, we discuss our results against the literature and provide some suggestions of how to explore real-world communicative interaction further.

4.1 Gaze distributions as cognitive offloading

In our experiment, participants showed an increase in the gaze proportions associated with the target object in trials with a referential competitor in common ground (*pair* condition). This outcome somehow contrasts with our initial prediction that competitors would be attended as informative cues for perspective-taking. Instead, the data suggest that in our simplified design, participants adapted their strategy by prioritizing direct information from targets and disregarding competitors, which provided no additional task-relevant information once their occlusion status was known. The results also differ from the previous ERP studies (Sikos et al., 2019; Richter et al., 2020), who both found a reduction in target gazes for referents with a privileged ground competitor. Still, they are in line with other studies demonstrating that the strength of the interference effect of referential competitors depends on a variety of factors, including (linguistic) context (e.g., Heller et al., 2008; Brown-Schmidt

et al., 2008), task (Ferguson et al., 2015; Ryskin et al., 2020), culture (Wu & Keysar, 2007; Wu et al., 2013; Wang et al., 2019), and the availability of cognitive resources (Epley et al., 2004a; Cane et al., 2017). Clearly, our very simple experimental design and the fact that competitors remained in their compartment for the entire 10-15 minutes of a block made the referential conflict very explicit. Participants likely learned certain aspects of the placement of objects, so that "ignoring" the referential competitors was possible without increasing the effort of solving the task. Moreover, our use of *identical* competitors also likely reduced the number of gazes to competitor objects further, since there was no information gain from gazing at the competitors, as compared to gazing at the target objects. As we discussed above at several points, the target objects in the *pair* condition were the most salient and most informative stimuli in terms of task resolution, since they provide all necessary information for task resolution *and* they are the target object that is supposed to be moved. This makes gazing at the competitor object (of which participants know where it is) uneconomical. This is different from Richter et al. (2020) or Sikos et al. (2019), where participants had to compare sizes or accessories of referential candidates in ever-changing, randomized visual displays to reach a decision. We argue that this is a behavior called cognitive offloading (Risko & Gilbert, 2016): participants gaze at the stimulus display to gather the information that is necessary for task performance, since the alternative, memorizing the array, is cognitively very demanding. While this likely happened in Richter et al. (2020) and Sikos et al. (2019), for the reasons outline above, this behavior was not necessary and even infelicitous in the present study, since learning certain aspects of the visual display was connected with less effort in the present study.

Further evidence for our claim that targets in the *pair* condition require more attentional resources in terms of reorientation comes from the effects in the fixation-related potentials.

Previous research on fixation-related potentials has identified a P300 effect for fixation-related potentials, that distinguishes between fixations on singleton target stimuli within a display of distractors and fixations on the distractors (Brouwer et al., 2013; Kamienkowski et al., 2012). This receives strong evidence also from intracranial recordings, showing that the brain dissociates between target and distractor stimuli at the single-neuron level (Wang et al., 2018). The present fixation-related potentials show very early positive responses following target fixations in the *pair* condition, given that (partial) linguistic information has already been provided. That is, fixations that occurred during or after the target noun has been uttered, elicit a stronger early positivity effect in the *pair* condition, but not in the *singleton* condition. This provides indirect evidence that, compared to targets in the *singleton* condition, participants seemed to attribute greater attentional resources to the targets in the *pair* condition. This finding in the fixation-related potentials helps us in the interpretation of the gaze behavior and the language related event-related potentials, as we can make assumptions about participants' attentional states that are informed by data acquired from the participants themselves.

4.2 The N400-P300 complex and multimodal integration

In the language-related potentials, we found a P300 following nouns, with higher amplitude in the *pair* condition. It is thus in line with the ERPs in Richter et al. (2020) and Sikos et al. (2019), who also found an increased positivity for referents with common ground competitors. However, given that we found the reversed effect in the gaze proportions for targets, as compared to their study, this is surprising. The finding therefore suggests that P300 amplitude is not straightforwardly tied to the *number* of target fixations, but rather reflects the detection of referential conflicts that require attentional reorientation. Given the nature of the

P300 as a decision related brain potential that seems to be at the intersection of cognition and behavior (Aston-Jones & Cohen, 2005; Aston-Jones & Waterhouse, 2016; Chennu et al., 2009; Makeig et al., 2004; Nieuwenhuis et al., 2005; Twomey et al., 2015; Verleger et al., 2005), the reduction in amplitude for the *singleton* targets becomes readily explainable: the target object can be understood similar to the repeated presentation of a stimulus in a controlled experiment (e.g. a standard): the *singleton* targets can be understood as the unsurprising, predicted and easily identifiable "standard" and the *pair* as the surprising, harder-to-detect deviant, requiring attentional reorientation and more effort (i.e. more gazes) to be identified and selected correctly. As a consequence, they elicit the stronger P300 response. This interpretation is further strengthened by the effect of mean target fixation time for both conditions: the later participants fixated the target, i.e. the longer they needed to identify it, the larger the amplitude of the P300 in both conditions. The fact that e.g. Richter et al. (2020) report fewer gazes at the target in their conflict condition is, from our perspective, explainable by cognitive offloading (Risko and Gilbert, 2016): in Richter et al. (2020), participants compare sizes of distractors and targets, in order to arrive at a decision. By contrast, in our design, participants solve the more difficult task by making more gazes to the actual target, since the distractor is uninformative and its position known. In both cases, participants use their gaze to collect the evidence needed to reach a decision, i.e. they employ the same action with different strategies. That is, what seems as a differences between our studies, is actually none.

However, at least since Roehm et al. (2007) we know that the P300 and N400 response overlap during language comprehension. As a consequence, it is difficult to decide, whether a given effect in the P300/N400 time range is a real P300 effect without N400, a real N400 effect without P300, or a mixture of both, an N400 effect *and* a P300 effect. Although we

acknowledge this problem, and although we acknowledge that solving it would be very helpful, yet, given that we cannot entirely solve it as of today, we do not think that this discussion is very fruitful, especially given that what we are gazing at are averaged responses that do not reflect the underlying processes, but rather 'peaky' responses of different latency, which we summed and averaged to a sustained positivity (cf. Aliko et al., 2023 for an interesting analysis and discussion on the topic). However, regardless of this discussion on component overlap, we argue that it is nonetheless interesting that, just as in controlled laboratory studies on multimodal integration, in our interactive, real-world study, we find the N400/P300 time window to be highly sensitive to the integration of visual and auditory cues during language comprehension – for fixation-related *and* language-related potentials. In both cases, we found significant increases in the N400/P300 range (and the following time window) for the time at which fixations to the target object occurred. The fixation-related potentials during natural reading mentioned in the introduction behave in a very similar way: semantically congruent parafoveal preview gradually reduces P300/N400 amplitude of the following fixation on a target word (Antúnez et al., 2022; Dimigen et al., 2012; Li et al., 2024; Niefind & Dimigen, 2016). Thus, our results, together with previous results on multimodal integration, which all report effects of multimodal integration in the N400 time window (e.g. Özyürek et al., 2007; Peeters et al., 2015; Willems et al., 2008; Zhang et al., 2021), strongly underline the adequacy of Kutas and Federmeier's (2011: 22) description of the N400 as "[a] region [that] is more accurately described as reflecting the activity in a multimodal long-term memory system that is induced by a given input stimulus during a delimited time window", and highlights the need for a multimodal view on language and the brain, as put forward by (Hasson et al., 2018). Clearly, and this becomes obvious from our and previous results, participants use multimodal information for reaching referential

decisions, and the integration of this information seems to be reflected in the N400-P300 time window. With the right analytic tools, we think that we can tap into this time window during real-time language comprehension in real-world interaction.

4.4 Conclusion, open questions and outlook

With the current study we demonstrated that the real-time dynamics of language in interaction can be successfully investigated using mobile recording equipment. We demonstrated that even under naturalistic conditions, with freely-moving participants and large amounts of experimental noise, it is possible to calculate event-related potentials from noisy EEG data that are comparable to previous, controlled studies, with effects of multimodal integration taking place in the N400/P300 time window, as has been repeatedly demonstrated in controlled studies. Overall, our results required us to revise several of our original assumptions. For instance, although we predicted that competitors would be attended as informative cues, participants in fact ignored them; although we predicted that pair targets would be easier, the data revealed stronger P300 effects indicating greater attentional demand. These deviations from our predictions are not contradictions but highlight how real-world data can challenge and refine theoretical expectations. We want to stress at this point, that it is not necessary for this purpose and, given the multimodal nature of our stimuli, not really adequate, to offer isolated linguistic explanations for our results. As pointed out by Hasson et al. (2018), we showed that it is possible to explain the present results with reference to domain general processes related to attention and memory (learning of distractor positions in shelf, cognitive offloading, attention) and prediction (pre-noun gaze at target objects, semantic prediction), alone. At least for this kind of linguistic input (very simple, reduced utterances with noun phrases and without verbs), it is not necessary to draw back to

language specific brain functions that should or should not be reflected in the N400 time window.

However, having spent considerable time examining the dataset before writing this manuscript, we want to stress that the number of questions the data raise is far higher than the number of answers they provide. Here is a non-exhaustive list of questions we did not discuss in the present manuscript so far, since admittedly, we do not have satisfying answers to them. First, it would be very interesting to further subdivide the data into groups based on individual gazing strategies. However, a much higher number of participants would be needed for such an analysis. Second, future research should further examine the role of the director and their interaction with the participant, for instance how joint attention (participants following the gaze of the director) is used by participants to make inferences about the upcoming object. Here, we were only able to track the face of the director in a subset of participants and trials, making it difficult to analyze these gazes. However, we observed that on average participants followed a very clear gazing strategy: gaze at the director's face, gaze at the target object, gaze at the goal (Figure 12).

[FIGURE 12 HERE]

Taking this information into account would be very helpful for data interpretation, e.g. how long participants gaze at the director's face before noun onset, or whether participants follow the gaze of the director to the target object, while the director is planning their instructions (cf. Egurtzegi et al., 2022; Sauppe, 2016, 2017), and whether and to which degree they use this visual cue provided by the director. Related to this, third, where do the language-related ERP effects before noun onset come from? How could there be a difference between conditions when the noun phrases used in the instructions are ambiguous until the noun?

