The role of co-speech gestures in retrieval and prediction during naturalistic multimodal narrative processing

Sergio Osorio[a,b,*,1], Benjamin Straube [c], Lars Meyer [d,e,+], & Yifei He [c,+]

[a] Laboratory for Cognitive and Evolutionary Neuroscience, School of Medicine, Pontificia Universidad Católica de Chile, Santiago de Chile, Chile

[b] Interdisciplinary Centre for Neuroscience, Pontificia Universidad Católica de Chile, Santiago de Chile, Chile

[c] Translational Neuroimaging Group, Department of Psychiatry and Psychotherapy Philipps-Universität Marburg, Marburg, Germany

[d] MPRG Language Cycles, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[e] Clinic for Phoniatrics and Pedaudiology, University Hospital Münster, Germany

* Corresponding author: srosorio@uc.cl

+ Shared senior authors

[1] Current affiliation: Department of Neurology and Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA.

Abstract

During daily communication, visual cues such as gestures accompany the speech signal and facilitate semantic processing. However, how gestures impact lexical retrieval and semantic prediction, especially in a naturalistic setting, remains unclear. Here, participants watched a naturalistic multimodal narrative, where an actor narrated a story and spontaneously produced co-speech gestures. For all content words, word frequency and lexical surprisal were regressed against the EEG using temporal response functions (TFRs), which were fitted separately, additively, and interactively for words accompanied and not accompanied by gestures. Results from our analyses suggest a robust modulation effect of gesture on the frequency-dependent regression N400. Besides, we also observed some evidence of modulative effect of gesture on the surprisal-N400 effect based on the single-predictor model. Our finding thus suggests that, on a neural level, the presence of co-speech gestures facilitates lexical retrieval and potentially semantic prediction during the processing of naturalistic multimodal stimuli.

Keywords: multimodality, co-speech gestures, surprisal, word frequency, regression ERP, lexical retrieval, semantic prediction, N400

Introduction

Most human communication is essentially multimodal—besides speech, we use a range of visual signals such as eye-gaze, body orientation, and hand gestures, to convey meaning and social communicative intent. Multimodality has been proposed to facilitate information transmission (Holler & Levinson, 2019; McNeill, 2008). Amongst the modalities, hand gestures (hereafter referred to as gestures) stand out as particularly unique and effective means of expressing complex meaning together with accompanied speech (He et al., 2015; Kelly et al., 2010; Özyürek, 2014; Özyürek et al., 2007). More importantly, they are reported to have a positive impact on speech perception and auditory sentence comprehension, by facilitating speech processing at various levels (Alibali & Kita, 2010; Bosker & Peeters, 2021; Cuevas et al., 2019; Drijvers & Özyürek, 2017; Holle et al., 2012; Kelly et al., 2010; Y. Zhang et al., 2021).

Recent electrophysiological studies suggest that gestures modulate amplitudes of evoked activities for both speech perception and sentence processing. At lower perceptual levels, co-speech gestures modulate the early N1–P2 components when single words are being processed (Kelly et al., 2004; Sun et al., 2021). At the semantic level, humans automatically integrate gesture and speech semantics during online processing, as reflected in the N400 component (Fabbri-Destro et al., 2015; Kelly et al., 2004; Özyürek et al., 2007; Willems et al., 2007; Wu & Coulson, 2005). Thus, increasing evidence suggests that gestures modulate N400 amplitude when word or sentence semantics are being processed (He et al., 2020; Holle & Gunter, 2007; Morett et al., 2020; Wang & Chu, 2013; Y. Zhang et al., 2021).

However notably, these previous studies have predominantly made use of factorial designs where the brain response to single words and sentences is investigated under carefully controlled conditions. Although such designs have long been the basis for our understanding of speech processing and sentence comprehension, questions have been raised about whether observed effects actually hold during ecologically valid communication scenarios (Hamilton & Huth, 2020; Kandylaki & Bornkessel-Schlesewsky, 2019; Meyer et al., 2020; Willems et al., 2020). More recent studies on auditory language processing have therefore investigated how different lexical and contextual features modulate the N400 by using experimental paradigms that employ naturalistic speech stimuli such as long auditory narratives (Alday et al., 2017; Broderick et al., 2018; Goldstein et al., 2022; Sassenhagen, 2019; Yan & Jaeger, 2020). In line with the classic sentence processing literature using factorial experiments (Kutas & Federmeier, 2011; Van Petten & Kutas, 1990), these studies suggest that the N400 is modulated by either the lexical frequency of words that reflects an automatic, bottom-up retrieval mechanism (Sassenhagen, 2019), or by metrics that measure more context-dependent predictive mechanisms, such as semantic similarity and lexical surprisal (Broderick et al., 2018; Goldstein et al., 2022).

In the EEG literature on the semantic processing of co-speech gestures, in contrast, despite heterogeneity in design, most studies either directly compared the N400 amplitude of single words as accompanied by different types of gestures (e.g., Wang & Chu, 2013), or measured the semantic N400 effect in response to match/mismatch between speech and gestures (e.g., He et al., 2020; Özyürek et al., 2007). In this regard, these studies provide us with insights into either semantic processing in general (as reflected in the N400 amplitude), or rather the interplay between semantic prediction and integration as reflected in the N400 effect arisen from semantic mismatch (Lau et al., 2008; Nieuwland et al., 2020). As a result, it remains unclear how different sub-stages of semantic processes such as lexical retrieval or semantic prediction is influenced by gestures. Also importantly, the naturalistic approach as implemented in the auditory speech processing literature has not yet been adapted to shed light on these issues.

Here, we conducted an EEG study employing a naturalistic paradigm to investigate whether and how co-speech gestures modulate the N400 as reflecting word-by-word lexical retrieval and semantic prediction. To this end, we obtained measures of frequency (which quantifies lexical retrieval) and surprisal (which quantifies a word's unpredictability as a measure of semantic prediction) for each word in a multimodal narrative. We then isolated brain responses to individual content words while controlling for overlapping responses to neighboring lexical items using multivariate time-resolved regression (Crosse et al., 2016; Ehinger & Dimigen, 2019; Sassenhagen, 2019). For lexical retrieval, it has been shown that words that occur more frequently in daily communication generally elicit less pronounced N400s than rare words (Sassenhagen, 2019; Van Petten & Kutas, 1990). Word unpredictability, which can be measured using human ratings (i.e., cloze probability) or computational measures of conditional probability derived from Probabilistic Language Models (PLMs), has in turn been also extensively used as a predictor of cognitive load during semantic processing (Frank et al., 2015; Huizeling et al., 2022; Kutas & Federmeier, 2011; Kutas & Hillyard, 1980; Monsalve et al., 2012; Rayner & Duffy, 1986). Here, we quantified words' unpredictability with lexical surprisal, an information-theoretic measure derived from a deep neural network (GPT-2), as a measure of lexicosemantic pre-activation during naturalistic language processing. GPT-2 is a pre-trained transformer-based model that can estimate the unexpectedness of a lexical item considering its previous context, based on a large training database of written text. GPT-2 outperforms other types of language models in generating upcoming words (i.e., lexical-semantic prediction), including simple embedding and recurrent-neural network models. It was also found that metabolic activity during language processing is fit well by GPT-2's lexical-semantic predictions (Schrimpf et al., 2021).

Lexical surprisal is defined as the negative conditional log-probability of a word given its preceding context. It quantifies the cognitive demand associated with the construction of semantic representations on the basis of incremental probabilistic disambiguation (Hale, 2001; Levy, 2008). Lexical surprisal values predict reaction times during self-paced reading of single sentences (Monsalve et al., 2012) and gaze duration during natural reading (Goodkind & Bicknell, 2018). Neurobiologically, the N400—as a marker of cognitive effort during lexical-semantic prediction—increases in amplitude with higher lexical surprisal (Frank et al., 2015; Frank & Willems, 2017; Yan & Jaeger, 2020). More recently, GPT-2-derived measures of surprisal have also been shown to be associated with an N400-like regression effect during natural language comprehension (Goldstein et al., 2022; Heilbron et al., 2022).

While it is well-established that both lexical frequency and surprisal correlate with the amplitude of N400 component, whether and how multimodal cues can also modulate this electrophysiological response during naturalistic speech processing is less clear. In a recent study (Y. Zhang et al., 2021), by investigating the processing of isolated multimodal sentences, the authors showed that the regression-based N400 associated with surprisal may be modulated by co-speech gestures, among other multimodal cues (e.g., prosody). Taking a step further, we aimed at extending this work from several novel perspectives: (1) We tested if the facilitative effect of co-speech gestures can be observed in more ecologically valid narratives (Willems et al., 2020). (2) We employed a time-resolved regression technique (multivariate Temporal Response Functions, mTRFs) instead of conventional ERPs, which is tailored to analyze EEG recordings from the processing of continuous speech (Crosse et al., 2016). (3) We examined the potential facilitative effects of gestures on both lexical retrieval (word frequency) and semantic prediction (GPT2-surprisal). Notably, although both measures have been linked to the N400, they are usually highly correlated and may interact under certain contexts (Halgren et al., 2002; Huizeling et al., 2022; Kretzschmar et al., 2015). Consequently, to what extent the frequency–derived or surprisal-derived N400s may be affected by gestures remains unknown.

We hypothesized that, during the processing of a naturalistic multimodal narrative, the presence of gestures would result in a decreased amplitude of the N400 response to both frequency and surprisal, analogous to the dampening of the N400 in the presence of co-speech gestures in factorial experiments. Twenty participants were presented with audiovisual clips in which an actor narrated a story presented in German using co-speech gestures spontaneously (figure 1b). Content words in the speech stimuli were coded according to whether their meanings were either repeated, emphasized, or complemented by accompanying co-speech gestures (*Gesture present*) or not (*Gesture absent*), for all types of gestures that are produced by the actor (see methods). Word frequency was included as the corpus rank bin count from the Projekt Deutscher Wortschatz (Goldhahn et al., 2012), with higher ranks (1–24) indicating less frequent words. Words not found in the corpus were coded as 2 + the highest rank (Sassenhagen, 2019). The GPT-2 transformer model was used to estimate surprisal values. Responses associated with lexical retrieval and semantic predictions were obtained through mTRFs for frequency and surprisal. We estimated model fit for two single-predictor models (one for surprisal and the other one for frequency), as well as for an additive model (frequency + surprisal) and an interaction model (frequency + surprisal + frequency * surprisal), with and without gestures as a categorical predictor. We found that adding gesture information improved model fit regardless of model type, but the additive model showed the best fit to the data. Then, by hypothesis, we set up separate models for frequency and surprisal for gesture and no-gesture words to investigate how the EEG correlates of retrieval and semantic prediction error scale with the presence of co-speech gestures. Individual mTRF responses were modelled as the sum of the product of unknown beta coefficients with individual values of frequency or surprisal per word (see figure 1a for a visual summary of our analytical protocol). Results are in line with our hypotheses: words that were not accompanied by gestures during a naturalistic narrative showed both frequency– and surprisal-related N400 responses, which were reduced for words that were accompanied by gestures. Further analysis including both frequency and surprisal as regressors in the same additive model only revealed significant modulation effect of gestures on frequency-dependent N400 responses. To summarize, our findings provide—for the first time—clear evidence that co-speech gestures are associated with modulated frequency-dependent N400 effects, thus indicating a potentially facilitative role of gesture on bottom-up lexical retrieval. For surprisal, results from our surprisal-only model agree with prior observations showing the facilitative effect of gesture on probabilistic disambiguation during lexical-semantic prediction. However, to what extent this effect can be dissociated from lexical retrieval requires further examination.

(Figure 1 here)

Materials and methods

Participants

Participants (*n* = 20, *mean age* = 24.1 years, *range* = 19-34 years, 14 females) were native German speakers recruited from the Marburg-Giessen area, Germany. Participants were right-handed, had normal hearing and normal or corrected to normal vision. None of them reported any medical or neuropsychiatric condition. Participants read and signed an informed consent before participating in the study. They were compensated with seven euros per hour for participation. The research protocol and procedures were approved by the Ethics Committee at Phillips University Marburg and were conducted in accordance with the Declaration of Helsinki.

