More than words: The online orchestration of word predictability, prosody, gesture, and mouth movements during natural language comprehension

Ye Zhang[1], Diego Frassinelli[2], Jyrki Tuomainen[3], Jeremy I Skipper[1], Gabriella Vigliocco[1]

[1]Experimental Psychology, University College London, UK
[2]Institute of Natural Language Processing, University of Stuttgart, Germany
[3]Speech, Hearing and Phonetic Sciences, University College London, UK

Abbreviated title: Electrophysiology of multimodal comprehension

Corresponding Author
Ye Zhang
Experimental Psychology
University College London
26 Bedford Way
London WC1H OAP
United Kingdom
E-mail: y.zhang.16@ucl.ac.uk

Number of pages: 30
Number of Figures: 7
Number of words: 226 abstract, 646 introduction, 1,469 discussion

Acknowledgements: The work reported here was supported by a European Research Council Advanced Grant (ECOLANG, 743035) to Gabriella Vigliocco.

The authors declare no competing financial interests.

Abstract

Communication naturally occurs in dynamic face-to-face environments where spoken words are embedded in linguistic discourse and accompanied by multimodal cues. Existing research supports predictive brain models where prior discourse but also prosody, mouth movements, and hand-gestures individually contribute to comprehension. In electroencephalography (EEG) studies, more predictable words show reduced negativity, peaking at approximately 400ms after onset of the word (N400). Meaningful gestures and prosody also reduce the N400, while the effects of rhythmic gestures and mouth movements remain unclear. However, these studies have only focused on individual cues while in the real world, communicative cues co-occur and potentially interact. We measured EEG elicited by words while participants watched videos of a speaker producing short naturalistic passages. For each word, we quantified the information carried by prior linguistic discourse (surprisal), prosody (mean pitch), mouth informativeness (lip-reading), and presence of meaningful and/or rhythmic gestures. Discourse predictability reduced the N400 amplitude and multimodal cues impacted and interacted with this effect. Specifically, higher pitch and meaningful gestures reduced N400 amplitude, beyond the effect of discourse, while rhythmic gestures increased the N400 amplitude. Moreover, higher pitch overall reduced N400 amplitude while informative mouth movements only showed effects when no gestures were present. These results can constrain existing predictive brain models in that they demonstrate that the brain uses cues selectively in a dynamic and contextually determined manner in real-world language comprehension.

Significance Statement

Language originated, is learnt and is mostly used in face-to-face contexts comprising linguistic information and multimodal cues such as intonation of the speech, meaningful/rhythmic hand gestures and mouth movements. Yet, the rich multimodal context of real-world language use is usually stripped away in language studies. Here, we developed quantifications of the visual and acoustic non-linguistic cues accompanying speech and we investigate how, jointly, the cues shape brain activity. We find that the brain activity associated with processing is always modulated by the multimodal non-linguistic cues and their interactions. Thus, this work questions the validity of language comprehension models only based on speech. It further presents new methods for the study of real-world language.

Introduction

Language originated, is learned and is most often used in face-to-face contexts. These contexts include linguistic information like prior discourse and multimodal cues such as speech intonation (prosody), hand gestures, and mouth movements. Behavioural, electrophysiological and neuroimaging work has shown that these cues each individually improve speech perception and language comprehension (Cutler et al., 1997; Kuperberg & Jaeger, 2015; Massaro & Jesse, 2007; Özyürek, 2014). However, most of what we know about language comes from studies focused on only one (usually linguistic) cue. A neurobiological account of how multiple cues are used and how they interact is missing.

There is increasing evidence for prediction across distributed brain regions as a mechanistic account of brain functioning in general, and (audiovisual) speech perception and language comprehension in particular (Arnal, Wyart and Giraud, 2011; Clark, 2013; Friston and Kiebel, 2009). In some predictive models, multimodal cues are argued to help predict upcoming sounds or words (Holler & Levinson, 2019). The N400, a negative-going event-related potential (ERP) recorded from centro-parietal electrodes 200-600ms after word onset, is a biological marker of semantic processing difficulty, at least partly associated with word predictability (Kutas & Federmeier, 2010). Words that are less predictable in linguistic context elicit more negative N400 (Frank et al., 2015; Haggort & Brown, 2000; Kutas & Hillyard, 1980).