There is likely some source of information which we have not considered thus far. The gaze direction of the director would, as mentioned above, be a potential candidate, but intonation patterns might also allow the participants to infer the next object prior to noun onset. Fourth, can we predict gaze behavior based on the language related ERPs using naturalistic data? This list could be continued endlessly and the data open countless opportunities for further investigations.

Overall, the present study demonstrates that it is possible to collect EEG and eye tracking data simultaneously in interactive, real-world settings. Our study shows the potential of such real-world interactions with multimodal data recordings: they enable us to study the coordination of multimodal processes, such as gaze control, language comprehension and production, as well as discourse and social phenomena during natural interaction situations. Future research must not only work out the details of this interplay, but must also establish new standards in the preprocessing, analysis and interpretation of such multimodal data sets by means of innovation, replication and scientific exchange.

Declarations

Acknowledgments

We want to thank Claudia Kilter, Brita Rietdorf, Leonie Latza, Lisa Lubomierski and Robert Voigt for their help in recruiting participants, data collection, data annotation and analysis.

Funding

The research presented in this manuscript has been funded by the *VolkswagenStiftung* (Volkswagen Foundation) in the Momentum project "Communication electrified - towards a natural investigation of real-time language processing“ of the third author.

Conflicts of interest/Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics approval

The study involved human participants. It was reviewed and approved by the Ethikkommission der Deutschen Gesellschaft für Sprachwissenschaft (Ethics Committee of the German Linguistic Society).

Consent to participate

The participants provided their written informed consent to participate in this study.

Consent for publication

The participants provided their written informed consent to their data being used in a publication. All participants were given the possibility to chose what kind of data (EEG, ET, video, audio) they allow to be published for what kind of purpose (e.g. as part of group results in a journal, individual video/audio/EEG/ET in a journal, at a conference, with the scientific community etc.).

Availability of data and materials

Due to data protection rules, the data cannot be made publicly available, but documented code for the analyses in this article is available at OSF (https://osf.io/5ds4z) together with two complete example data sets. Upon personal request, all data sets can be made available to other researchers.

Code availability

All code used for the preprocessing, analysis and plotting of the present data is publicly available at OSF: https://osf.io/5ds4z (there is also a modified and optimized Python implementation available on github: https://github.com/XlinCLab/DGAME).

Authors' contributions Ingmar Brilmayer: design of the experiment, code design and writing code, data collection, data analysis, writing; Philip Georgis: coding, writing; Petra B. Schumacher: writing

References

Alday, P. M. (2019). How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology*, *56*(12), Article 12. https://doi.org/10.1111/psyp.13451

Aliko, S., Wang, B., Small, S. L., & Skipper, J. I. (2023). *The entire brain, more or less, is at work: 'Language regions' are artefacts of averaging* (p. 2023.09.01.555886). bioRxiv. https://doi.org/10.1101/2023.09.01.555886

Antúnez, M., Milligan, S., Hernández-Cabrera, J. A., Barber, H. A., & Schotter, E. R. (2022). Semantic parafoveal processing in natural reading: Insight from fixation-related potentials & eye movements. *Psychophysiology*, *59*(4), e13986. https://doi.org/10.1111/psyp.13986

Aston-Jones, G., & Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus–norepinephrine system in optimal performance. *Journal of Comparative Neurology*, *493*(1), 99–110. https://doi.org/10.1002/cne.20723

Aston-Jones, G., & Waterhouse, B. (2016). Locus coeruleus: From global projection system to adaptive regulation of behavior. *Brain Research*, *1645*, 75–78. https://doi.org/10.1016/j.brainres.2016.03.001

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), Article 1. https://doi.org/10.1016/j.cognition.2008.07.005

Barr, D. J. (2016). Visual world studies of conversational perspective taking: Similar findings, diverging interpretations. In *Visually Situated Language Comprehension* (pp. 261–290). John Benjamins Publishing Company. https://www.degruyterbrill.com/document/doi/10.1075/aicr.93.10bar/html

Benetti, S., Ferrari, A., & Pavani, F. (2023). Multimodal processing in face-to-face interactions: A bridging link between psycholinguistics and sensory neuroscience. *Frontiers in Human Neuroscience*, *17*. https://doi.org/10.3389/fnhum.2023.1108354

Biau, E., Morís Fernández, L., Holle, H., Avila, C., & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio–visual alignment are modulated by the communicative nature of the stimuli. *NeuroImage*, *132*, 129–137. https://doi.org/10.1016/j.neuroimage.2016.02.018

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00298

Brouwer, A.-M., Reuderink, B., Vincent, J., van Gerven, M. A. J., & van Erp, J. B. F. (2013). Distinguishing between target and nontarget fixations in a visual search task using fixation-related potentials. *Journal of Vision*, *13*(3), 17. https://doi.org/10.1167/13.3.17

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*(3), 1122–1134. https://doi.org/10.1016/j.cognition.2007.11.005

Cane, J. E., Ferguson, H. J., & Apperly, I. A. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), Article 4. https://doi.org/10.1037/xlm0000345

Chennu, S., Craston, P., Wyble, B., & Bowman, H. (2009). Attention Increases the Temporal Precision of Conscious Perception: Verifying the Neural-ST2 Model. *PLOS Computational Biology*, *5*(11), e1000576. https://doi.org/10.1371/journal.pcbi.1000576

Debreslioska, S., Özyürek, A., Gullberg, M., & Perniss, P. (2013). Gestural Viewpoint Signals Referent Accessibility. *Discourse Processes*, *50*(7), 431–456. https://doi.org/10.1080/0163853X.2013.824286

Degno, F., & Liversedge, S. P. (2020). Eye Movements and Fixation-Related Potentials in Reading: A Review. *Vision*, *4*(1), Article 1. https://doi.org/10.3390/vision4010011

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dikker, S., Michalareas, G., Oostrik, M., Serafimaki, A., Kahraman, H. M., Struiksma, M. E., & Poeppel, D. (2021). Crowdsourcing neuroscience: Inter-brain coupling during face-to-face interactions outside the laboratory. *NeuroImage*, *227*, 117436. https://doi.org/10.1016/j.neuroimage.2020.117436

Dimigen, O., & Ehinger, B. V. (2021). Regression-based analysis of combined EEG and eye-tracking data: Theory and applications. *Journal of Vision*, *21*(1), 3–3.

Dimigen, O., Kliegl, R., & Sommer, W. (2012). Trans-saccadic parafoveal preview benefits in fluent reading: A study with fixation-related brain potentials. *NeuroImage*, *62*(1), 381–393. https://doi.org/10.1016/j.neuroimage.2012.04.006

Drijvers, L., & Holler, J. (2022). Face-to-face spatial orientation fine-tunes the brain for neurocognitive processing in conversation. *iScience*, *25*(11), Article 11. https://doi.org/10.1016/j.isci.2022.105413

Egurtzegi, A., Blasi, D. E., Bornkessel-Schlesewsky, I., Laka, I., Meyer, M., Bickel, B., & Sauppe, S. (2022). Cross-linguistic differences in case marking shape neural power dynamics and gaze behavior during sentence planning. *Brain and Language*, *230*, 105127. https://doi.org/10.1016/j.bandl.2022.105127

Ehinger, B. V., & Dimigen, O. (2019). Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ*, *7*, e7838. https://doi.org/10.7717/peerj.7838

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective Taking as Egocentric Anchoring and Adjustment. *Journal of Personality and Social Psychology*, *87*(3), Article 3. https://doi.org/10.1037/0022-3514.87.3.327

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*(6), 760–768. https://doi.org/10.1016/j.jesp.2004.02.002

Ferguson, H. J., Apperly, I., Ahmad, J., Bindemann, M., & Cane, J. (2015). Task constraints distinguish perspective inferences from perspective use during discourse interpretation in a false belief task. *Cognition*, *139*, 50–70. https://doi.org/10.1016/j.cognition.2015.02.010

Ferguson, H. J., & Breheny, R. (2012). Listeners' eyes reveal spontaneous sensitivity to others' perspectives. *Journal of Experimental Social Psychology*, *48*(1), Article 1. https://doi.org/10.1016/j.jesp.2011.08.007

Forbes, S., Dink, J., & Ferguson, B. (2025). *eyetrackingR*. http://www.eyetracking-r.com/

Goodwin, C. (2003). Pointing as Situated Practice. In *Pointing*. Psychology Press.

Gorjan, D., Gramann, K., De Pauw, K., & Marusic, U. (2022). Removal of movement-induced EEG artifacts: Current state of the art and guidelines. *Journal of Neural Engineering*, *19*(1), 011004. https://doi.org/10.1088/1741-2552/ac542c

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, *10*(1), 57–67. https://doi.org/10.1080/00207597508247319

Gramann, K. (2024). Mobile EEG for neurourbanism research - What could possibly go wrong? A critical review with guidelines. *Journal of Environmental Psychology*, *96*, 102308. https://doi.org/10.1016/j.jenvp.2024.102308

Gregori, A., Amici, F., Brilmayer, I., Ćwiek, A., Fritzsche, L., Fuchs, S., Henlein, A., Herbort, O., Kügler, F., Lemanski, J., Liebal, K., Lücking, A., Mehler, A., Nguyen, K. T., Pouw, W., Prieto, P., Rohrer, P. L., Sánchez-Ramón, P. G., Schulte-Rüther, M., … Von Eiff, C. I. (2023). A Roadmap for Technological Innovation in Multimodal Communication Research. In V. G. Duffy (Ed.), *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management* (Vol. 14029, pp.