Stimuli

Participants watched 16 video clips (individual clips lasting between 1:02 and 3:31 min; story duration = 32:12 minutes) of a professional male actor narrating an adapted version of the story *Der Kuli Kimgun* as naturally as possible. Consent was obtained from the actor to use his image for research and publication purposes. Foreign words in the original story were replaced by German synonyms. We analyzed EEG responses to all content words in the story. Firstly, content words whose semantic representations that were either associated, emphasized, or complemented by gestures were coded as *Gesture present*; all other content words were coded as *Gesture absent*. For example, in figure 1, the screenshots depict two consecutive beat gestures that were used by the actor to emphasize the two “three” in “Samir stayed three days and three nights on the hills”). In this case, even if the gesture onsets from “stay” and offsets around “on”, we only coded the two “three”s as words in the *Gesture* *present* condition, and all other words were coded as *Gesture absent*. This coding applies analogously for non-referential beat gestures. For iconic, metaphoric, and emblems, it is then their lexical affiliates that were coded as gesture present. The coding of the Gesture present/absent conditions were double-checked by two independent expert coders. From the total number of content words in the story (n = 3582), 466 were coded as G*esture present*. The actor was free to decide when and how to make use of gestures. Throughout the story, a total of 493 hand gestures were conducted by the actor. Table 1 reports the frequency of each type of gestures. For a sample description (screenshot and transcript) of the multimodal story, please refer to Cuevas et al., 2019. In supplementary figure s1, we also illustrate a few additional examples of the hand gestures and the corresponding Gesture present words. Notably, despite prior research showing mixed effect of beat gestures on semantic processing and potentially differential effect of iconic gestures and other gestures (Hintz et al., 2022; Morett et al., 2020; Wang & Chu, 2013; Y. Zhang et al., 2021), for the purpose of testing the facilitative role of gestures in general on semantic processing, we collapsed across all gesture types for all analyses for the main analysis (see Cuevas et al., 2019 for an fMRI study using the same stimuli). We nevertheless reported specific effects of both iconic and beat gestures in the supplement.

(Table 1 about here)

Word frequency was included as the corpus rank bin count from the Projekt Deutscher Wortschatz (Goldhahn et al., 2012), with higher ranks (1–24) indicating less frequent words. Words not found in the corpus were coded as 2 + the highest rank. To model word-by-word lexical-semantic processing demands, we employed a deep neural network transformer-based model (GPT-2). Whereas earlier language models (e.g., n-gram models, recurrent neural networks) derive surprisal from serial cumulative calculations on prior word sequences of varying length (e.g., two words, all words from sentence onset), transformers incorporate an analogue of attention to those parts of the context that do maximize word prediction performance (e.g., Radford et al., 2019; Ryu & Lewis, 2021; Schrimpf et al., 2021). While attention can be bidirectional, GPT-2 is a unidirectional forward-attention transformer, consistent with the contingency of human speech. Note also that instead of using raw text, GPT-2 performs next-word prediction based on word embeddings. We here employed the German DBMZ GPT-2 model provided via Huggingface at https://huggingface.co/dbmdz/german-gpt2. Surprisal was calculated from GPT-2's word probabilities using the standard formula. The base of the logarithm was 2.

Summary statistics are provided in table 2. We compared the difference for both frequency and surprisal between the *Gesture present* and *Gesture absent* conditions with a Wilcoxon rank sum test. A significant difference was observed for both surprisal (z = -2.50, *rank sum* = 1.43e+06, p = 0.012) and frequency (z = -3.726, *rank sum* = 1.42e+06, p = 1.94e-04).

(Table 2 about here)

Procedure

Participants sat in a sound attenuated room at 70 centimetres from a monitor and were instructed to watch a professional actor telling the story. Auditory stimuli were delivered using a pair of loudspeakers with volume being constant for all participants.

Data acquisition

Electrophysiological data was acquired using a Brain Products 32-channel EEG system (Brain Products GmbH, Gilching, Germany). Electrodes were positioned according to the 10-20 international standard. EEG data was collected at sampling rate of 250Hz without any on-line filters, referenced to FCz. The EEG dataset has been reported by Sassenhagen (2019) for a different research question.

Data analysis

EEG data preprocessing

Data were preprocessed using a modified version of the Harvard Automated Preprocessing Pipeline (Gabard-Durnam et al., 2018) together with the EEGlab toolbox (Delorme & Makeig, 2004). EEG data were re-referenced to the average of electrodes TP9/10 (mastoids). Bad channels were removed for later interpolation based on joint probability (Delorme & Makeig, 2004). Line noise was removed with ZapLine (de Cheveigné, 2019), and data were lowpass-filtered using an 10-Hz one-pass Hamming sinc FIR filter. For artifact detection and rejection, data underwent Independent Component Analysis (Delorme et al., 2007). Components underwent frequency-domain thresholding (Castellanos & Makarov, 2006). Artifact components were then selected automatically by ICLABEL (Pion-Tonachini et al., 2019), MARA ($p < 0.05$; Winkler et al., 2011) and ADJUST (Mognon et al., 2011). On average across participants, 2.85 components (SD = 1.89) were rejected. Bad channels identified initially were interpolated. Linear detrending was applied. Any remaining data exceeding 50 μV with a moving window of 2 seconds were excluded from statistical analysis. On average, across participants, this resulted in 3.59 % of words (SD = 5.08) being excluded. We defined a group of representative centro-parietal electrodes (electrodes Cz, Pz, C3, C4, P3, P4, CP1, CP2) as our region of interest (ROI), and a 300-500 ms time window as our time window of interest. This is consistent with the topographical features and latency of the N400 response previously reported elsewhere in the literature.

Multivariate time-resolved regression

For multivariate time-resolved regression analyses, we used the mTRF toolbox (Crosse et al., 2016). This method models brain responses using ridge-regression by fitting a multivariate temporal response function (mTRF) to brain signals, which allows mapping between stimulus features and neural activity. For this reason, multivariate time-resolved regression is particularly well-suited to investigate brain responses to continuous, naturalistic stimuli. Another important advantage of ridge-regression is that it is robust against collinearity. This is particularly important because lexical surprisal and lexical frequency values are highly correlated ($r = 0.64$, $p < 0.001$).

Brain responses were modelled for each subject as the sum of the product of unknown beta coefficients with individual values per content word for word onset, lexical frequency, GPT2-derived surprisal, and the interaction between frequency and surprisal, estimated as the product between these two predictors. We analyzed content words only, as function words do not induce much word-by-word N400 (Frank et al., 2015). All predictors except word onset were z normalized before mTRF estimation. We implemented four separate encoding models, one for surprisal using the formula *y ~ word onset + surprisal*, another one for frequency using the formula *y ~ word onset + frequency*, one additive model for the combined effect of surprisal and frequency (*y ~ word onset + surprisal + frequency*) and one last model for the interaction between surprisal and frequency (*y ~ word onset + surprisal + frequency + surprisal*frequency*). Here, importantly, the first two single-predictor models evaluate the effects of frequency and surprisal separately, and how these effects were modulated by gestures (see below). The latter two models consider additionally the situations where both the effect of frequency and surprisal maybe dependent on, and interact with each other (Dambacher et al., 2006; Huizeling et al., 2022; Payne et al., 2015; Van Petten & Kutas, 1990). For model fit evaluation, the four mTRF models were obtained from all content words collapsed regardless of whether they were accompanied by gestures or not. Next, the model's prediction accuracy was evaluated via cross-validation (see next section). Then, the four models were estimated and evaluated again after adding a categorical predictor indicating the presence or absence of gestures. A linear mixed effect model was used to test the effect of model type and gesture information as an additional regressor on model fit.

Model optimization and evaluation

Model optimization and model evaluation were conducted via a *leave-one-out* 10-fold cross-validation procedure using the mTRF toolbox (Crosse et al., 2016). For single predictor models, the lambda (i.e., ridge) parameter was consistently set to zero. For all the other models, optimal lambda parameters were identified by evaluating model fit for a logarithmic space of 31 ridge values between 0.01 and 10 and between 100-600ms after word onset. This resulted in a p*artition-by-lambda-by-sensor* matrix of correlation coefficients. The optimal lambda parameter was programmatically set as the mean spearman r value across the electrodes within the pre-defined ROI that maximized model fit (supplementary table s1). For model evaluation, the predictive power of the trained data after cross-validation was tested against the test data partition, which returns a set of 27 correlation coefficients, one for each channel. The mean spearman r value was then obtained for the set of electrodes within our predefined ROI (supplementary table s2).

Pair-wise comparison between gesture absent and gesture present conditions

In a next step, we separately obtained mTRFs for gesture absent and gesture present content words based on the four models above. This set of analyses, in addition to the model fit comparisons, provides additionally information on if the presence of gesture *enhances* or *modulates* the frequency– or surprisal-dependent N400s. For statistical analyses, we separately obtained each subject's median beta values for gesture present and gesture absent mTRF models in the group of representative centro-parietal electrodes within our ROI, and within a predefined time-window of 300-500ms. This is consistent with the topographical features and latency of the N400 response previously reported elsewhere in the literature. Within each model, we compared the extracted frequency– and surprisal-N400 beta amplitudes by means of a Wilcoxon signed-rank test. We used median values and a non-parametric test because Kolmogorov-Smirnov normality tests indicated that not all distributions were normal (e.g., surprisal-dependent N400s in the gesture absent, $p = 0.02$). Given our strong hypothesis for a decreased N400 amplitude in the presence of a co-speech gesture for both frequency and surprisal, we opted for one-tailed testing, which increases statistical power (Cho & Abe, 2013).

Results

To investigate the overall effect of gestures on mTRFs, we modelled brain responses to all content words. We computed four different models: two single-predictor models using lexical frequency and surprisal separately, one additive model for the combined effect of frequency and surprisal, and a model testing the interaction of the two regressors. We then repeated this procedure after adding to all models an additional categorical predictor indicating the presence or absence of gestures. Because model fit varies numerically from one run to another, we implemented these analyses 10 times and averaged model fit values across runs. We report the model fit averaged across all subjects in table 3. Results show that model fit improves when gestures are included as an additional regressor, regardless of model type. Among the four models that include gestures as a categorical predictor, the additive model shows the best fit. For statistical comparison between model fits, we conducted a linear mixed effect model for model type and gesture information, controlling for the within-subject nature of the design by including random effects for subject and the interaction of subject with model type and subject with gesture information. Results revealed a statistically significant main effect of gesture information ($F = 33.718$, $p = 1.36e\text{-}05$) and model type ($F = 9.68$, $p = 2.94e\text{-}05$) in predicting a better model fit. No statistically significant effect was found for the model type by gesture interaction. For the main effect of model type, post hoc Tukey contrasts revealed significant better model fit for the additive model in comparison to the frequency-only ($z = 2.97$, $p = 0.015$) and the surprisal-only ($z = 3.17$, $p = 8.25e\text{-}03$) models, with no difference between the additive and the interaction models ($z = 0.425$ $p = 0.974$).

(Table 3 here)

Having established the importance of gesture information in model fit, we next investigated the frequency-dependent and the surprisal-dependent mTRF responses for words accompanied by gestures and for words not accompanied by gestures separately. Regarding the effect of frequency, figure 2a shows the mTRF for *Gesture absent* and *Gesture present* words, averaged across a group of representative centroparietal electrodes (see methods). Beta values for *Gesture absent* words are more negative than beta values for *Gesture present* words within a time-window that is consistent with the latency of the classic N400 responses (300–500 ms). We then visually inspected the topographical distribution of the mean mTRF responses between 300 and 500 ms after word onset. Unlike *Gesture present* words, *Gesture absent* words showed a centro-parietal negativity that is also reminiscent of the classical N400 effect (figure 2b). Finally, we extracted the median beta values between 300–500 ms for this group of electrodes (figure 2c). Given our directional hypothesis, we expected the beta values for the *Gesture absent* words to be more negative than those for *Gesture present* words in a time window consistent with the N400. A non-parametric Wilcoxon signed-rank test ($z = -3.49$, signed rank = 11, $p = 2.40e\text{-}04$, one-tailed) showed that the median beta values for *Gesture absent* words ($n = 20$, median = –5.37, $SD = 3.96$) was significantly more negative than the median beta values for *Gesture present* words ($n = 20$, median = 3.07, $SD = 9.01$), with a large effect size ($r = 0.78$). Besides the mTRF results, we also reported the conventional ERP effects for high– vs. low-frequency words (with median split), both collapsed across both conditions and within the *Gesture present* and *Gesture absent* conditions in supplementary figure s2.