Like prior linguistic context, multimodal cues provide information that might support prediction of upcoming words. New information, introduced by prosodic stress elicits smaller N400 amplitude compared to unstressed words (e.g. Magne et al, 2005). Meaningful co-speech gestures make upcoming words more predictable. For example, activating the less predictable meaning of the homonymous word “ball” using a “dancing” gesture, reduces the N400 response to a later mention of “dance” (Holle and Gunter, 2007; Obermeier et al., 2011; Obermeier et al., 2012). In line with the EEG findings, meaningful gestures result in a significant reduction in cortical activity in language regions, a hallmark of prediction (Skipper, 2014). Very few studies have investigated beat gestures (meaningless gestures time-locked to the speech rhythm, McNeil, 1992). Beats have been argued to enhance saliency of associated speech (Krahmer & Swerts, 2007) like prosodic stress (Hubbard et al.,

2009; Wang and Chu, 2013). Finally, while many studies have focused on the sensory mechanisms underscoring the use of mouth movements in speech, less is known about whether mouth movements affect word predictability (but see Brunellière et al., 2013; Drijvers and Özyürek, 2017; Hernández-Gutiérrez et al., 2018; Skipper et al 2007; 2009).

Crucially, most previous studies focused on single cues, rather than considering their natural co-occurrence. One important reason for this pertains the challenge of doing experimental research with naturalistic materials (Alday, 2019). Thus, many have used a reductionist approach to ensure experimental control (e.g., "normalising" prosody and avoiding audiovisual presentation when studying speech; showing only the mouth when studying audiovisual speech perception; hiding the face when studying gestures). This approach, however, risks breaking the natural (and predictive) correlation among cues with unknown consequences on processing (Hasson et al., 2018; Skipper, 2015).

Here we ask whether/how the multimodal cues affect words predictability in more naturalistic settings. In doing so, we also address the methodological issues above. We measured the electrophysiological responses to each cue and their interactions, elicited by words in naturalistic speech videos. We quantified word-predictability using surprisal, a measure of the probability that a word follows from previous words (Frank et al., 2015; Frank, 2017; Smith and Levy, 2013). We also quantified prosody, gesture and mouth informativeness. We predict that, as in prior studies, word-predictability will modulate N400 amplitude. Prosody and meaningful gestures will reduce the N400 amplitude as both provide meaningful information that makes upcoming words, especially those less predictable based on linguistic context, more predictable and, therefore, easier to process. Finally, we expect an interaction between meaningful gestures and mouth movement based on previous behavioral and fMRI work (Drijvers and Özyürek, 2017; Skipper et al 2007; 2009).

Methods

Participants

Thirty-six native English speakers with normal hearing and normal or corrected to normal vision were paid £8 to participate in the present study after giving written consent. Five participants were excluded, three due to technical issues, one for falling asleep, and one for excessive muscle noise, leaving thirty-one participants (mean age=27, SD=9.95, 17 women). The experimental procedure was approved by the UCL Division of Psychology and Language Science ethics committee.

Material

Two-hundred and forty-six naturalistic sentence pairs (two consecutive sentences) were extracted from the British National Corpus (BNC, University of Oxford, 2007 http:// www.natcorp.ox.ac.uk/). Sentences were selected in a semi-random fashion with the only constraints the second sentence had to be at least five words long, and contain at least one verb that could be easily gestured (e.g. turn the pages). If necessary, we edited slightly the first sentence to facilitate readability and resolved all ambiguities (e.g. proper nouns without a clear reference were changed into pronouns), while the second sentence was kept unmodified. Twelve native English speakers were paid £2 each to evaluate the sentence pairs for grammaticality, meaningfulness and gesturability on a 1-5 scale. We selected 103 sentence pairs that had averaged gesturability > 2 (and SD < 2.5); and had no grammatical errors or semantic anomalies. Three sentence pairs were used for practice, and 100 were used as stimuli (Mean gesturability = 2.67, SD = 0.58).

A native British English-speaking actress produced the 103 sentence pairs. She stood in front of a dark-blue background, wearing black T-shirt and trousers to keep her arms and hands

visible, and did not wear glasses to keep her face visible. She was instructed to read out the sentences presented behind the camera at a natural speed, with natural prosody and facial expressions. Each sentence pair was recorded with and without gestures. For videos with gestures, the actress was instructed to gesture as she naturally would. For videos without gestures, she was asked to stand still keeping her arms along her body.