402–438). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35748-0_30

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615. https://doi.org/10.1016/j.jml.2007.01.008

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), Article 1. https://doi.org/10.1016/S0749-596X(03)00022-6

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, *180*, 135–157. https://doi.org/10.1016/j.cognition.2018.06.018

Heller, D., & Brown-Schmidt, S. (2023). The Multiple Perspectives Theory of Mental States in Communication. *Cognitive Science*, *47*(7), Article 7. https://doi.org/10.1111/cogs.13322

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), Article 3. https://doi.org/10.1016/j.cognition.2008.04.008

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743–747. https://doi.org/10.1038/s41562-017-0208-0

Hirschfeld, G., Zwitserlood, P., & Dobel, C. (2011). Effects of language comprehension on visual processing – MEG dissociates early perceptual and late N400 effects. *Brain and Language*, *116*(2), 91–96. https://doi.org/10.1016/j.bandl.2010.07.002

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, *30*(3), 1028–1037. https://doi.org/10.1002/hbm.20565

Huizeling, E., Alday, P. M., Peeters, D., & Hagoort, P. (2023). Combining EEG and 3D-eye-tracking to study the prediction of upcoming speech in naturalistic virtual environments: A proof of principle. *Neuropsychologia*, *191*, 108730. https://doi.org/10.1016/j.neuropsychologia.2023.108730

Kamienkowski, J. E., Ison, M. J., Quiroga, R. Q., & Sigman, M. (2012). Fixation-related potentials in visual search: A combined EEG and eye tracking study. *Journal of Vision*, *12*(7), 4. https://doi.org/10.1167/12.7.4

Keitel, A., Ince, R. A. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *NeuroImage*, *147*, 32–42. https://doi.org/10.1016/j.neuroimage.2016.11.062

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking Perspective in Conversation: The Role of Mutual Knowledge in Comprehension. *Psychological Science*, *11*(1), Article 1. https://doi.org/10.1111/1467-9280.00211

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial

thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kliegl, R., Grabner ,Ellen, Rolfs ,Martin, & and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1–2), 262–284. https://doi.org/10.1080/09541440340000213

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35. https://doi.org/10.1037/0096-3445.135.1.12

Knoeferle, P., Urbach, T. P., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification. *Psychophysiology*, *48*(4), 495–506. https://doi.org/10.1111/j.1469-8986.2010.01080.x

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural Entrainment Determines the Words We Hear. *Current Biology*, *28*(18), Article 18. https://doi.org/10.1016/j.cub.2018.07.023

Kothe, C. A. E., & Jung, T.-P. (2016). *Artifact removal techniques with signal reconstruction* (United States Patent No. US20160113587A1). https://patents.google.com/patent/US20160113587A1/en

Kothe, C., Miyakoshi, M., & Delorme, A. (2019). *Clean_rawdata* (Version 2.7) [Computer software]. https://github.com/sccn/clean_rawdata

Kothe, C., Shirazi, S. Y., Stenner, T., Medine, D., Boulay, C., Grivich, M. I., Mullen, T., Delorme, A., & Makeig, S. (2024). *The Lab Streaming Layer for Synchronized*

Multimodal Recording (p. 2024.02.13.580071). bioRxiv. https://doi.org/10.1101/2024.02.13.580071

Krogius, M., Haggenmiller, A., & Olson, E. (2019). Flexible Layouts for Fiducial Tags. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1898–1903. https://doi.org/10.1109/IROS40897.2019.8967787

Kuperman, V., Matsuki, K., & Van Dyke, J. A. (2018). Contributions of reader- and text-level characteristics to eye-movement patterns during passage reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(11), 1687–1713. https://doi.org/10.1037/xlm0000547

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(Volume 62, 2011), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Levinson, S. C., & Holler, J. (2014). *The origin of human multi-modal communication | Philosophical Transactions of the Royal Society B: Biological Sciences*. https://royalsocietypublishing.org/doi/full/10.1098/rstb.2013.0302

Li, N., Wang, S., Kornrumpf, F., Sommer, W., & Dimigen, O. (2024). Parafoveal and foveal N400 effects in natural reading: A timeline of semantic processing from fixation-related potentials. *Psychophysiology*, *61*(5), e14524. https://doi.org/10.1111/psyp.14524

Liu, B., Wang, Z., & Li, J. (2011). The influence of matching degrees of synchronous auditory and visual information in videos of real-world events on cognitive integration: An event-related potential study. *Neuroscience*, *194*, 19–26. https://doi.org/10.1016/j.neuroscience.2011.08.009

Mai, G., & Wang, W. S.-Y. (2019). *Delta and theta neural entrainment during phonological and semantic processing in speech perception* (p. 556837). bioRxiv. https://doi.org/10.1101/556837

Makeig, S., Delorme, A., Westerfield, M., Jung, T.-P., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2004). Electroencephalographic Brain Dynamics Following Manually Responded Visual Targets. *PLOS Biology*, *2*(6), e176. https://doi.org/10.1371/journal.pbio.0020176

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mazzini, S., Holler, J., & Drijvers, L. (2023). Studying naturalistic human communication using dual-EEG and audio-visual recordings. *STAR Protocols*, *4*(3), 102370. https://doi.org/10.1016/j.xpro.2023.102370

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, *20*(3), 336–366. https://doi.org/10.1111/josl.1_12177

Mozuraitis, M., Stevenson, S., & Heller, D. (2016). Combining Multiple Perspectives in Language Production: A Probabilistic Model. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *38*(0). https://escholarship.org/uc/item/0q95n4qq

Mozuraitis, M., Stevenson, S., & Heller, D. (2018). Modeling Reference Production as the Probabilistic Combination of Multiple Perspectives. *Cognitive Science*, *42*(S4), Article S4. https://doi.org/10.1111/cogs.12582

Mullen, T. R., Kothe, C. A. E., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.-P., & Cauwenberghs, G. (2015). Real-time neuroimaging and cognitive monitoring using

wearable dry EEG. *IEEE Transactions on Biomedical Engineering*, *62*(11), 2553–2567. https://doi.org/10.1109/TBME.2015.2481482

Niefind, F., & Dimigen, O. (2016). Dissociating parafoveal preview benefit and parafovea-on-fovea effects during reading: A combined eye tracking and EEG study. *Psychophysiology*, *53*(12), 1784–1798. https://doi.org/10.1111/psyp.12765

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus—Norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532. https://doi.org/10.1037/0033-2909.131.4.510

Ojeda, A., Bigdely-Shamlo, N., & Makeig, S. (2014). MoBILAB: An open source toolbox for analysis and visualization of mobile brain/body imaging data. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00121

Olson, E. (2011). AprilTag: A robust and flexible visual fiducial system. *2011 IEEE International Conference on Robotics and Automation*, 3400–3407. https://doi.org/10.1109/ICRA.2011.5979561

Özyürek, A. (2021). *Considering the Nature of Multimodal Language from a Crosslinguistic Perspective* (No. 1). *4*(1), Article 1. https://doi.org/10.5334/joc.165

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line Integration of Semantic Information from Speech and Gesture: Insights from Event-related Brain Potentials. *Journal of Cognitive Neuroscience*, *19*(4), Article 4. https://doi.org/10.1162/jocn.2007.19.4.605

Peeters, D., Chu, M., Holler, J., Hagoort, P., & Özyürek, A. (2015). Electrophysiological and kinematic correlates of communicative intent in the planning and production of pointing gestures and speech. *Journal of Cognitive Neuroscience*, *27*(12), Article 12.

Peeters, D., Snijders, T. M., Hagoort, P., & Özyürek, A. (2017). Linking language to the visual world: Neural correlates of comprehending verbal reference to objects through pointing and visual cues. *Neuropsychologia*, *95*, 21–29. https://doi.org/10.1016/j.neuropsychologia.2016.12.004

Pérez, A., Carreiras, M., & Duñabeitia, J. A. (2017). Brain-to-brain entrainment: EEG interbrain synchronization while speaking and listening. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-04464-4

Perniss, P. (2018). Why We Should Study Multimodal Language. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01109

Richter, M., Paul, M., Höhle, B., & Wartenburger, I. (2020). Common Ground Information Affects Reference Resolution: Evidence From Behavioral Data, ERPs, and Eye-Tracking. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.565651

Ries, A. J., Slayback, D., & Touryan, J. (2018). The fixation-related lambda response: Effects of saccade magnitude, spatial frequency, and ocular artifact removal. *International Journal of Psychophysiology*, *134*, 1–8. https://doi.org/10.1016/j.ijpsycho.2018.09.004

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), Article 9. https://doi.org/10.1016/j.tics.2016.07.002

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., & Schlesewsky, M. (2007). To Predict or Not to Predict: Influences of Task and Strategy on the Processing of Semantic Relations. *Journal of Cognitive Neuroscience*, *19*(8), Article 8. https://doi.org/10.1162/jocn.2007.19.8.1259

Ryskin, R., Stevenson, S., & Heller, D. (2020). Probabilistic weighting of perspectives in dyadic communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*(0). https://escholarship.org/uc/item/35m9s3fp

Sassenhagen, J. (2019). How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. *Language, Cognition and Neuroscience*, *34*(4), 474–490. https://doi.org/10.1080/23273798.2018.1502458

Sauppe, S. (2016). Verbal Semantics Drives Early Anticipatory Eye Movements during the Comprehension of Verb-Initial Sentences. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00095

Sauppe, S. (2017). Word Order and Voice Influence the Timing of Verb Planning in German Sentence Production. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01648

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

Schiel, F. (2015). A Statistical Model for Predicting Pronounciation. *Proc. of the ICPhS*, *195*.

Sikos, L., Tomlinson, S. B., Heins, C., & Grodner, D. J. (2019). What do you know? ERP evidence for immediate use of common ground during online reference resolution. *Cognition*, *182*, 275–285. https://doi.org/10.1016/j.cognition.2018.10.013

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two Neurocognitive Mechanisms of Semantic Integration during the Comprehension of Visual Real-world Events. *Journal of Cognitive Neuroscience*, *20*(11), 2037–2057. https://doi.org/10.1162/jocn.2008.20143

Smith, N. J., & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*(2), 157–168. https://doi.org/10.1111/psyp.12317

Smith, N. J., & Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, *52*(2), 169–181. https://doi.org/10.1111/psyp.12320

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the Hands to Identify Who Does What to Whom: Gesture and Speech Go Hand-in-Hand. *Cognitive Science*, *33*(1), 115–125. https://doi.org/10.1111/j.1551-6709.2008.01006.x

Somashekarappa, V., Howes, C., & Sayeed, A. (2020). An Annotation Approach for Social and Referential Gaze in Dialogue. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 759–765). European Language Resources Association. https://aclanthology.org/2020.lrec-1.95

Staudte, M., Ankener, C., Drenhaus, H., & Crocker, M. W. (2021). Graded expectations in visually situated comprehension: Costs and benefits as indexed by the N400. *Psychonomic Bulletin & Review*, *28*(2), 624–631. https://doi.org/10.3758/s13423-020-01827-3

Tromp, J., Peeters, D., Meyer, A. S., & Hagoort, P. (2018). The combined use of virtual reality and EEG to study language processing in naturalistic environments. *Behavior Research Methods*, *50*(2), 862–869. https://doi.org/10.3758/s13428-017-0911-9