(Figure 2 here)

The mTRF results for GPT-2-surprisal are illustrated in figure 3. Similar to the frequency-dependent results, beta values for *Gesture absent* words are more negative than beta values for *Gesture present* words within a time-window that is consistent with the latency and the topography of the classic N400 responses. Again, we extracted the median mTRF beta values between 300–500 ms for the same *a priori* defined centro-parietal electrodes. Given our directional hypothesis, we expected the mTRF response for the *Gesture absent* words to be more negative than mTRF for *Gesture present* words in a time window consistent with the N400. A Wilcoxon signed-rank test ($z = –2.07$, signed rank = 49, $p = 0.019$, one-tailed) showed that the median beta values between 300 and 500 ms for *Gesture absent* words ($n = 20$, median = -2.43, $SD = 3.96$) were significantly more negative than the median beta values for *Gesture present* words ($n = 20$, median = 0.085, $SD = 10.15$), with a moderate effect size ($r = 0.46$). Similar to lexical frequency, conventional ERP effects for high– vs. low-surprisal words (with median split), both collapsed across and within the *Gesture present* and *Gesture absent* conditions, are reported in the supplementary figure s3.

(Figure 3 here)

Notably, within the same time window and for identical electrodes, the modulation effects for the *Gesture absent* and *Gesture present* comparisons are statistically significant for both frequency and GPT-2 surprisal, although the beta coefficients appear to be larger for the frequency-dependent effect by visual inspection. Thus, we compared if gestures elicit a stronger modulatory effect for the regression-based N400 for lexical frequency than for surprisal. To this end, we conducted a Wilcoxon signed-rank test to directly compare the median difference between *Gesture present* and *Gesture absent* responses for frequency (median = 9.54, $SD = 9.77$) and surprisal (median = 5.77, $SD = 9.19$). We found that this difference is indeed significant ($z = 2.20$, signed rank = 164, $p = 0.028$, two-tailed). Thus, it appears that gesture modulates the frequency-derived N400 more strongly than the surprisal-derived N400.

In a further step, based on the fact that the additive model outperforms the single predictor models, we additionally compared the regression coefficient between the Gesture present and absent conditions for both frequency and surprisal, when they were both entered as regressors in the additive model. Results of this analysis are illustrated in figure 4. Here, beta values differ significantly between *Gesture present* and *Gesture absent* conditions for lexical frequency ($z = –2.967$, signed rank = 25, $p = 0.0015$) with a moderate effect size ($r = 0.66$), whereas no significant difference was observed for surprisal ($z = 0.952$, signed rank = 130, $p = 0.829$, $r = 0.21$). This analysis corroborates the effect of gesture on frequency-dependent N400s. Here, however, in comparison to the single predictor effects, the scalp distribution of the frequency effect appears to be more anterior. This pattern seems to be reminiscent to the report on the effect of gesture on speech processing being more anterior (Kandana Arachchige et al., 2021). Therefore, based on the additive model together with the gesture information (with gesture) binary regressor, we additionally compared the model fit between our N400 parietal ROI and a set of frontal electrodes (F1/2/3/4/z, FC1/2). A t-test for the difference between beta values extracted from the parietal and frontal ROIs showed that this difference was not statistically significant ($t = 2.06$, $p = 0.053$).

(Figure 4 here)

Further, even though the model fit did not differ between the additive and the interaction models, we nevertheless conducted a control analysis to investigate the effect of gesture on a potential interaction between frequency and lexical frequency (figure s4, supplementary materials). Results of the mTRF analysis for the interaction model suggest an effect of frequency ($z = -2.86$, signed rank = 28, $p = 0.002$) for a moderate effect size ($r = 0.64$), where words not accompanied by gestures ($z = -3.863$, $SD = 7.25$) are associated to more negative beta coefficients than words accompanied by gestures ($z = 7.70$, $SD = 17.11$). However, the gesture effect on the interaction term in this model was not statistically significant ($z = 1.47$, signed rank = 144, $p = 0.93$, $r = 0.33$, figure s3).

We also explored if iconic and beat gestures may have differential impacts of either frequency-dependent or surprisal-dependent N400. This comparison, however, needs to be treated with caution given the much lower numbers of data points for each gesture type. We evaluated both gesture type’s potential impacts also in two steps. Firstly, for model comparison, for both frequency-only and surprisal-only models separately, we included a binary regressor regarding the presence of either an iconic or a beat gesture, and evaluated if the inclusion of iconic/beat gesture regressor improves the model fit, and if they differ from each other. For the frequency model, results showed that including coding for both types of gestures generally improves model fit (iconic: $t = 3.79$, $p = 0.02$, beat: $t = 2.29$, $p = 0.024$). However, no difference between both types of gestures were observed ($t = 1.57$, $p = 0.137$). For surprisal, including coding for both types of gestures also improved model fit (iconic: $t = 2.29$, $p = 0.038$, beat: $t = 1.74$, $p = 0.015$). We also observed no difference between both types of gestures ($t = 1.75$, $p = 0.096$).

To investigate the effect of gesture types on frequency-derived N400s, we estimated two separate models for frequency, based on words that were marked as gesture present with iconic gestures and beat gestures respectively. We then analyzed their corresponding mTRFs to gesture present and gesture absent content words for each gesture type. Results (figure s5, supplementary materials) indicated that N400s are similarly modulated by both gesture types, as both iconic gesture absent words (median = -5.37, SD = 3.96) and beat gesture absent words (median = -5.37, SD = 3.96) are associated with more negative beta coefficients than their gesture present counterparts (iconic: median = 7.03, SD = 13.33; beat: median = -0.12, SD = 15.30). Both effects were statistically significant (iconic: $z = -2.93$, signed rank = 26, $p = 0.002$, $r = 0.66$; beat: $z = -2.40$, signed rank = 40, p value = 0.008, $r = 0.54$, figure s4). We repeated this analysis for surprisal derived N400s. For surprisal (figure s6, supplementary materials), only beat gestures show a statistical effect ($z = -2.632$, signed rank = 34, $r = 0.589$) where gesture absent words (median = -2.43, SD = 3.96) are statistically more negative than gesture present words (median = 4.50, SD = 13.24). In contrast, iconic gestures do not have a statistically significant effect on the surprisal derived N400 amplitude ($z = -0.690$, signed rank = 86, $p = 0.245$, $r = 0.154$).

Discussion

We investigated whether the presence of gestures modulates the cognitive demands associated with lexical retrieval and semantic prediction during the processing of naturalistic multimodal stimuli. With a set of mTRF analyses, we found that providing the gesture coding significantly improves the mTFR model fit. Most importantly, extending a prior study (Sassenhagen, 2019), across all models, we observed robust evidence that co-speech gestures reduced the amplitude of the frequency-dependent N400, thus suggesting a facilitative role of gestures on lexical retrieval. This finding significantly elaborates the semantic processing literature of multimodal language: prior studies typically employ classic semantic violation paradigms (He et al., 2020; Morett et al., 2020; Wang & Chu, 2013), or disambiguation paradigms (Holle & Gunter, 2007) to derive the N400 effect. Consequently, although results from these studies speak strongly for a facilitative role of gesture, it is, more specifically, semantic *integration* that benefits from the visual modality. Here, building on the well-established link between lexical frequency and the N400, we showed that bottom-up lexical retrieval, as indexed by the N400, benefits from a complementary visual modality, just as how it interacts with sentence context (Van Petten & Kutas, 1990). Our results also align with an extensive line of literature on the facilitative effect of gesture on lexical retrieval during production (Hadar et al., 1998; Hadar & Butterworth, 1997; Lanyon & Rose, 2009). To our knowledge, this is the first time that an interaction between frequency and co-speech gestures was observed, although directly from a naturalistic paradigm.

Regarding semantic prediction, GPT-2 surprisal indeed models an apparent N400 during the processing of a naturalistic multimodal narrative. This is in line with studies showing a surprisal N400 effect using auditory-only stimuli (Frank et al., 2015; Frank & Willems, 2017; Yan & Jaeger, 2020) and corroborate a similar effect of surprisal during the processing of multimodal stimuli (Y. Zhang et al., 2021). We extended these findings by using speech stimuli from a long and continuous multimodal narrative rather than single sentences, thus establishing the modulatory effect of co-speech gestures on the N400 during naturalistic comprehension. Similar to more recent studies, we also implemented a time-resolved version of regression-based ERPs, namely multivariate Temporal Response Functions, which allowed us to unmix the ERP responses to words of interest from that to preceding and succeeding words in the continuous EEG signal (Crosse et al., 2016). With this in mind, our findings imply that surprisal can quantify a word's unpredictability not only at the level of single lexical items or sentences (Hale, 2001; Levy, 2008), but also during word-by-word processing of naturalistic narratives that are multimodal in nature.

However notably, our analyses in the additive and interactive models only showed a significant effect of gesture on frequency-dependent N400, but not on the surprisal-dependent N400. This finding may be considered evidence of the modulation effect of gestures on surprisal being potentially dependent on frequency. In the literature on unimodal (visual or auditory) naturalistic language processing, word frequency is commonly, but not always input as a covariate when modelling the effect of context-dependent measures such as surprisal or semantic similarity (Armeni et al., 2019; Weissbart et al., 2019; Willems et al., 2016, but see Broderick et al., 2018). In the multimodal language processing literature, to date, the effect of gestures on semantic prediction (as indexed by surprisal) has been investigated in only one recent study (Y. Zhang et al., 2021), but there the potential effect of word frequency was not controlled for. In our study, the null effect for surprisal in the *frequency + surprisal* model does not necessarily imply no effect of gestures on semantic prediction: for example, the absence of a significant effect of surprisal in the additive model may be alternatively explained by the fact that frequency captures the shared variance between frequency and surprisal; and after all, in the single-predictor model, we still observed significant effect of gesture on the surprisal-N400. Thus, although our results are in accordance with prior studies in suggesting a modulatory effect of gestures on the N400s (which may reflect semantic processing in general), they would need further validation, especially regarding how lexical retrieval, semantic prediction, and semantic integration are interactively affected by gestures.

An important contribution of the current study is that the effect of gestures on the frequency- and/or surprisal-dependent N400 is observed in a multimodal narrative (Willems et al., 2020). This extends previous results from factorial studies (Fabbri-Destro et al., 2015; Kelly et al., 2004; Özyürek et al., 2007; Willems et al., 2007; Wu & Coulson, 2005) to the processing of more naturalistic stimuli. A possible reason for this facilitative effect is that gesture onset in the current experiment preceded the onset of critical words, preactivating semantic representations or facilitating lexical retrieval, and thus alleviating the cognitive demand associated to the decoding of meaning (He et al., 2020; Maess et al., 2016; Szewczyk & Schriefers, 2018; ter Bekke et al., 2020; Y. Zhang et al., 2021) . This effect could be especially highlighted in a more naturalistic setting where the actor of the video was able to freely produce the spontaneous gestures during recording. Thus, it could be that there is a bias to use gestures for words that are contextually more salient (Pouw et al., 2021; Trujillo et al., 2021), or that gestures co-occur with enhanced articulation, or even with a slow-down of the speaking rate, both of which could potentially facilitate semantic processing (e.g. Broderick et al., 2018). Consequently, when perceived by the comprehender, the onset of a gesture would automatically signal the ease of processing of the upcoming key word and would jointly reduce the effort of lexical retrieval and/or semantic prediction *together* with more enhanced articulation. However, as we did not control for these factors (e.g., artificially cover the actor’s mouth) for the purpose of maintaining maximal naturalness of the stimuli, these potential confounds would need to be further assessed in more controlled conditions.