Quantification of Cues

The onset and offset of each word were automatically detected using a word-phoneme aligner based on Hidden Markov Models (Rapp, 1995). The timing was then checked and corrected manually if needed. The mean word duration was 440.32ms (SD = 375.69ms). Next, for each

	Mary	thought	of	joining	Clara.	They	might	pick	flowers	together.
Surprisal	-	6.73	-	9.02	-	-	-	8.63	9.41	7.40
Prosody	-	389.29	-	342.19	-	-	-	432.18	364.69	482.95
Meaningful Gesture	-	0	-	1	-	-	-	1	0	0
Beat Gesture	-	1	-	0	-	-	-	0	0	0
Mouth Movement	-	0.39	-	0.02	-	-	-	0.1	0.21	0.32

Figure1. Example of stimuli and annotations. Annotation was carried out for content words only. Each frame corresponds to an image during each such word. Face area was covered to comply with BioRxiv policy only and was visible to participants.

content word (i.e., nouns, adjectives, verbs and adverbs) we quantified the informativeness of different multimodal cues. We did not quantify measures of informativeness for function words (i.e., articles, pronouns, auxiliary verbs and prepositions) because Frank et al. (2015) failed to show any effect of the predictability (measured as surprisal) for such words. Figure 1 gives an example sentence.

Linguistic predictability was measured using surprisal (Mean surprisal = 7.92, SD = 2.10, range: 2.98 - 17.30), defined as the negative log likelihood of the probability of a word to follow a sequence of other words (Shannon, 1949). Previous work has shown that surprisal provides a good measure of predictability, especially for low predictability words (Smith & Levy, 2013) and predicts reading times (e.g. Hale, 2001; Levy, 2008; Smith and Levy, 2013) and N400 amplitude (Frank et al., 2015). Here, surprisal was generated using a bigram language model trained on the lemmatized version of the first slice (~19-million tokens) of the ENCOW14-A corpus (Schäfer, 2015), a 16-billion token English web corpus. We chose a bigram model to reduce data sparsity and, consequently, increase the robustness of our surprisal measures. Moreover, Frank and colleagues (2015) showed that bigram models perform equally well, if not better than more complex models - trigram, recurrent neural networks (RNN) and probabilistic phrase-structure grammar (PSG) - in fitting N400 data. Once trained, the bigram model was used to calculate the surprisal of each word in the sentence pairs based on previous content words in the two sentences using the following formula:

$$\text{Surprisal}(w_{t+1}) = -\log P(w_{t+1}|w_{1\ldots t})$$

where w_{t+1} indicates the current word, and $w_{1\ldots t}$ stands for previous content words in the two sentences. We also developed models in which the number of content words used in

computing surprisal was varied. Given the minor differences observed, we decided to include all previous content words in the two-sentence stimuli (from the first word in the first sentence to the word preceding the target word, in the Extended Data (Figure 2-1) we show results for different window sizes to justify this choice).

Prosodic information for each word was quantified as the mean F0 of the word (mean F0=298.39Hz, SD=84.19, range: 90.75-602.56). We automatically extracted mean F0, maximum F0, minimum F0, mean intensity and F0 change per each word using Praat (version 6.0.29, Boersma and Weenink, 2018). A comparison of results obtained with these different pitch measurements showed that they are similar, therefore we chose to report results for mean F0 because it has been used most often as a measure of prosody in previous work (e.g. Kakouros et al., 2018).

Gestures were coded in ELAN 5.0.0 (Sloetjes and Wittenburg, 2008) as meaningful gestures or beats by an expert coder. Coding was checked for reliability by asking a second expert coder to annotate 10% of the productions, resulting in an intercoder reliability of 85.1%. Meaningful gestures (N = 359) comprised iconic gestures (e.g. drawing movements for the word “drawing”) and deictic gestures (e.g. pointing to the hair for “hair”). Beat gestures (N = 229) comprised rhythmic movements of the hands without clear meaning (McNeill, 1992), but regarded to enhance the salience of the speech (Krahmer and Swerts, 2007).

Mouth informativeness was quantified per word in a separate online experiment. The video-clips (from the ‘without gesture’ condition) were muted, cropped to leave only the face and segmented at the word level. The resulting video clips were presented using the Gorilla Experiment Builder (https://gorilla.sc) to 95 native English speaking subjects who did not participate in the main experiment. Participants were recruited through Prolific Academic

(https://www.prolific.ac) and were paid £2 for participation. Each video clip was presented at the top of the screen with four words presented at the bottom, including the target word and three foils randomly selected from all words in the experiment with the same length as the target. Participants were instructed to select the word matching the video. They were allowed to watch the video for as many times as they needed before making their choice. The four words were not shown until after the participant played the video. Each word received 10 responses, and we calculated the mouth informativeness of each word using its mean accuracy (mean=0.77, SD=0.42) divided by the average number of times the clip was played (mean=2.83, SD=2.52; mean informativeness = 0.32, SD = 0.15, range: 0-1).