Twomey, D. M., Murphy, P. R., Kelly, S. P., & O'Connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *European Journal of Neuroscience*, *42*(1), 1636–1643. https://doi.org/10.1111/ejn.12936

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158–171. https://doi.org/10.1016/j.brainres.2006.06.091

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Verleger, R., Jaśkowski, P., & Wascher, E. (2005). Evidence for an Integrative Role of P3b in Linking Reaction to Perception. *Journal of Psychophysiology*, *19*(3), 165–181. https://doi.org/10.1027/0269-8803.19.3.165

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130292. https://doi.org/10.1098/rstb.2013.0292

Wang, J. J., Tseng, P., Juan, C.-H., Frisson, S., & Apperly, I. A. (2019). Perspective-taking across cultures: Shared biases in Taiwanese and British adults. *Royal Society Open Science*, *6*(11), Article 11. https://doi.org/10.1098/rsos.190540

Wang, J., & Olson, E. (2016). AprilTag 2: Efficient and robust fiducial detection. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4193–4198. https://doi.org/10.1109/IROS.2016.7759617

Wang, Q., Zhang, Q., Sun, W., Boulay, C., Kim, K., & Barmaki, R. L. (2023). A scoping review of the use of lab streaming layer framework in virtual and augmented reality

research. *Virtual Reality*, *27*(3), 2195–2210. https://doi.org/10.1007/s10055-023-00799-8

Wang, S., Mamelak, A. N., Adolphs, R., & Rutishauser, U. (2018). Encoding of Target Detection during Visual Search by Single Neurons in the Human Brain. *Current Biology*, *28*(13), 2058-2069.e4. https://doi.org/10.1016/j.cub.2018.04.092

Willems, R. M., Özyürek, A., & Hagoort, P. (2008). Seeing and Hearing Meaning: ERP and fMRI Evidence of Word versus Picture Integration into a Sentence Context. *Journal of Cognitive Neuroscience*, *20*(7), 1235–1249. https://doi.org/10.1162/jocn.2008.20085

Wu, S., Barr, D. J., Gann, T. M., & Keysar, B. (2013). How culture influences perspective taking: Differences in correction, not integration. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00822

Wu, S., & Keysar, B. (2007). The Effect of Culture on Perspective Taking. *Psychological Science*, *18*(7), 600–606. https://doi.org/10.1111/j.1467-9280.2007.01946.x

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, *42*(6), 654–667. https://doi.org/10.1111/j.1469-8986.2005.00356.x

Zappa, A., Bolger, D., Pergandi, J.-M., Mallet, P., Dubarry, A.-S., Mestre, D., & Frenck-Mestre, C. (2019). Motor resonance during linguistic processing as shown by EEG in a naturalistic VR environment. *Brain and Cognition*, *134*, 44–57. https://doi.org/10.1016/j.bandc.2019.05.003

Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: Word predictability, prosody, gesture and mouth movements in natural

language comprehension. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1955), 20210500. https://doi.org/10.1098/rspb.2021.0500

Figures

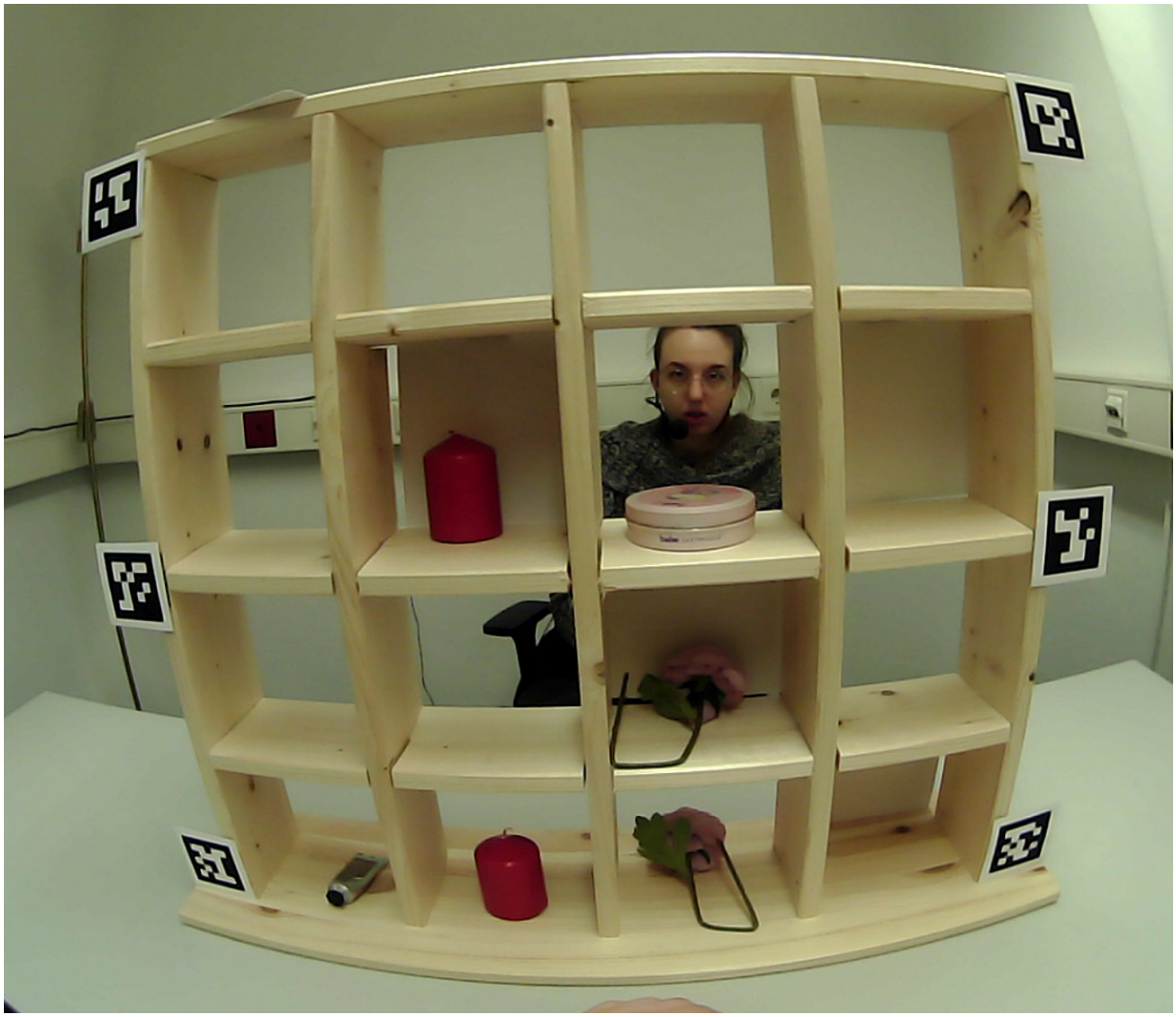

Figure 1: Experimental setup in the current study.

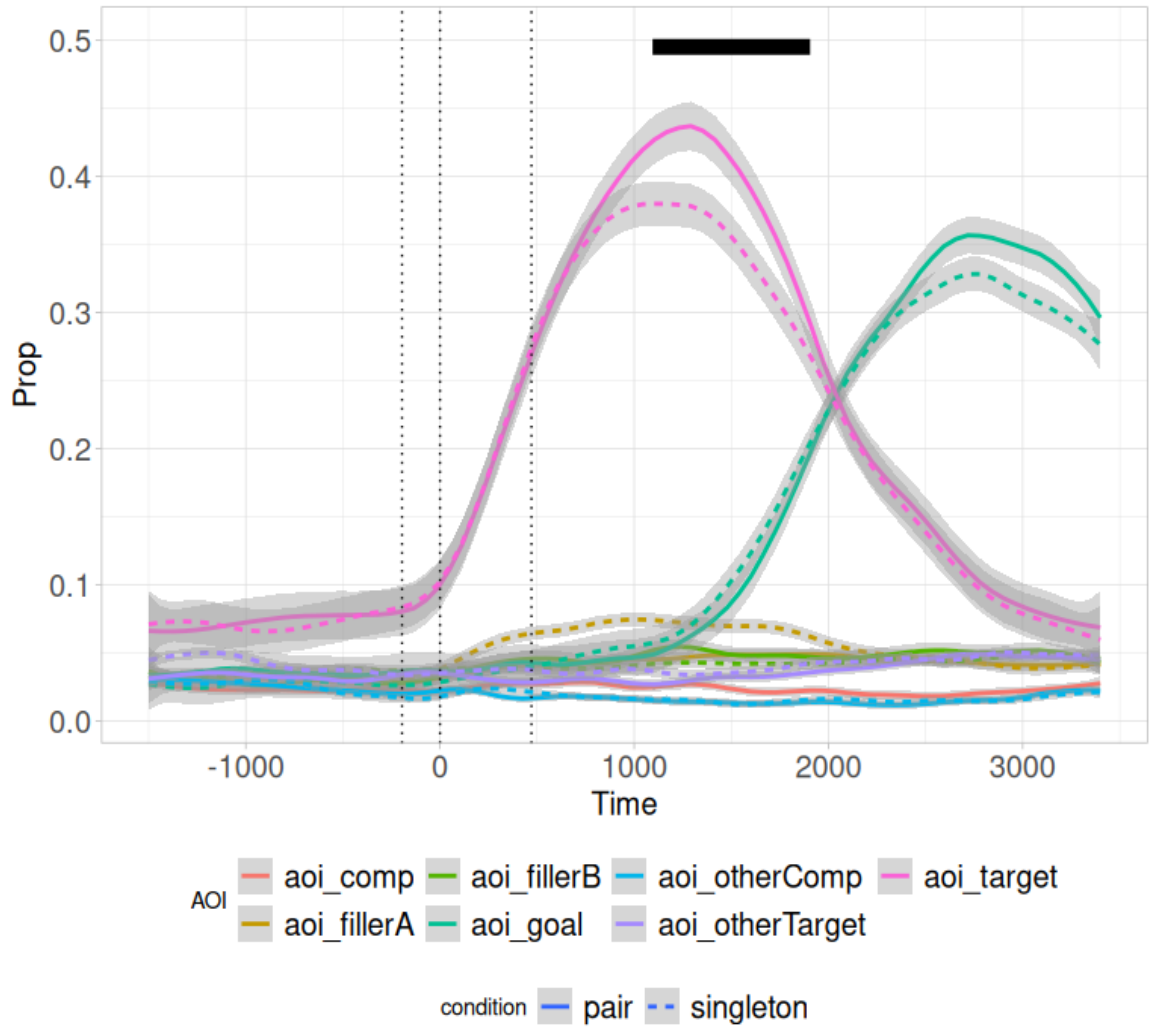

Figure 2: Gaze proportions by condition (solid: pair, dashed: singleton) and AOI (color). The black bar marks significant differences between conditions. Abbreviations: target: the target object; comp: competitor; fillerA/B: singleton, non-target objects; otherComp: the competitor of the non-target pair; otherTarget: the non-hidden target of the other pair; goal: the goal compartment, i.e. the intended end point of the target object in the shelf.