Alternatively, from a neurobiological perspective, it has been suggested that a left-lateralized, modality-independent system exists in anterior and posterior temporal regions that maps semantic information into common conceptual representations (Andric et al., 2013; Straube et al., 2012, 2013; Xu et al., 2009, but see Jouravlev et al., 2019 for an alternative view). Therefore, it could also be possible that the facilitative effect of gestures during lexicosemantic retrieval/prediction indexed by the attenuation of the frequency and surprisal N400 responses reflects the integration of matching acoustic and visual symbolic representations encoded by multisensory neuronal populations in this supramodal semantic system. This would be similar to what has been documented for low-level perceptual features during audiovisual speech perception (Park et al., 2018) and to additive and supra-additive effects during multisensory integration (Stein & Stanford, 2008). Both scenarios, however, highlight the relevance of speech gestures in human communication and raise interesting questions about the evolutionary origins and relevance of their facilitative effect.

Our study has a number of limitations. First, for practical considerations, we did not set up an auditory-only condition that could potentially serve as unimodal baseline. Although this approach is used by a number of recent naturalistic language processing studies (S. Zhang et al., 2022), it is nevertheless vulnerable to the lack of control of stimulus-related features between the *Gesture present* and *Gesture absent* conditions. Secondly, in the current study, even if we have compared iconic and beat gestures on their effects on frequency and surprisal, and have found potentially differential effects on the surprisal-dependent N400s (see figure s5, supplement), the interpretation of this set of analysis should be cautious because of the low number of datapoints available in the current study. Here, we found that both iconic and beat gestures consistently modulate the frequency-dependent N400; but for surprisal, although the mTRF model fit generally improved when both iconic and beat gesture coding were additionally included as regressors, the pairwise comparison between gesture present and gesture absent words only showed a significant modulation effect of beat gesture on the surprisal-dependent N400. Other recent literature has provided evidence of dissociable effects of gesture types on electrophysiological responses using both naturalistic and factorial approaches, with iconic (i.e., meaningful) gestures being associated to a reduction in the amplitude of the N400 and beat gestures being associated to increased N400 effects as derived from semantic violation or surprisal (Hintz et al., 2022; Y. Zhang et al., 2021). On the other hand, there is another line of literature showing that the N400 amplitude of single words in a sentence may still be modulated by the presence of beat gestures, suggesting a potentially facilitative effect of beat gestures (Morett et al., 2020; Wang & Chu, 2013). Clearly, given the mixture of current literature, future experiments are necessary to shed light on this issue. Moreover, in this study we used GPT-2 for analyzing surprisal. While it was informative, subsequent large language models may outperform GPT-2 due to their expanded training datasets and refined architectures, and may provide better fit to the EEG data (Digutsch & Kosinski, 2023; Mahowald et al., 2023; Michaelov et al., 2023). Further, in the current study we interpolated bad electrodes based on a relatively sparse 32-channel system, this potential distortion may be best handled with a conceptual replication with an independent dataset. Lastly, like many other studies that investigate semantic processing using narratives, we did not employ any behavioral tasks during the experiential procedure (e.g., Broderick et al., 2018; Goldstein et al., 2022; Willems et al., 2016). As a result, the interpretation of any degree of facilitation, either of lexical retrieval or semantic prediction, can be interpreted on a neural level at best. For this reason, a naturalistic paradigm that most optimally combines behavior and its neural substrates becomes imperative for further research (Gratton et al., 2022).

Acknowledgments

This project was funded by the Deutsche Forschungsgemeinschaft (DFG), funding number HE8029/2-1, the von Behring-Röntgen-Stiftung (funding number 59-0002, 64-0001), and the Excellence Program 'The Adaptive Mind' of the Hessian Ministry of Higher Education. SO received funding from Agencia Nacional de Investigación y Desarollo (ANID), national grant for doctoral studies N° 21181786.

Competing interests

The authors declare no competing interests.

Data availability

Data will be made available upon request by contacting YH at yifei.he@staff.uni-marburg.de.

Author contributions

SO implemented EEG preprocessing and data analysis scripts, implemented multivariate regression analyses, created figures, interpreted results, and wrote and edited the manuscript. BS designed the experiment, acquired funding, and reviewed the manuscript. LM implemented multivariate regression analyses, GPT2 lexical-surprisal analyses, interpreted results, wrote, reviewed, and edited the manuscript. YH designed the experiment, acquired data, created figures, interpreted results, and wrote, reviewed, and edited the manuscript.

References

Alday, P. M., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2017). Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: Event- related potentials reflect continuous model updates. *ENeuro*, *4*(6). https://doi.org/10.1523/ENEURO.0311-16.2017

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture*, *10*(1), 3–28. https://doi.org/10.1075/gest.10.1.02ali

Andric, M., Solodkin, A., Buccino, G., Goldin-meadow, S., Rizzolatti, G., & Small, S. L. (2013). Brain function overlaps when people observe emblems, speech, and grasping. *Neuropsychologia*, *51*(8), 1619–1629. https://doi.org/10.1016/j.neuropsychologia.2013.03.022

Armeni, K., Willems, R. M., van den Bosch, A., & Schoffelen, J. M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, *198*(September 2018), 283–295. https://doi.org/10.1016/j.neuroimage.2019.04.083

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1943). https://doi.org/10.1098/rspb.2020.2419

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803-809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Castellanos, N. P., & Makarov, V. A. (2006). Recovering EEG brain signals: Artifact suppression with wavelet enhanced independent component analysis. *Journal of Neuroscience Methods*, *158*(2), 300–312. https://doi.org/10.1016/j.jneumeth.2006.05.033

Cho, H. C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*(9), 1261–1266. https://doi.org/10.1016/j.jbusres.2012.02.023

Crosse, M. J., Liberto, G. M. Di, Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) toolbox : A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*(November), 1–14. https://doi.org/10.3389/fnhum.2016.00604

Cuevas, P., Steines, M., He, Y., Nagels, A., Culham, J., & Straube, B. (2019). The facilitative effect of gestures on the neural processing of semantic complexity in a continuous narrative. *NeuroImage*, *195*(March), 38–47. https://doi.org/10.1016/j.neuroimage.2019.03.054

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103. https://doi.org/10.1016/j.brainres.2006.02.010

de Cheveigné, A. (2019). ZapLine: a simple and effective method to remove power line artifacts. *NeuroImage*, *1*(1), 1–13. https://doi.org/http://dx.doi.org/10.1101/782029

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

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, *34*(4), 1443–1449. https://doi.org/10.1016/j.neuroimage.2006.11.004

Digutsch, J., & Kosinski, M. (2023). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Scientific Reports*, *13*(1), 1–7. https://doi.org/10.1038/s41598-023-32248-6

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(January), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Ehinger, B. V., & Dimigen, O. (2019). Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ*, *2019*(10), 1–33. https://doi.org/10.7717/peerj.7838

Fabbri-Destro, M., Avanzini, P., De Stefani, E., Innocenti, A., Campi, C., & Gentilucci, M. (2015). Interaction between words and symbolic gestures as revealed by N400. *Brain Topography*, *28*, 591–605. https://doi.org/10.1007/s10548-014-0392-4

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

Gabard-Durnam, L. J., Leal, A. S. M., Wilkinson, C. L., & Levin, A. R. (2018). The harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data. *Frontiers in Neuroscience*, *12*, 97. https://doi.org/10.3389/fnins.2018.00097

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 759–765.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, 10–18. https://doi.org/10.18653/v1/w18-0102

Gratton, C., Nelson, S. M., & Gordon, E. M. (2022). Brain-behavior correlations: Two paths toward reliability. *Neuron*, *110*(9), 1446–1449. https://doi.org/10.1016/j.neuron.2022.04.018

Hadar, U., & Butterworth, B. (1997). Iconic gesture, imagery and word retrieval in speech. *Semiotica*, *115*, 147–172.

Hadar, U., Wenkert-Olenik, D., Krauss, R., & Soroker, N. (1998). Gesture and the processing of speech: Neuropsychological evidence. *Brain and Language*, *62*(1), 107–126.

Hale, J. T. (2001). A probabilistic earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. https://doi.org/10.3115/1073336.1073357

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage*, *17*(3), 1101–1116. https://doi.org/10.1006/nimg.2002.1268

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27–42. https://doi.org/10.1016/j.neuropsychologia.2015.04.018

He, Y., Luell, S., Muralikrishnan, R., Straube, B., & Nagels, A. (2020). Gesture's body orientation modulates the N400 for visual sentences primed by gestures. *Human Brain Mapping*, *41*(17), 4901–4911. https://doi.org/10.1002/hbm.25166

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(32), 1–12. https://doi.org/10.1073/pnas.2201968119

Hintz, F., Khoe, Y. H., Strauß, A., Psomakas, A., & Holler, J. (2022). Electrophysiological evidence for the enhancement of gesture-speech integration by linguistic predictability during multimodal discourse comprehension. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/avudx

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–1192. https://doi.org/10.1162/jocn.2007.19.7.1175

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in Psychology*, *3*(March), 1–12. https://doi.org/10.3389/fpsyg.2012.00074

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

Huizeling, E., Arana, S., Hagoort, P., & Schoffelen, J. M. (2022). Lexical Frequency and Sentence Context Influence the Brain's Response to Single Words. *Neurobiology of Language*, *3*(1), 149–179. https://doi.org/10.1162/nol_a_00054

Jouravlev, O., Zheng, D., Balewski, Z., Le Arnz Pongos, A., Levan, Z., Goldin-Meadow, S., & Fedorenko, E. (2019). Speech-accompanying gestures are not processed by the language-processing mechanisms. *Neuropsychologia*, *132*(August 2018), 107132. https://doi.org/10.1016/j.neuropsychologia.2019.107132

Kandana Arachchige, K. G., Simoes Loureiro, I., Blekic, W., Rossignol, M., & Lefebvre, L. (2021). The Role of Iconic Gestures in Speech Comprehension: An Overview of Various Methodologies. *Frontiers in Psychology*, *12*(April), 1–15.

https://doi.org/10.3389/fpsyg.2021.634074

Kandylaki, K. D., & Bornkessel-Schlesewsky, I. (2019). From story comprehension to the neurobiology of language. *Language, Cognition and Neuroscience*, *34*(4), 405–410. https://doi.org/10.1080/23273798.2019.1584679

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253–260. https://doi.org/10.1016/S0093-934X(03)00335-3

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(6), 1648–1662. https://doi.org/10.1037/xlm0000128

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*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Lanyon, L., & Rose, M. L. (2009). Do the hands have it? The facilitation effects of arm and hand gesture on word retrieval in aphasia. *Aphasiology*, *23*(7–8), 809–822.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:

(De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Maess, B., Mamashli, F., Obleser, J., Helle, L., & Friederici, A. D. (2016). Prediction signatures in the brain: Semantic pre-activation during language comprehension. *Frontiers in Human Neuroscience*, *10*(November), 1–11. https://doi.org/10.3389/fnhum.2016.00591

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *ArXiv Preprint ArXiv:2301.06627*.

McNeill, D. (2008). *Gesture and thought*. Chicago University Press.

Meyer, L., Sun, Y., & Martin, A. E. (2020). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, *35*(9), 1089–1099. https://doi.org/10.1080/23273798.2019.1693050

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 1–29. https://doi.org/10.1162/nol_a_00105

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229–240. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, 398–408.

Morett, L. M., Landi, N., Irwin, J., & McPartland, J. C. (2020). N400 amplitude, latency, and variability reflect temporal integration of beat gesture and pitch accent during language processing. *Brain Research*, *1747*(August), 147059. https://doi.org/10.1016/j.brainres.2020.147059

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a largescale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791). https://doi.org/10.1098/rstb.2018.0522

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651). https://doi.org/10.1098/rstb.2013.0296

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), 605–616. https://doi.org/10.1162/jocn.2007.19.4.605

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS Biology*, *16*(8), 1–26.

https://doi.org/10.1371/journal.pbio.2006558

Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, *52*(11), 1456–1469.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197.

Pouw, W., Wit, J. D., Bögels, S., Rasenberg, M., Milivojevic, B., & Ozyurek, A. (2021). Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics. *International Conference on Human-Computer Interaction*, 269–287.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8). http://arxiv.org/abs/2007.07582

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201. https://doi.org/10.3758/BF03197692

Ryu, S. H., & Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. *CMCL 2021 - Workshop on Cognitive Modeling and Computational Linguistics, Proceedings*, 61–71. https://doi.org/10.18653/v1/2021.cmcl-1.6

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

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45). https://doi.org/10.1073/pnas.2105646118

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255–266. https://doi.org/10.1038/nrn2331

Straube, B., Green, A., Weis, S., & Kircher, T. (2012). A Supramodal Neural Network for Speech and Gesture Semantics: An fMRI Study. *PLoS ONE*, *7*(11). https://doi.org/10.1371/journal.pone.0051207

Straube, B., He, Y., Steines, M., Gebhardt, H., Kircher, T., Sammer, G., & Nagels, A. (2013). Supramodal neural processing of abstract information conveyed by speech and gesture. *Frontiers in Behavioral Neuroscience*, *7*(September), 1–14. https://doi.org/10.3389/fnbeh.2013.00120

Sun, J., Wang, Z., & Tian, X. (2021). Manual gestures modulate early neural responses in loudness perception. *Frontiers in Neuroscience*, *15*(September), 1–17. https://doi.org/10.3389/fnins.2021.634967

Szewczyk, J. M., & Schriefers, H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience*, *33*(6), 665–686. https://doi.org/10.1080/23273798.2017.1401101 ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. *PsyArXiv*. https://doi.org/10.31234/osf.io/b5zq7

Trujillo, J., Özyürek, A., Holler, J., & Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect in noise. *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-95791-0

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequencyinevent-related brainpotentials. *Memory & Cognition*, *18*(4), 380–393. https://doi.org/10.3758/BF03197127

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: an ERP study. *Neuropsychologia*, *51*(13), 2847–2855. https://doi.org/10.1016/j.neuropsychologia.2013.09.027

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2019). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. https://doi.org/10.1162/jocn_a_01467

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van Den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Willems, R. M., Nastase, S. A., & Milivojevic, B. (2020). Narratives for neuroscience. *Trends in Neurosciences*, *43*(5), 271–273. https://doi.org/10.1016/j.tins.2020.03.003

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*(10), 2322–2333.

https://doi.org/10.1093/cercor/bhl141

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, *7*(1), 30. https://doi.org/10.1186/1744-9081-7-30

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

Xu, J., Gannon, P. J., Emmorey, K., Smith, J. F., & Braun, A. R. (2009). Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(49), 20664–20669. https://doi.org/10.1073/pnas.0909197106

Yan, S., & Jaeger, T. F. (2020). (Early) context effects on event-related potentials over natural inputs. *Language, Cognition and Neuroscience*, *35*(5), 658–679. https://doi.org/10.1080/23273798.2019.1597979

Zhang, S., Jixing, L., Yang, Y., & Hale, J. (2022). Decoding the silence: Neural bases of zero pronoun resolution in Chinese. *Brain and Language*, *224*(November 2021), 105050. https://doi.org/10.1016/j.bandl.2021.105050

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). https://doi.org/10.1098/rspb.2021.0500

Table 1. Summary of gesture occurrences

Gesture Type	Occurrence	%
Iconic	181	36.71
Metaphoric	85	17.24
Emblematic	19	3.85
Beat	148	30.02
Deictic	60	12.17
Total	493	100

Table 2. Summary of word statistics

	N° of Words	Word duration				GPT-2 Surprisal				Word frequency			
		Mean	Median	SD	Range	Mean	Median	SD	Range	Mean	Median	SD	Range
Gesture Present	466	0.545	0.510	0.199	1.080	13.714	12.679	7.331	47.010	12.695	11	5.628	27
Gesture Absent	1491	0.491	0.460	0.195	2.810	13.056	11.538	7.932	53.972	11.674	11	5.586	27

Table 3. Model fit comparison

Model	**Fit (spearman r)**			
	Without gesture as regressor		*With gesture as regressor*	
	Mean	SD	Mean	SD
Surprisal	0.058	0.013	0.069	0.016
Frequency	0.056	0.015	0.069	0.013
Additive	0.066	0.017	**0.076**	0.015
Interaction	0.065	0.015	0.075	0.016