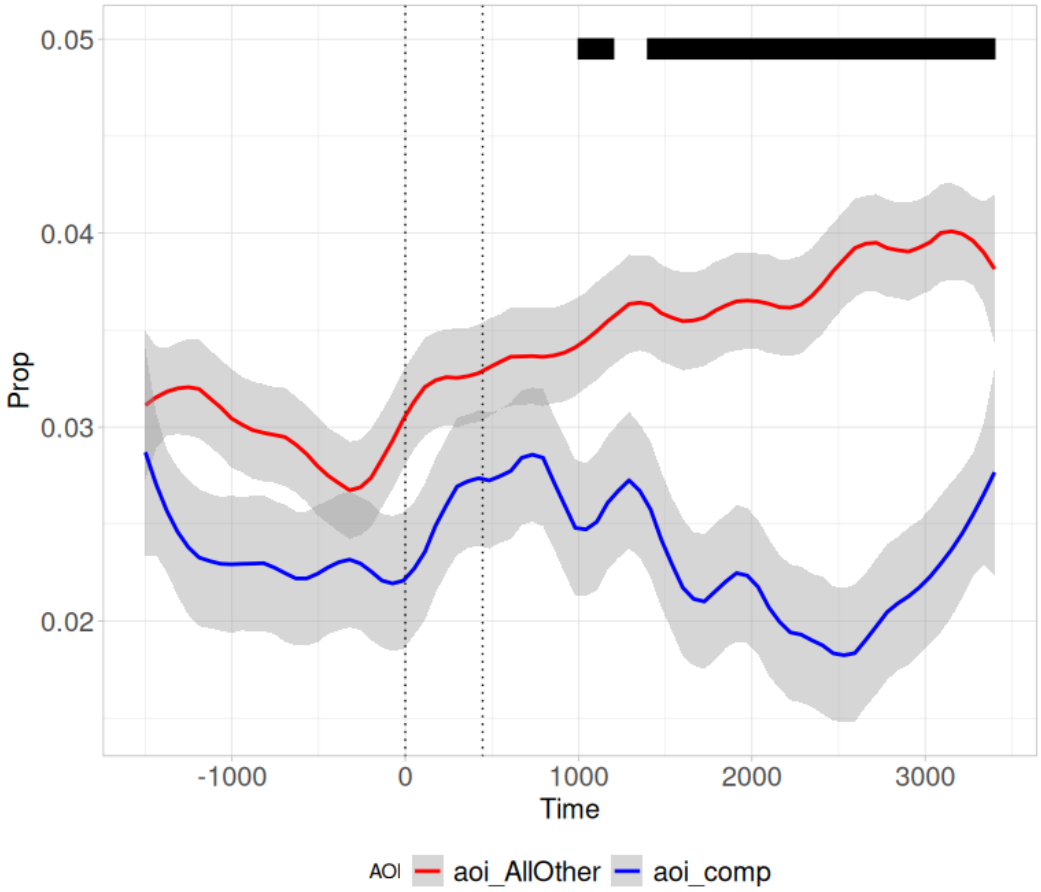

Figure 3 Gaze proportions to the hidden competitor (blue) and the mean over all other objects in the shelf (red) in the pair condition. Numerically, the gaze proportions to the hidden competitor are lower at all time points from -3.5 to 3.5 seconds after noun onset. Statistically significant differences are marked by black bars.

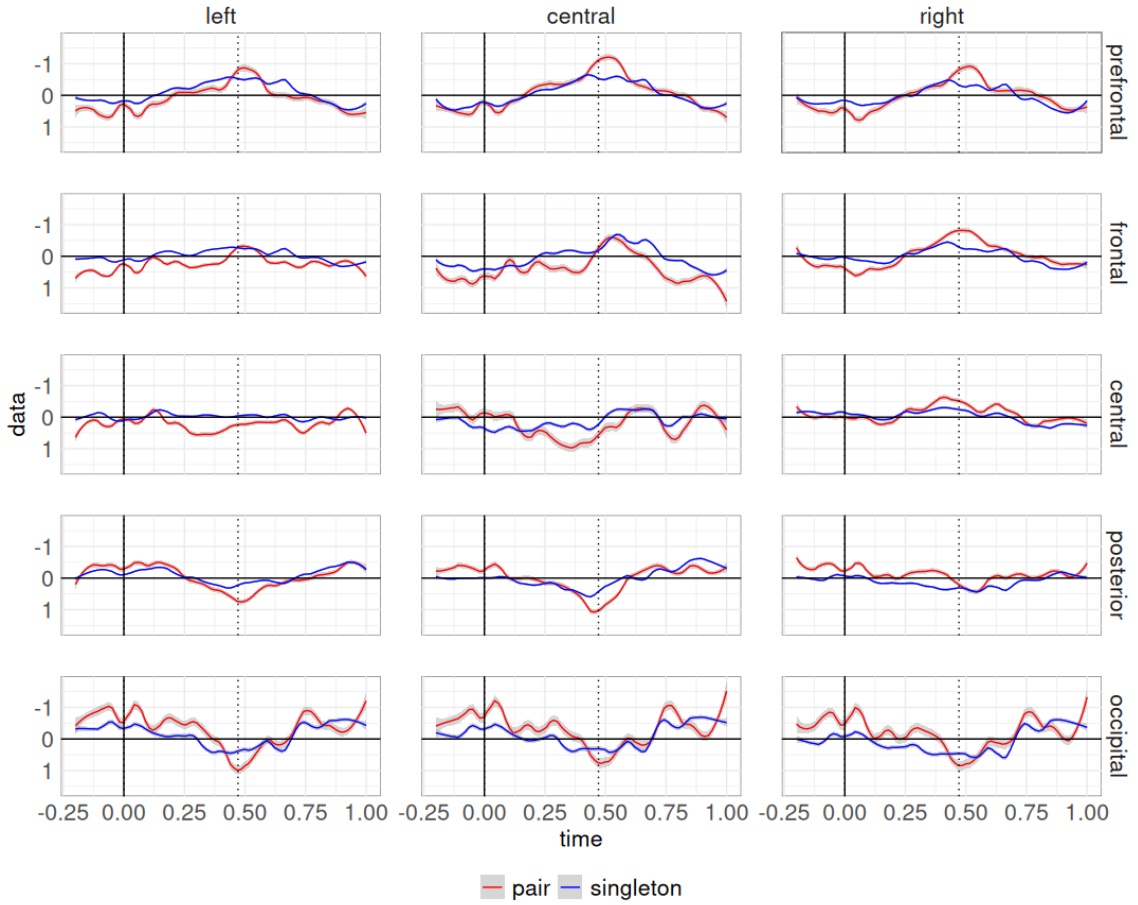

Figure 4: Overlap-corrected, reconstructed ERPs time-locked to noun onset by condition and region-of-interest. The dotted line marks the median noun duration.

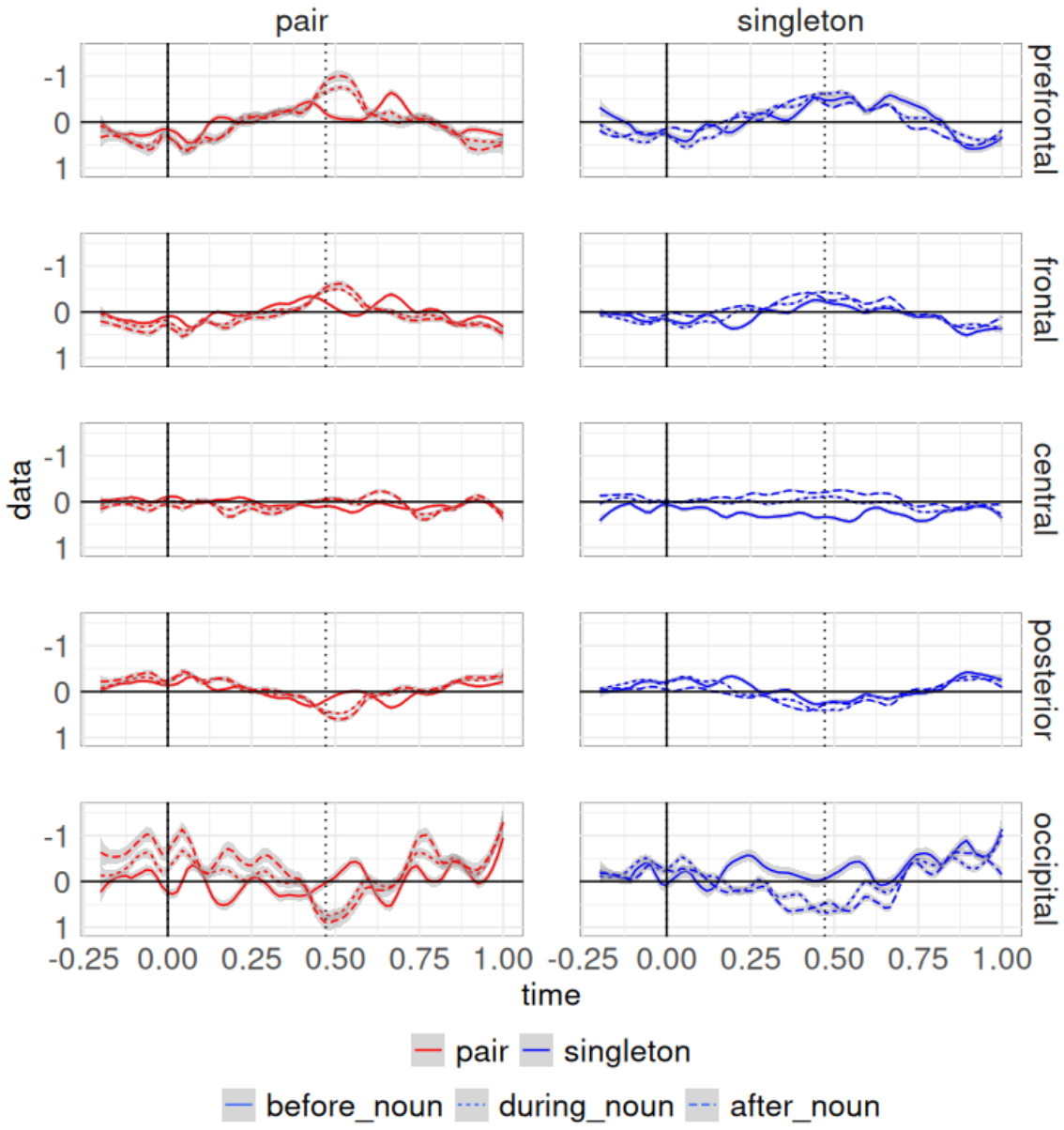

Figure 5: Overlap-corrected, reconstructed ERPs ime-locked to noun onset by condition, saggital regions-of-interest and mean target fixation time. The dotted line marks the median noun duration.