Figures

Figure 1

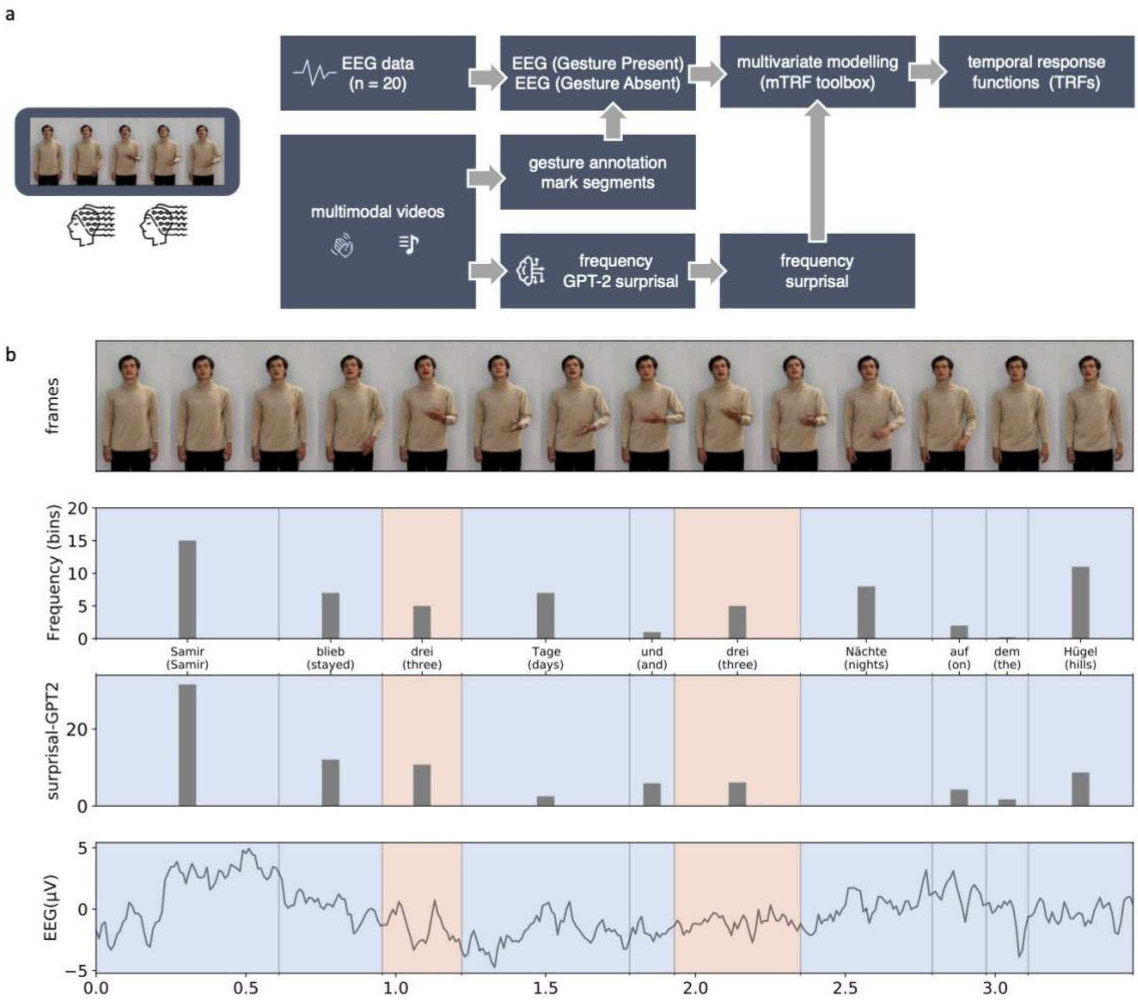


Figure 2

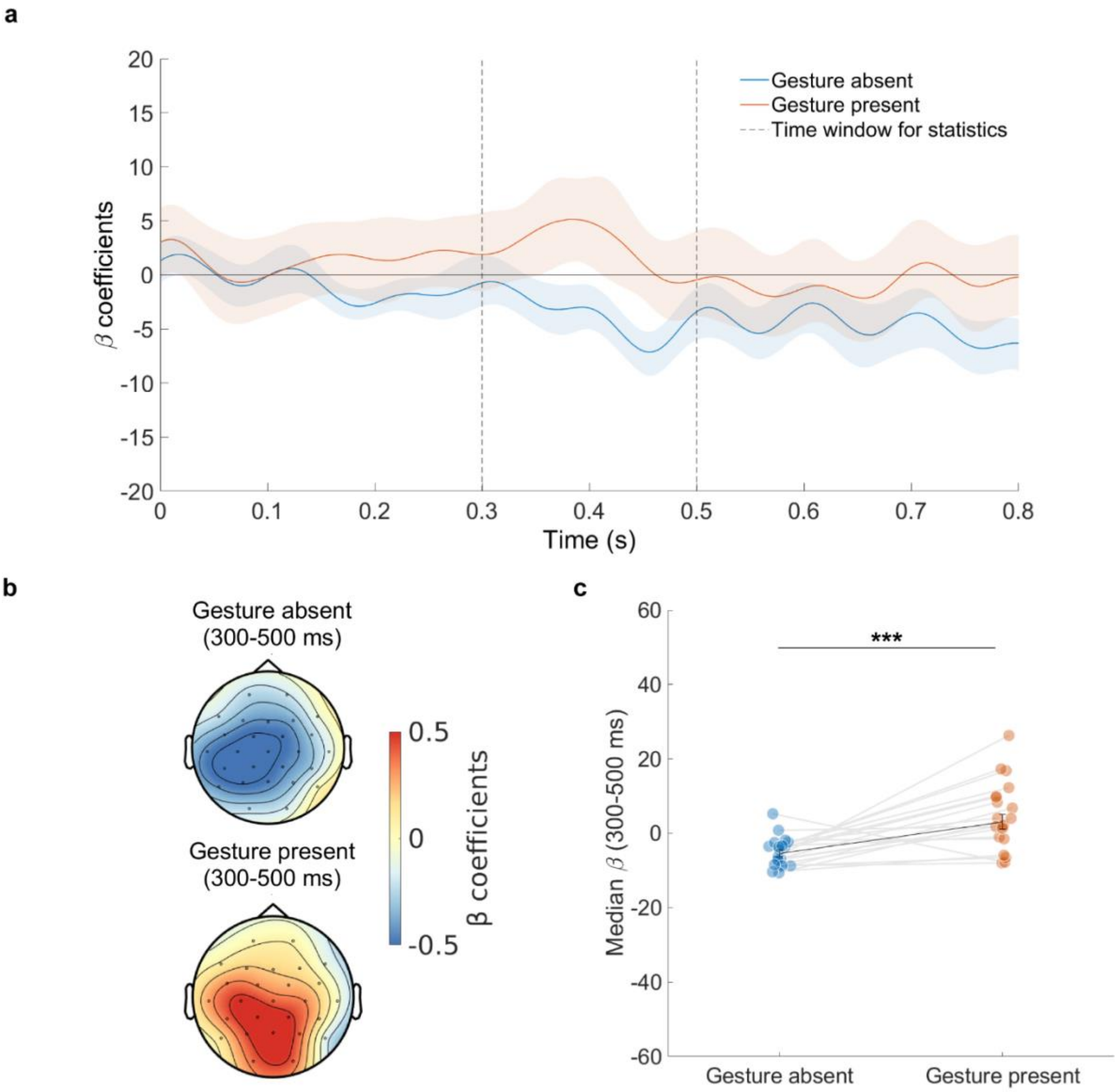
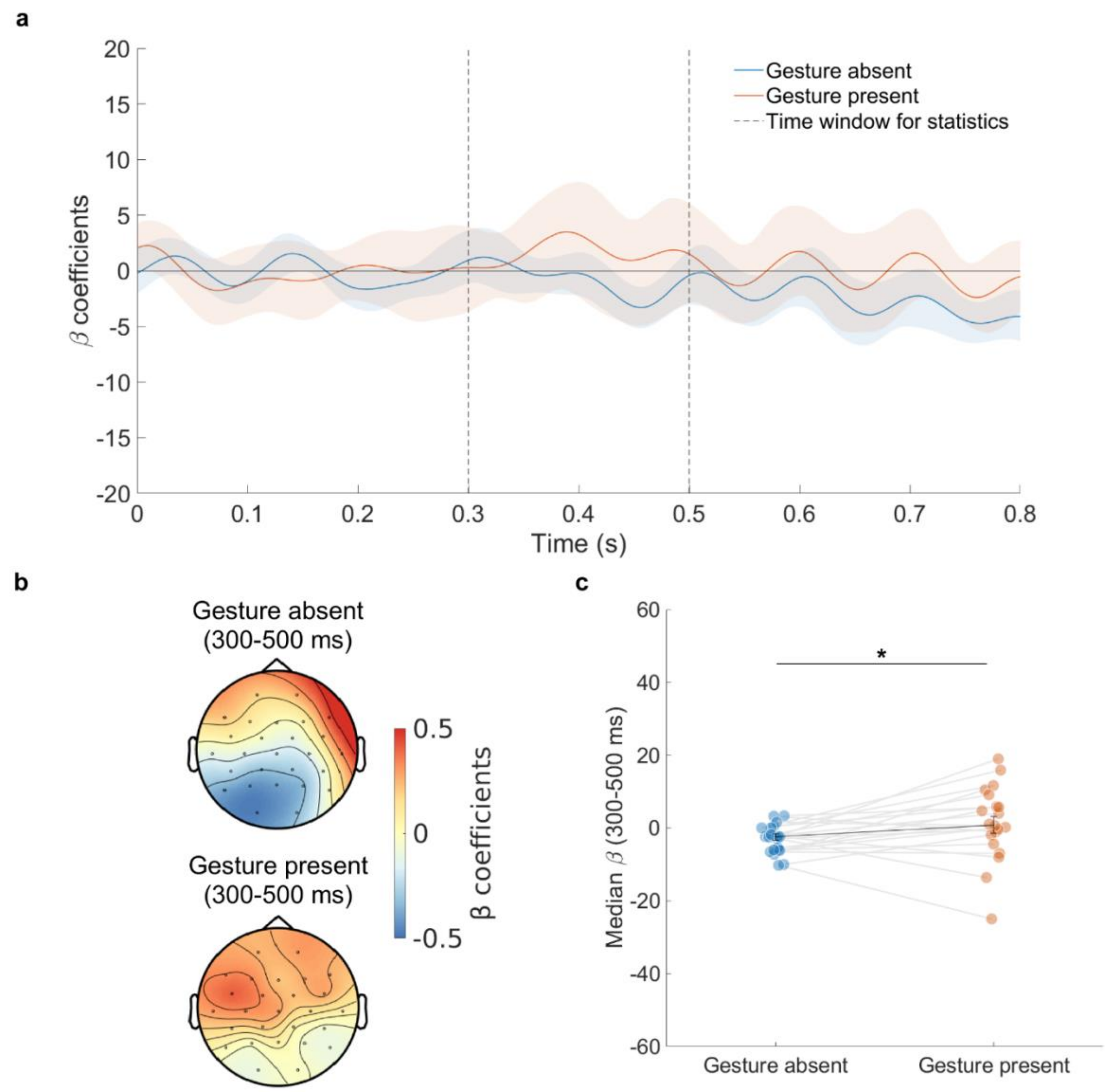
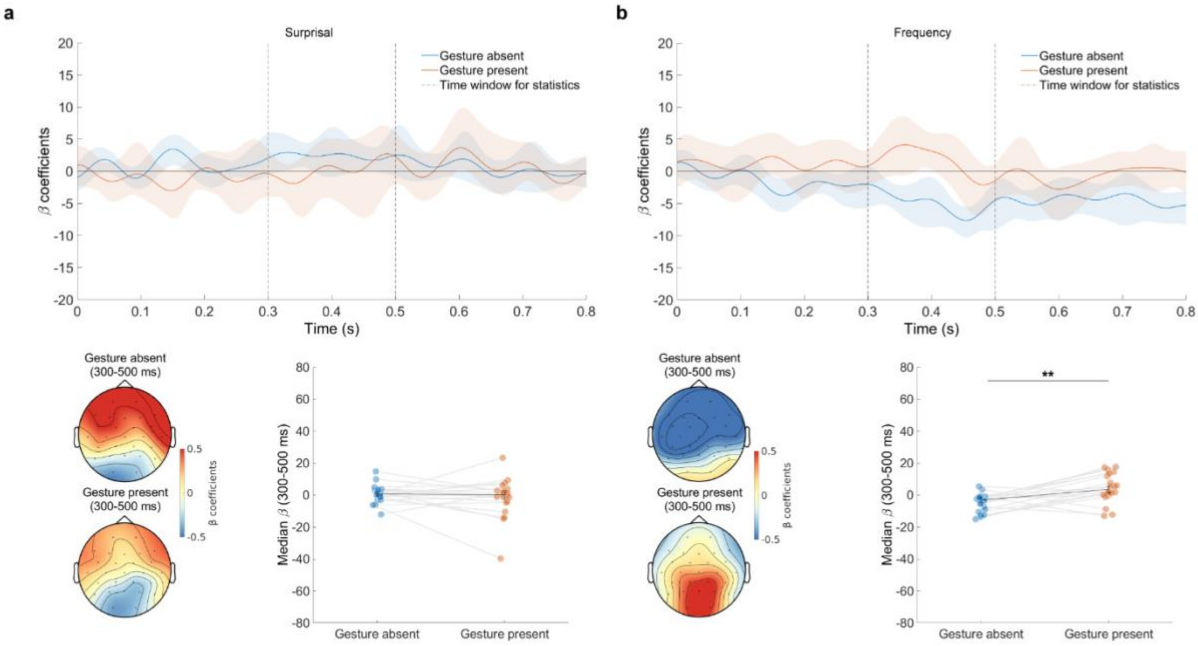

a
Gesture absent
Gesture present
Time window for statistics
β coefficients
Time (s)
b
Gesture absent
(300-500 ms)
Gesture present
(300-500 ms)
β coefficients
c
***
Median β (300-500 ms)
Gesture absent
Gesture present

Figure 3


a
Gesture absent
Gesture present
Time window for statistics
β coefficients
Time (s)
b
Gesture absent
(300-500 ms)
Gesture present
(300-500 ms)
β coefficients
c
Median β (300-500 ms)
*
Gesture absent
Gesture present

964 ***Figure 4***



965

966

967

968

969

970

971

972

973

974

975

Figure captions

- **Figure 1.** Schematic representation of experimental and analytic protocol. **a**. EEG data was acquired while participants were exposed to multimodal naturalistic speech stimuli. Besides lexical frequency, word-by-word surprisal was obtained using the GPT-2 model. Multimodal stimuli were annotated for the presence of absence of gestures. EEG data were then modelled using the mTRF toolbox to obtain regression-based ERPs (i.e., mTRFs) for gesture present and gesture absent data separately. **b**. Sample frames from the multimodal video stimuli (top) along with corresponding word-by-word frequency and GPT-2-derived surprisal values (middle) and the average EEG signal for a group of centro-parietal electrodes (bottom). Words marked as *Gesture present* are marked with red background.
- **Figure 2.** Results of TRF analyses for lexical frequency. **a**. Mean mTRF to *Gesture absent* (blue) and *Gesture present* (orange) words obtained from the average of our electrodes of interest. Coloured shades show the 95% bootstrapped confidence intervals. Dotted lines represent our time-window of interest for statistical analyses (300–500 ms). **b**. Topographical distributions of mean mTRF responses between 300–500 ms. **c**. Individual and group median mTRF beta values between 300–500 ms. The line and asterisks represent a statistically significant effect ($p < 0.001$, one-tailed).
- **Figure 3.** Results of TRF analyses for GPT-2 surprisal. **a**. Mean mTRF to *Gesture absent* (blue) and *Gesture present* (orange) words obtained from the average of a group of centro-parietal electrodes. Coloured shades show the 95% bootstrapped confidence intervals. Dotted lines represent our time-window of interest for statistical analyses (300–500 ms). **b**. Topographical distributions of mean mTRF responses between 300–500 ms. **c**. Individual and group median mTRF beta values between

300–500 ms. The line and asterisk represent a statistically significant effect ($p < 0.05$, one-tailed).

- **Figure 4.** Results of TRF analyses for the additive model **a**. Effect of GPT-2 surprisal. **b**. Effect of lexical frequency. For both panels, mTRF for *Gesture absent* (blue) and *Gesture present* (orange) words obtained as the average of an *a priori* defined group of centro-parietal electrodes. Coloured shades show the 95% bootstrapped confidence intervals. Dotted lines represent our time-window of interest for statistical analyses (300–500 ms). Topographical distributions are based on the mean mTRF responses between 300–500 ms. Point and line-plots show Individual and group median mTRF beta values between 300–500 ms. The line and asterisks represent a statistically significant effect ($p < 0.01$, one-tailed).

Supplementary materials

The role of co-speech gestures in retrieval and prediction during naturalistic multimodal narrative processing

This PDF includes:

Tables s1 and s2

Figures s1, s2, s3, s4, s5, s6

Table s1. Optimal Lamba values for additive and interaction models after cross-validation

Subject	**Additive**		**Interaction**	
	Gesture absent	*Gesture present*	*Gesture absent*	*Gesture present*
1	0.631	0.040	0.794	0.200
2	0.794	10.000	1.995	10.000
3	0.010	0.794	0.158	0.010
4	0.010	0.200	0.316	0.251
5	0.010	0.010	1.585	0.631
6	0.398	2.512	1.000	0.010
7	1.000	1.995	1.259	0.794
8	6.310	1.259	5.012	0.794
9	0.010	0.501	1.000	0.501
10	0.501	0.010	0.794	0.010
11	0.794	0.631	0.398	0.200
12	0.010	10.000	0.010	1.000
13	0.010	1.259	0.010	0.010
14	7.943	0.032	2.512	0.079
15	2.512	0.010	0.010	0.251
16	0.010	0.398	0.251	0.200
17	0.010	1.995	0.631	0.010
18	0.200	0.010	0.794	0.020
19	1.000	0.010	1.259	0.032
20	0.010	10.000	0.794	10.000

Table s2. Model fit values (spearman r) for the four models

Subject	**Surprisal**		**Frequency**		**Additive**		**Interaction**	
	Gesture absent	*Gesture present*	*Gesture absent*	*Gesture present*	*Gesture absent*	*Gesture present*	*Gesture absent*	*Gesture present*
1	0,088	-0,013	0,103	0,042	0,030	0,048	0,108	0,037
2	0,037	0,007	0,051	0,031	0,044	0,041	0,039	0,056
3	0,032	0,047	0,028	0,048	0,089	0,039	0,103	0,046
4	0,073	0,053	0,059	0,061	0,082	0,053	0,033	0,048
5	0,024	0,042	0,047	0,020	0,026	0,051	0,037	0,053
6	0,078	0,054	0,116	0,033	0,042	0,083	0,084	0,062
7	0,045	0,040	0,034	0,036	0,066	0,044	0,053	0,010
8	0,055	0,099	0,086	0,043	0,026	0,072	0,028	0,035
9	0,047	0,039	0,092	0,055	0,042	0,074	0,066	0,056
10	0,066	0,039	0,028	0,037	0,042	0,023	0,026	0,070
11	0,050	0,010	0,070	0,059	0,036	0,087	0,050	0,003
12	0,049	0,020	0,078	0,035	0,033	0,021	0,029	0,036
13	0,046	-0,008	0,064	0,054	0,036	0,004	0,093	0,048
14	0,046	0,064	0,068	0,045	0,100	0,070	0,100	0,069
15	0,057	0,019	0,039	0,050	0,075	0,028	0,065	0,102
16	0,123	0,092	0,090	0,100	0,089	0,078	0,084	0,082
17	0,041	0,032	0,033	0,045	0,135	0,039	0,110	0,067
18	0,067	0,035	0,021	0,054	0,086	0,046	0,046	0,045
19	0,022	0,052	0,072	0,060	0,039	0,075	0,074	0,066
20	0,075	0,028	0,043	0,004	0,080	0,021	0,050	0,033

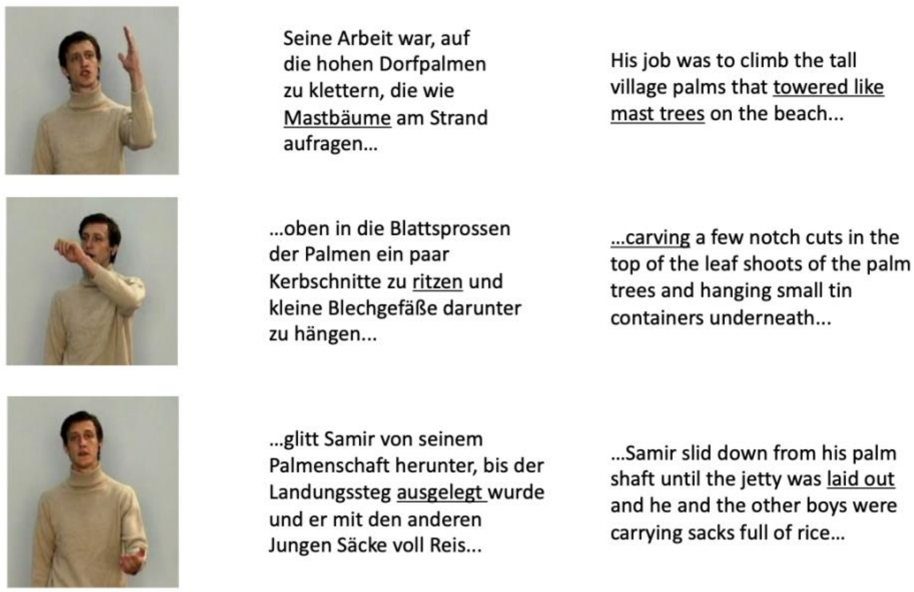


Figure s1. Example of the gestures as produced in the multimodal narrative (left panel). The middle panel shows the corresponding original German sentence and the *Gesture present* words (underlined). The right panel shows the English translation of the sentences and the *Gesture present* words.

Conventional ERP effects for high- vs. low-frequency words

We conducted a median-split analysis based on lexical frequency for all content words and compared the conventional ERP effects between high– and low-frequency words for (1) all words and (2) words within the *Gesture present* and the *Gesture absent* conditions. ERPs were all baseline-corrected based on the -200-0 ms time-locked to the onset of each word. We plot the ERP waveforms in figure s1. Analogous to the mTRF results, we directly extracted the ERP amplitudes between 300–500ms for centro-parietal electrodes and entered these values into a 2 x 2 parametric repeated-measures ANOVA with the factors FREQUENCY (high vs. low) and GESTURE (present vs. absent), as interaction is unable to be tested via non-parametric tests. ANOVA revealed a main effect of GESTURE ($F_{(1,19)}$= 41.90, $p < 0.00003$) and no main effect of frequency ($F_{(1,19)} = 1.04$, $p = 0.32$). The interaction between the two factors was significant ($F_{(1,19)}$= 5.92, $p < 0.00003$). We then conducted Wilcoxon signed-rank tests for the effect of frequency within *Gesture present* and *Gesture absent*. No effect of frequency was observed in the *Gesture absent* condition ($z = 2.25$, signed rank = 165, $p = 0.99$), but low-frequency content words (above median) showed more positive N400 amplitudes in comparison to high-frequency words ($z = –2.14$, signed rank = 47, $p = 0.015$) in the *Gesture present* condition. This effect, besides being significant, is not in accordance with the hypothesized pattern that low-frequency words should be more negative in the N400 window. Overall, the hypothesized effect of frequency was not reliably observed in the ERP results.

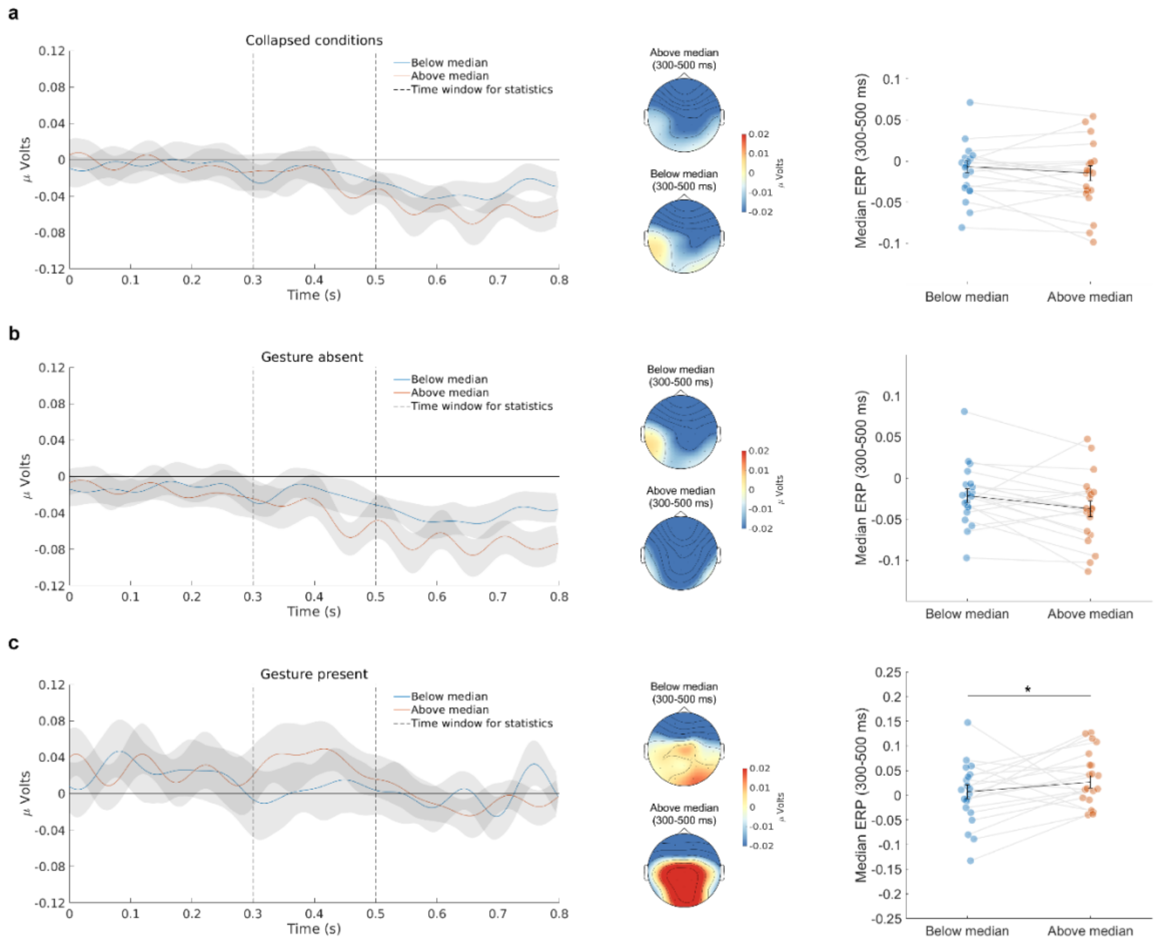


Figure s2. Results of the ERP analyses (median split) for lexical frequency for **a**. All words collapsed across *Gesture absent* and *Gesture present* conditions. **b**. Words in the *Gesture absent* condition. and **c**. Words in the *Gesture present* condition. For all panels, waveforms of low-frequency words (Above median) are depicted in red, and high-frequency words (Below median) are depicted in blue. ERPs were obtained from the average of all electrodes. Coloured shades show the 95% bootstrapped confidence intervals. Dotted lines represent our time-window of interest for statistical analyses (300-500 ms). Topographical distributions show ERP responses between 300-500 ms. Point- and line-plots show individual and group median ERP amplitudes between 300-500 ms. The line and asterisk represent a statistically significant effect ($p < 0.05$, one-tailed).

Conventional ERP effects for high- vs. low-surprisal words

For surprisal, we also conducted a median-split ERP analysis based on GPT-2 surprisal for all content words and compared the conventional ERP effects between high- and low-surprisal words for 1) all words and 2) words within the *Gesture present* and the *Gesture absent* conditions. ERPs were all baseline-corrected based on the -200-0 ms time-locked to the onset of each word. We plot the ERP waveforms in Figure s2. Analogous to the mTRF results, we directly extracted the ERP amplitudes between 300-500ms for centro-parietal electrodes and entered these values into a 2 x 2 parametric repeated-measures ANOVA with the factors SURPRISAL (high vs. low) and GESTURE (present vs. absent). ANOVA revealed a main effect of GESTURE ($F_{(1,19)} = 52.212$, $p = 7.34e-07$), but no main effect of SURPRISAL ($F_{(1,19)} = 0.078$, $p = 0.79$). The interaction between the factors was also non-significant ($F_{(1,19)} = 0.861$, $p = .365$). Non-parametric Wilcoxon signed-rank tests also showed no effect of SURPRISAL within the *Gesture absen*t ($z = 1.4$, signed rank = 142, $p = 0.919$), and the *Gesture present* conditions ($z = -0.35$, signed rank = 95, $p = 0.361$). Similar to lexical frequency, ERPs on high vs. low surprisal did not exhibit interpretable patterns.

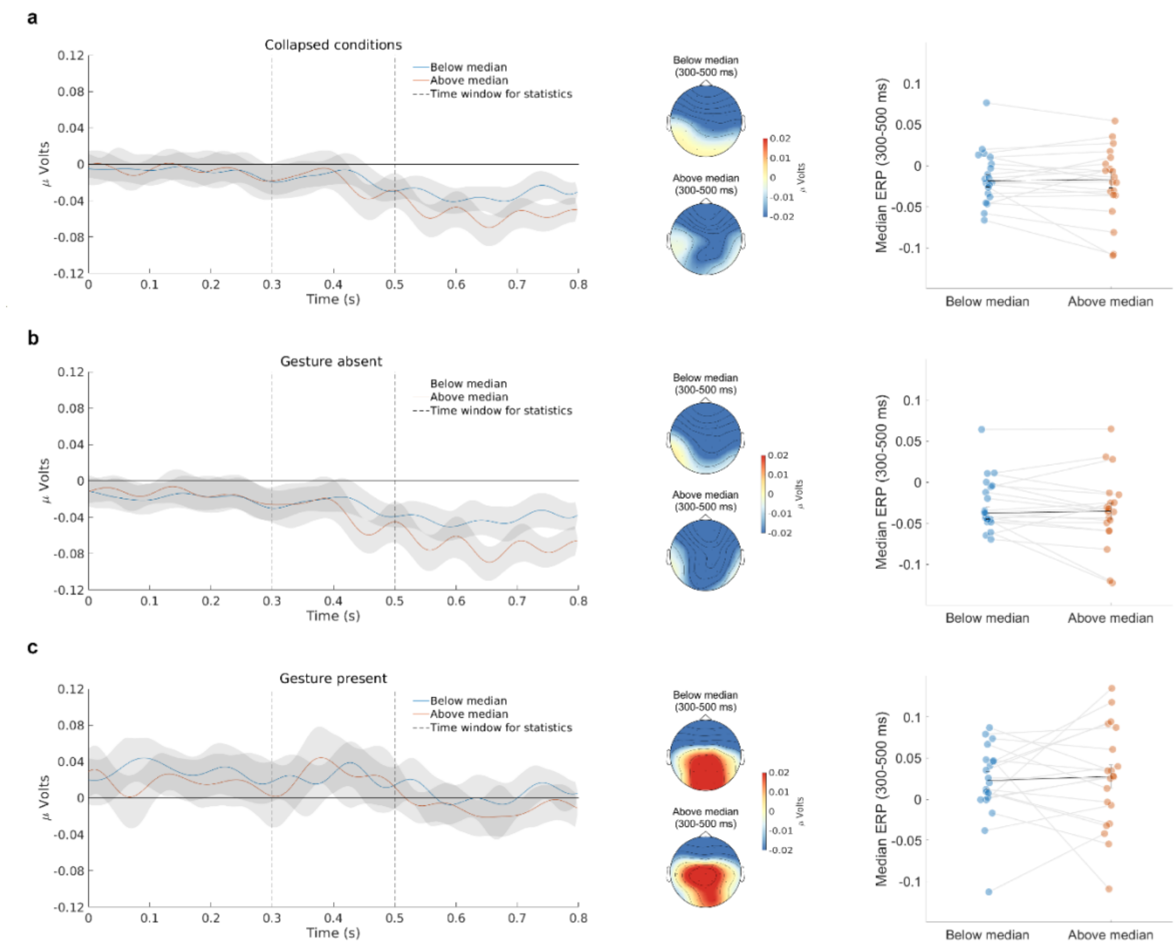


Figure s3: Results of the ERP analyses for GPT-2 surprisal for **a**. All words collapsed across *Gesture absent* and *Gesture present* conditions. **b**. Words in the *Gesture absent* condition. and **c**. Words in the *Gesture present* condition. For all panels, waveforms of high-surprisal words (Above median) are depicted in red, and low-surprisal words (Below median) are depicted in blue. ERPs were obtained from the average of all electrodes. Coloured shades show the 95% bootstrapped confidence intervals. Dotted lines represent our time-window of interest for statistical analyses (300–500 ms). Topographical distributions show ERP responses between 300–500 ms. Point- and line-plots show individual and group median ERP amplitudes between 300–500 ms.