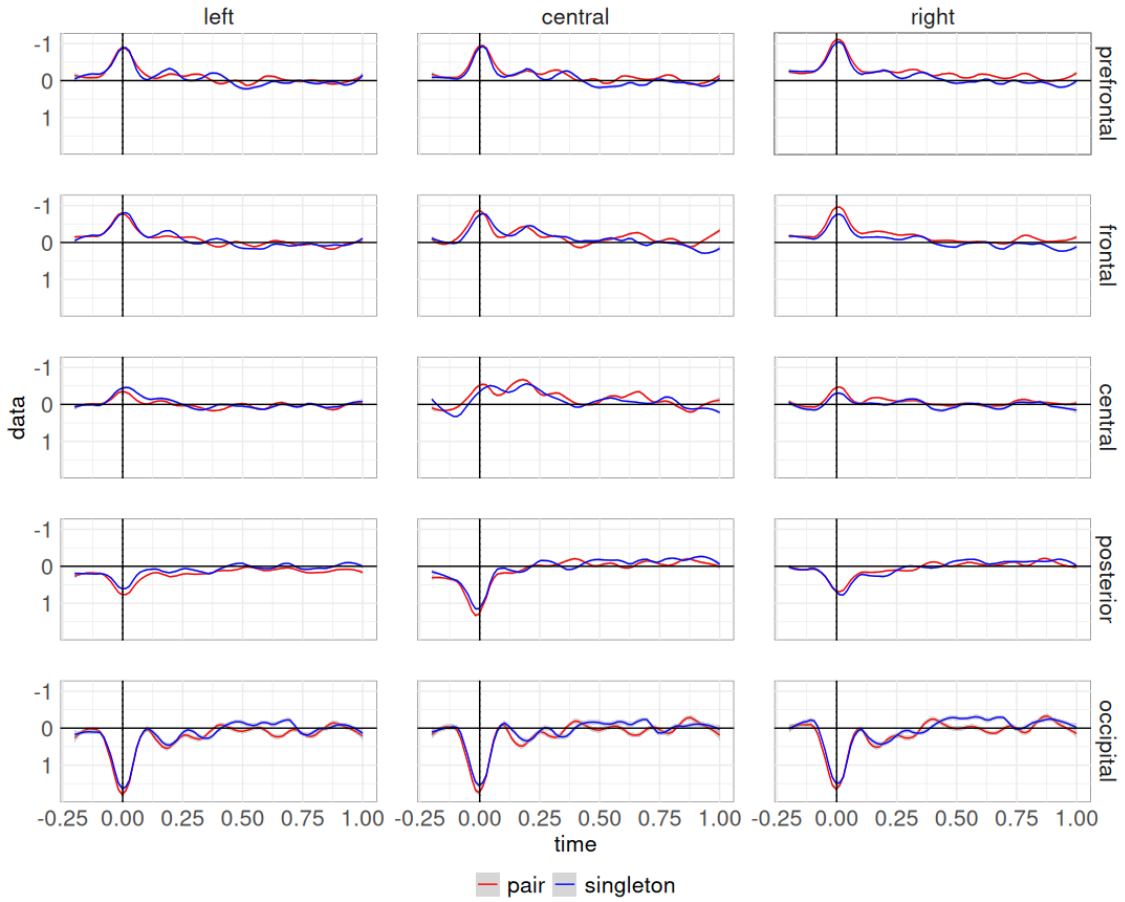

Figure 6: Fixation-related potentials from -200 to 1000 ms after fixation onset by condition and region-of-interest.

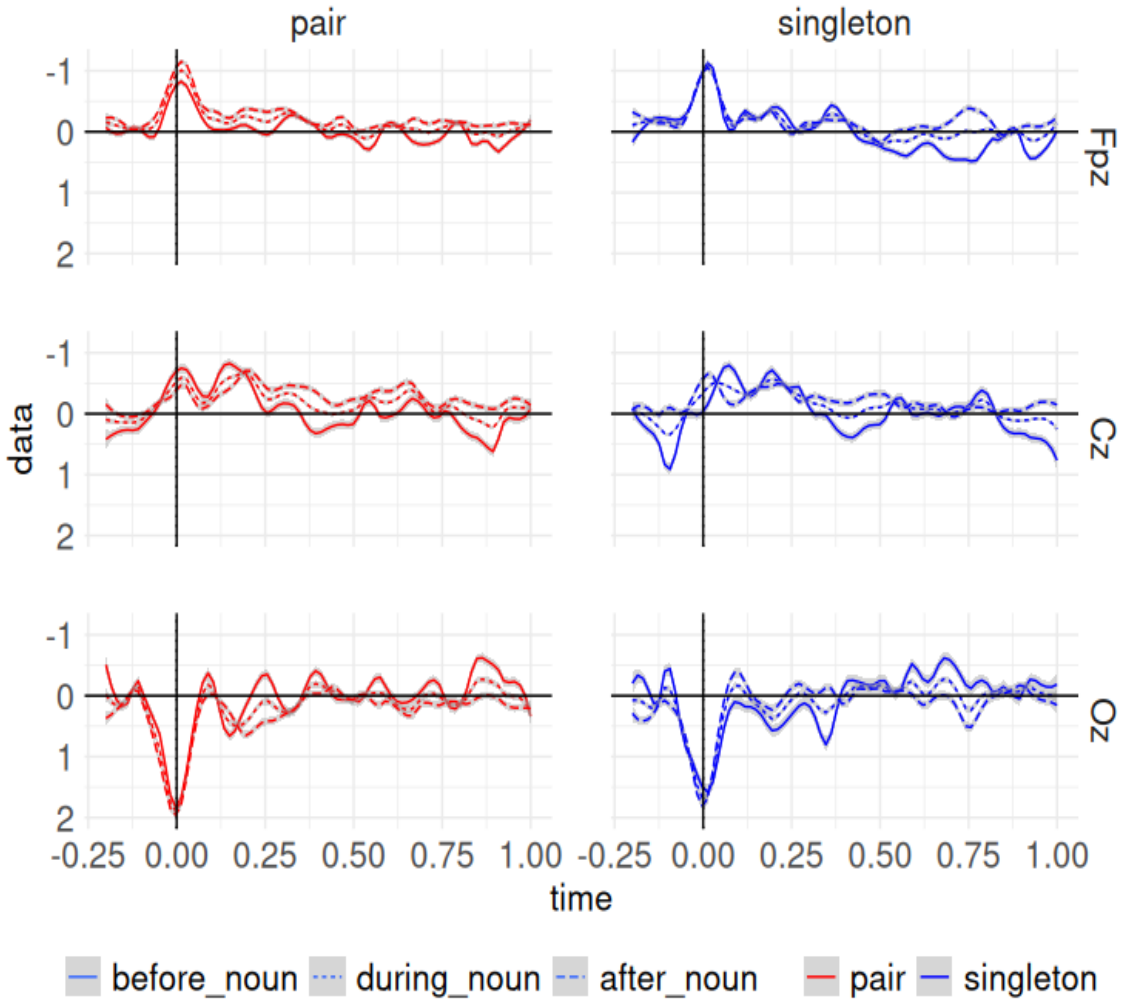

Figure 7: Fixation-related potentials of fixations to targets from -200 to 1000 ms after fixation by condition and fixation time at selected channels.

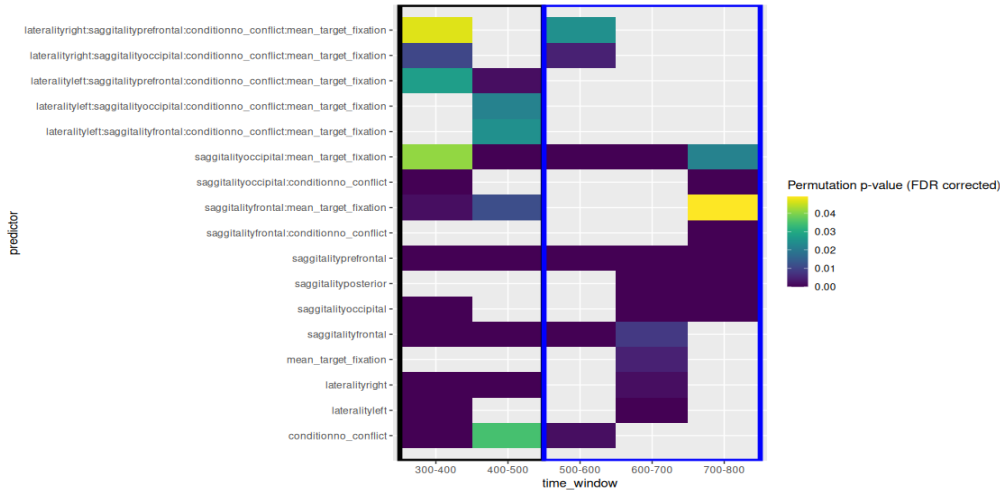

Figure 8: Significant terms as revealed by the permutation analysis for the language-related event-related potentials.

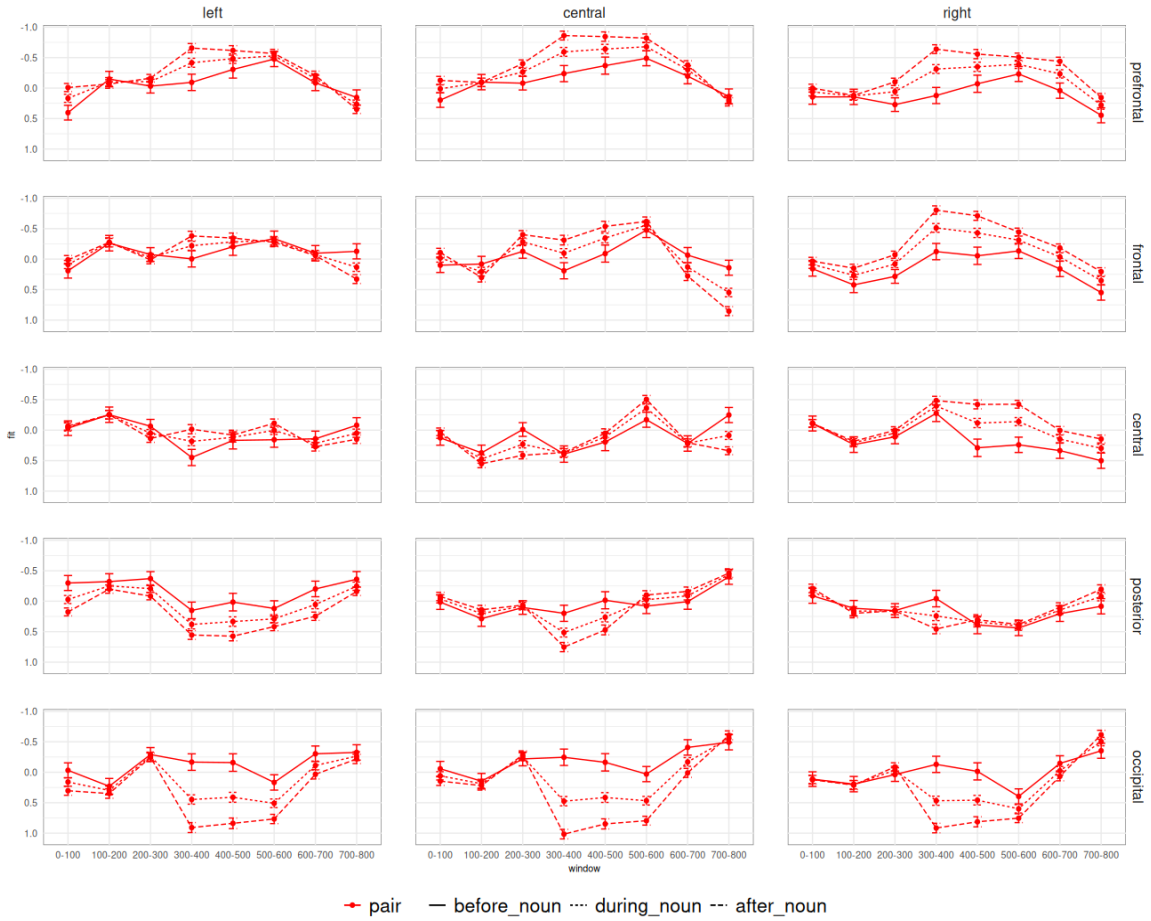

Figure 9: Effects of mean target fixation time in 100 ms time bins from 0-800 ms after noun onset for the pair condition. Error bars represent 83 % confidence intervals, corresponding roughly to 0.05 significance threshold. The continuous predictor was grouped for plotting purposes.

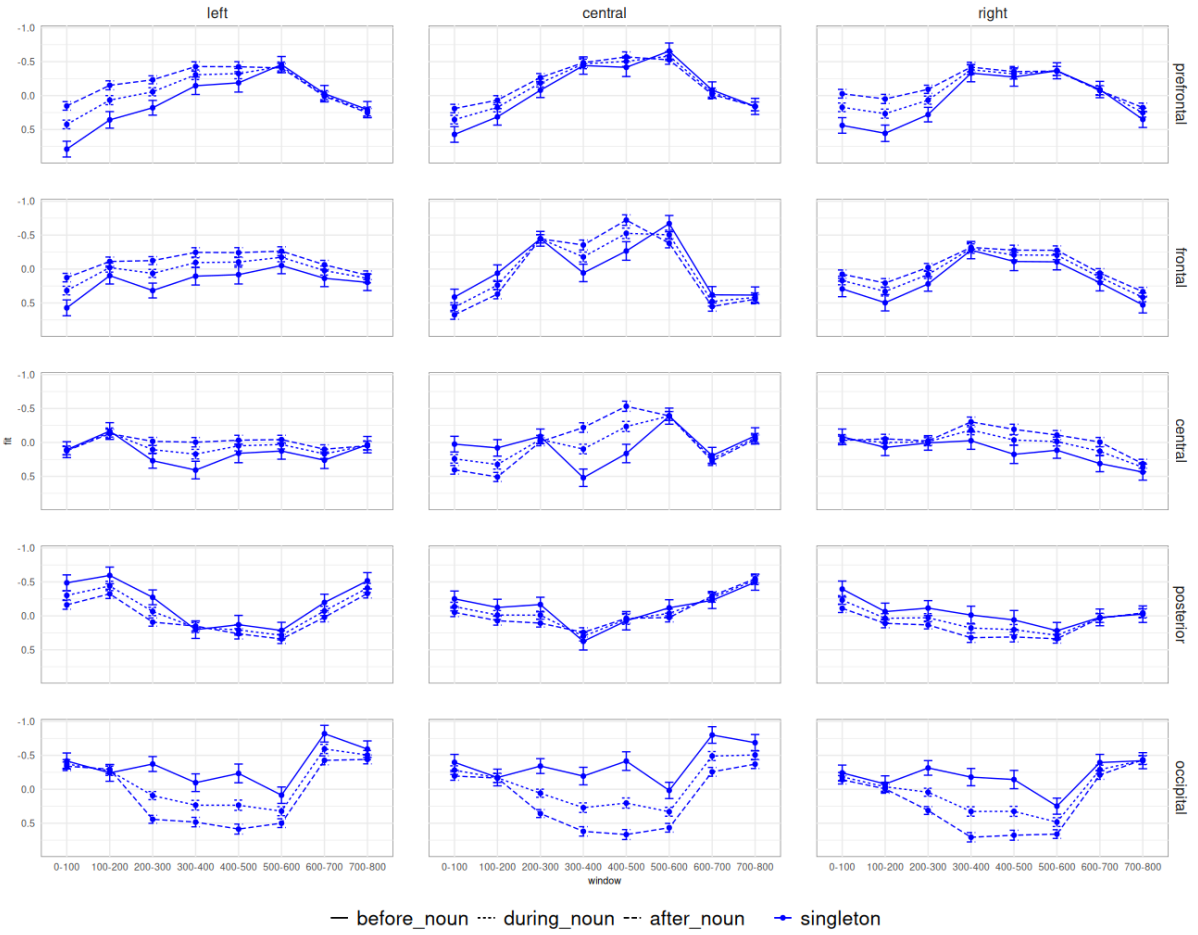

Figure 10: Effects of mean target fixation time in 100 ms time bins from 0-800 ms after noun onset for the singleton condition. Error bars represent 83 % confidence intervals, corresponding roughly to 0.05 significance threshold. The continuous predictor was grouped for plotting purposes.

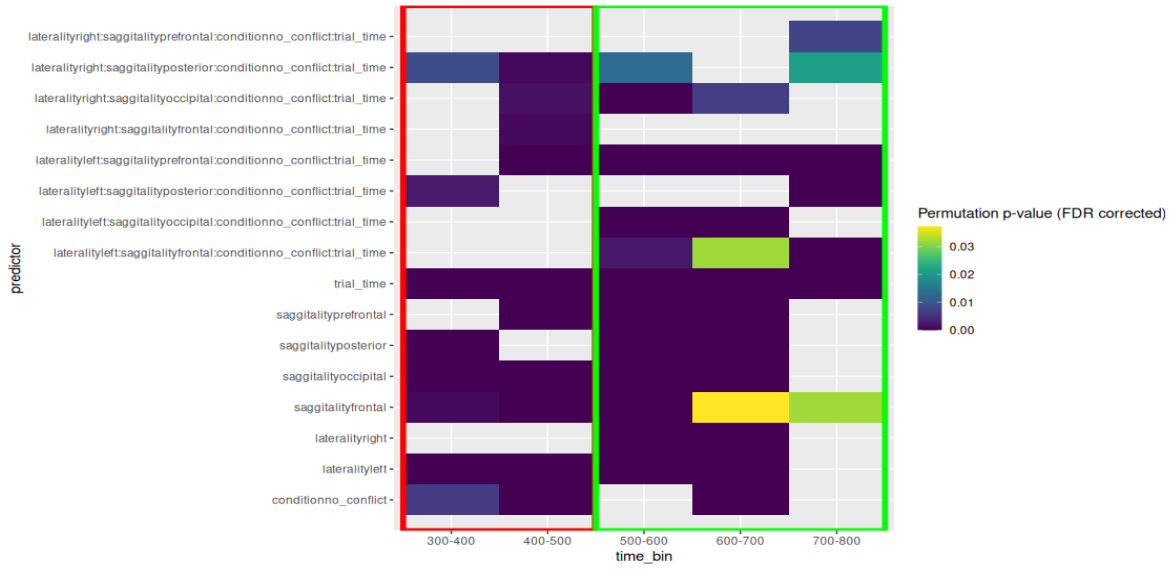

Figure 11: Significant terms as revealed by the permutation analysis for the fixation-related potentials.