Discussion on the median-split ERPs

Our conventional median-split ERP analyses revealed no systematic and interpretable effects for either frequency or surprisal. Our findings, at first glance, might be at odds with prior publications showing either an N400 effect of high vs. low surprisal using unimodal language stimuli e.g., visually presented sentences (Frank et al., 2015). Notably, in this study, the stimuli were presented in an RSVP manner, and only one critical word within one sentence were analysed for the ERPs. Regarding frequency, despite a long list of literature employing factorial ERP design (e.g., Van Petten and Kutas, 1990), no prior literature has used even unimodal but naturalistic stimuli with median-split to examine how the N400 ERPs vary as a function of word frequency. On the other hand, in the multimodal language processing literature, a number of studies have reported and N400 effect using conventional ERP methods (Kelly et al., 2004; Morett et al., 2020; Wang and Chu, 2013; Wu and Coulson, 2005). Similar to the study from Frank and colleagues (2015), the gesture-speech N400 studies also focused on a single critical word within sentences, so that there is no overlap of EEG time-windows between words. As a result, conventional ERP analysis focusing on the N400 window has led to reliable and interpretable data pattern. This is, however, not the case for the current study, as shown from our results. Here, the validity, and the comparability to prior studies when using conventional ERP methods is undermined by two important confounds: the temporal segments of words being analysed overlap with each other; and all these words were multimodal in nature (with facial expressions, lip movements, and where applicable, hand gestures). Thus, we argue that the conventional ERP analysis (here using the median-split approach) has limitations to answer our research questions; and the effect of surprisal and word frequency, and their interaction with co-speech gestures is best examined via the temporally-resolved regression approach, as we reported in the main text.

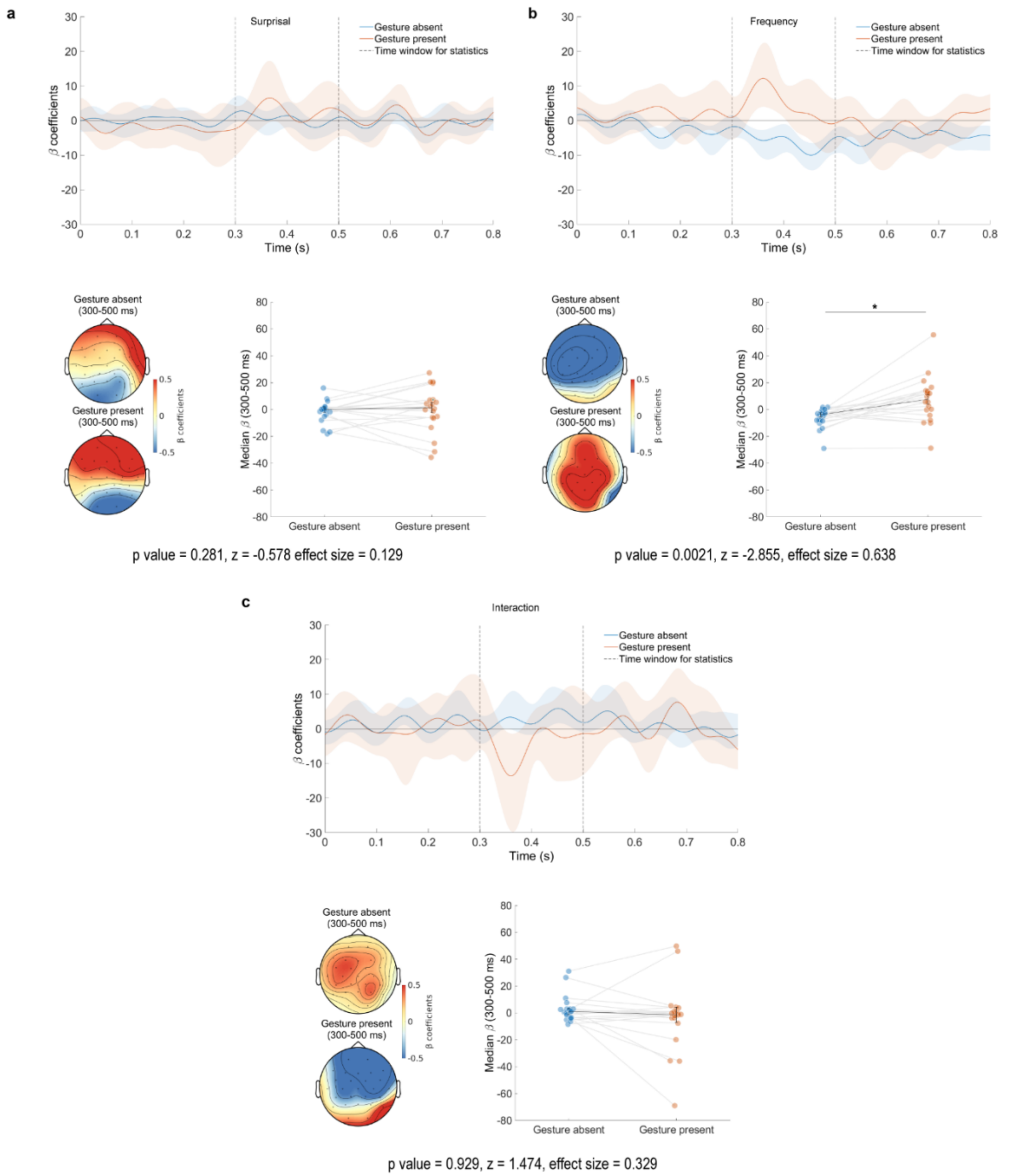


Figure s4. Results of mTRF analyses for surprisal (a), frequency (b) and their interaction (c). Mean mTRF to *Gesture absent* (blue) and *Gesture present* (orange) words obtained from the average of a group of centro-parietal electrodes with their 95% bootstrapped confidence intervals. Topographical distributions are based on the mean mTRF responses between 300–

500 ms. Point and line-plots show Individual and group median mTRF beta values between 300–500 ms. The line and asterisk represent a statistically significant effect ($p < 0.05$, one-tailed).

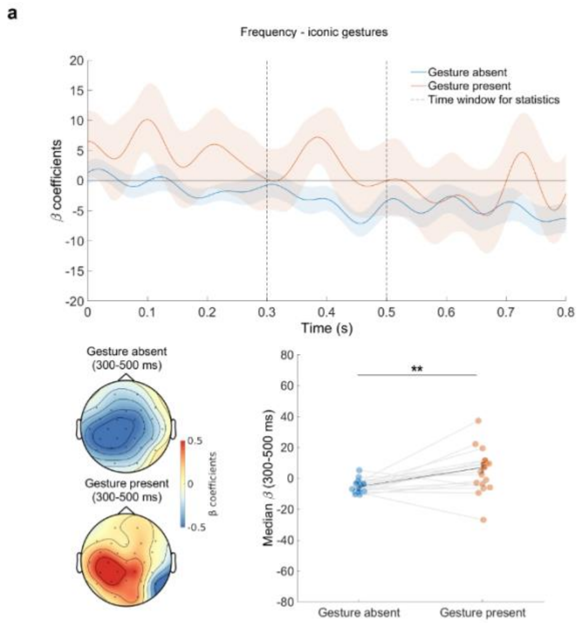


Figure s5. Results of mTRF analyses for frequency using iconic gestures (a) and beat gestures (b) only. Mean mTRF to *Gesture absent* (blue) and *Gesture present* (orange) words obtained from the average of a group of centro-parietal electrodes with their 95% bootstrapped confidence intervals. Topographical distributions are based on the mean mTRF responses between 300–500 ms. Point and line-plots show Individual and group median mTRF beta values between 300–500 ms. The line and asterisks represent a statistically significant effect ($p < 0.01$, one-tailed).

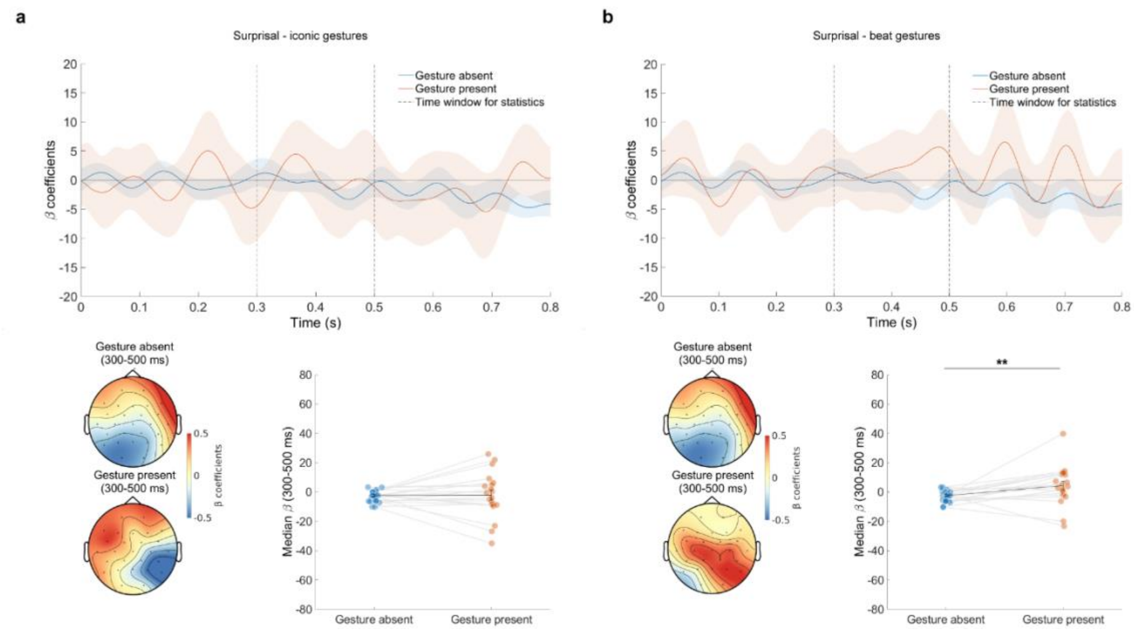


Figure s6. Results of mTRF analyses for surprisal using iconic gestures (a) and beat gestures (b) only. Mean mTRF to *Gesture absent* (blue) and *Gesture present* (orange) words obtained from the average of a group of centro-parietal electrodes with their 95% bootstrapped confidence intervals. Topographical distributions are based on the mean mTRF responses between 300–500 ms. Point and line-plots show Individual and group median mTRF beta values between 300–500 ms. The line and asterisks represent a statistically significant effect ($p < 0.01$, one-tailed).