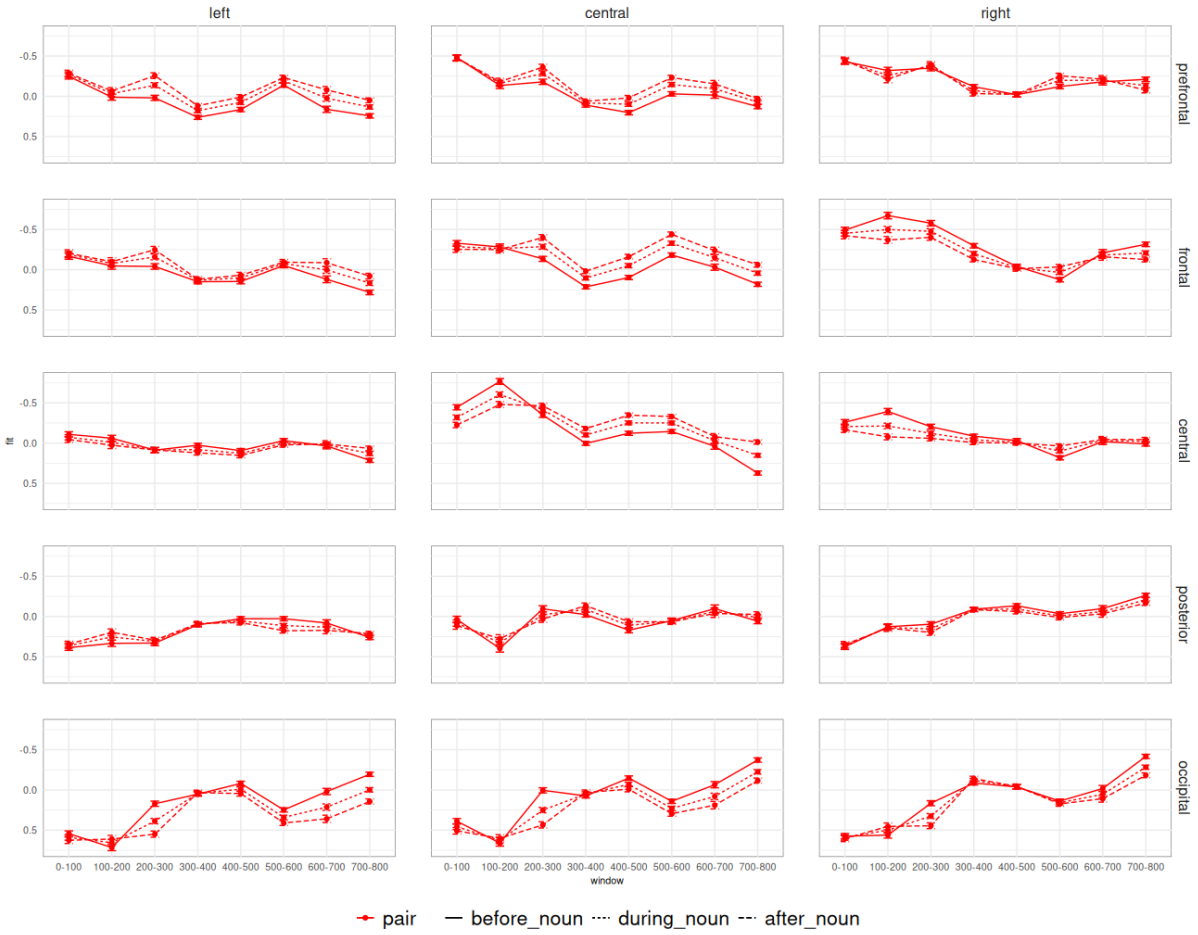

Figure 12: Effects of target fixation time in 100 ms time bins from 0-800 ms after fixation onset for the pair condition. Error bars represent 83 % confidence intervals, corresponding roughly to 0.05 significance threshold. The continuous predictor was grouped for plotting purposes.

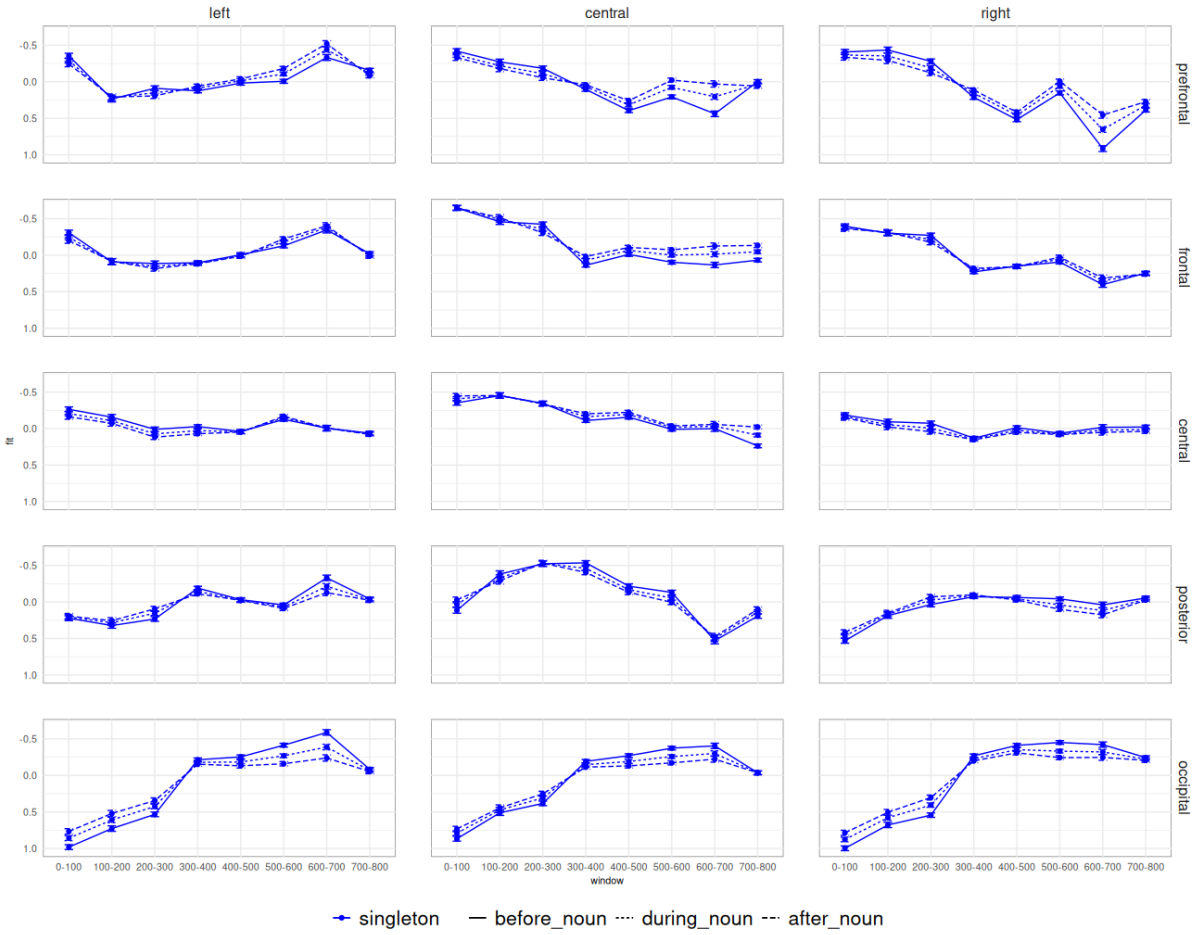

Figure 13: Effects of target fixation time in 100 ms time bins from 0-800 ms after fixation onset for the singleton condition. Error bars represent 83 % confidence intervals, corresponding roughly to 0.05 significance threshold. The continuous predictor was grouped for plotting purposes.

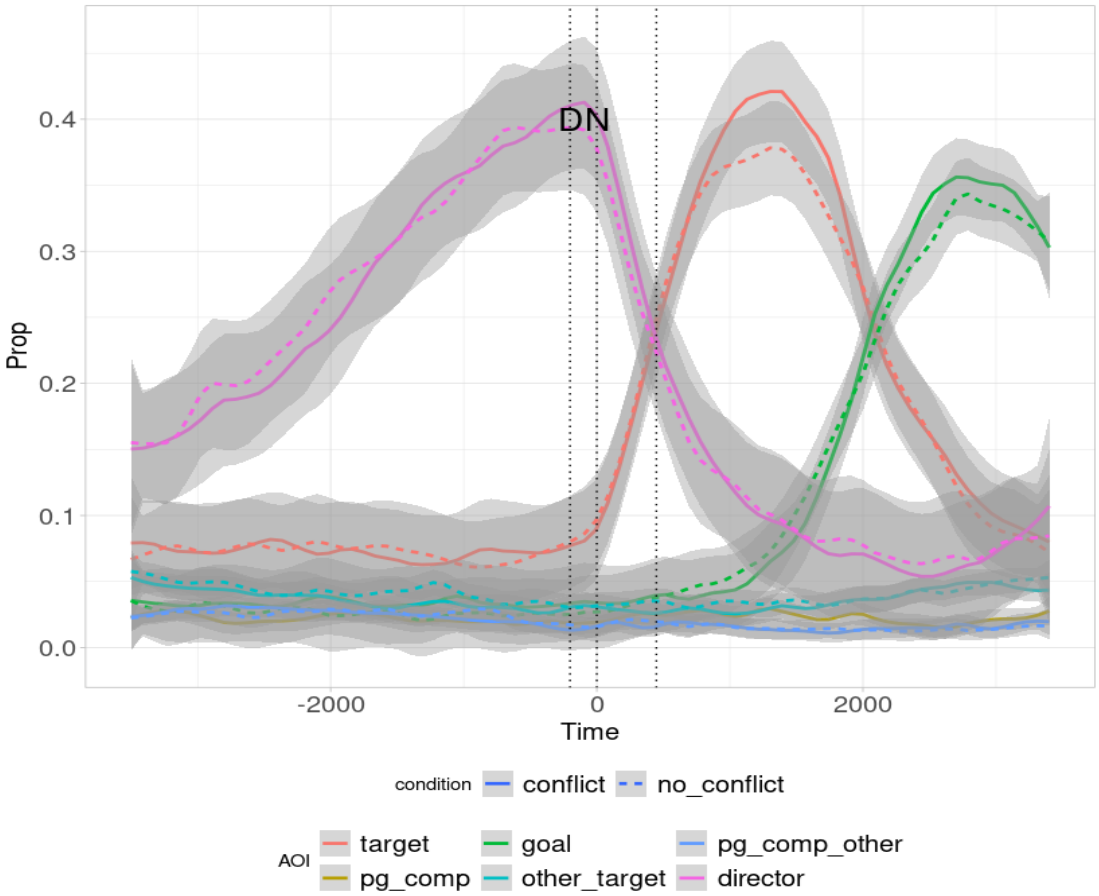

Figure 14: Gaze proportions including gazes at the face of the director for a subset of participants (n=19).