Procedure

After the three sentence-pairs given as practice trials, each participant was presented with 50 gesture and 50 no gesture videos in randomized order using Presentation software (V. 18.0 Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). Each sentence pair was separated by a 2000ms interval. The experiment was counterbalanced across every two participants so that they watched the videos in the same order but with counterbalanced gesture/no-gesture conditions. One-third of the videos (35) were followed by yes/no questions about the content of the video to ensure participants paid attention during the experiment and to acquire behavioral responses. For example, for the video “Emma screamed and swore at them. She was especially angry if the girls dared to eat any of her food or drink her coffee”, the question was “Is Emma going to share her sweets with the other girls?”. Participants sat comfortably one meter away from the screen with a resolution of 1024*768, wearing 50Ω headphones, and were instructed to watch the videos carefully and to answer questions as quickly and accurately as possible (prioritizing accuracy) by pressing the left

("Yes") or right ("No") control key. Participants were asked to avoid moving, keep their facial muscles relaxed and reduce blinking, but they were also told that it is better to blink then to tense up due to discomfort. Similar instructions were written on the screen. The recording took thirty minutes with three breaks.

EEG Recording

A 32-channel BioSimi system with silver-silver chloride electrodes and 24 bit resolution was used for the EEG recording, following a 10-10 international system layout. A common reference included the CMS electrode (serving as the online reference) and DRL electrode (serving as the ground electrode). Elastic head caps were used to keep the electrodes in place. Two external electrodes were attached under the left and right mastoids for off-line reference, while two other external eye electrodes were attached below the left eye and on the right canthus to detect blinks and eye movements. Electrolyte gel was inserted on each electrode to improve connectivity between the skull and the electrode. To check for relative impedance differences, the electrode offsets were kept between +/-25mV. The recording was carried out in a shielded room with the temperature kept at 18 °C.

Behavior Analysis

We used generalized mixed effect modeling to test whether surprisal had an effect on the accuracy and response time for questions following 35 sentences. The analysis was conducted using LME4 (Bates et al., 2015) package running under R Studio (version 3.4.1, RStudio Team, 2015). We used logistic regression in the accuracy analysis and linear regression for response time. In both, mean surprisal (averaged across all content words in the video) was the predictor variable, participant and sentence pair were added as random intercept to control for by participant and by video variation. All continuous variables

(response time and surprisal) per sentence were standardized (centered and scaled) using the "scale" function built in R so that each coefficient represents the effect size of the variable.

EEG pre-processing

The raw data were pre-processed with EEGLAB (v 14.1.1, Makeig et al., 2004) and ERPLAB version 7.0.0 (Lopez-Calderon and Luck, 2014) running under MATLAB (R2017b, The MathWorks, Inc.). All electrodes were included. Triggers were sent at the onset of each video, and word onset was subsequently calculated from the word boundary annotation. Any lag between trigger and stimuli presentation was also measured and corrected (Mean=0.21s, SD=0.07). The EEG file was re-referenced to average of the left and right mastoids (A1 and A2), down-sampled from 2048Hz to 256Hz to speed up preprocessing, and separated into epochs each containing data from -100 to 924ms around word onset (Frank et al., 2015). The data was filtered with a second order Butterworth 0.05-100Hz band-pass filter. Due to the likely overlap between any baseline period (-100 to 0ms) and the EEG signal elicited by the previous word, we did not perform baseline correction, but instead extracted the mean EEG amplitude in this time interval and later used it as a control variable in regression analysis (Frank et al., 2015). We conducted independent component analysis based artifact correction (ICA, Chaumon et al., 2015). Two independent experts manually labelled eye movement and other noise (e.g. heart beat, line noise) components that were subsequently removed from the data. Further artifact rejection was conducted by first using a moving window peak-to-peak analysis (Voltage Threshold=100 μV, moving window full width=200 ms, window step=20 ms) and then step-like artifact analysis (Voltage Threshold=35 μV, moving window full width=400 ms, window step=10 ms). This resulted in an average rejection of 12.43%

(SD=12.49) of the data. The ERP files were then computed from pre-processed data files, and were additionally filtered with a 30Hz low-pass filter.

EEG Analysis: Hierarchical Linear Modeling

It is an empirical question as to what time window the effect of linguistic context would emerge, given that no previous study has investigated the effect of surprisal in audiovisual multimodal communication. Therefore, rather than specifying the N400 window a priori, we first identified the time window where electrophysiological responses were sensitive to surprisal using hierarchical linear modeling.

While traditional ERP analysis compares different conditions and thus may require dichotomization of the predictor variable, recent regression based ERP analysis such as hierarchical linear modelling can linearly decompose an ERP into time-series of beta coefficient waveforms elicited by continuous variables. Significant differences between the beta coefficient waveforms and zero (flat line) represent the existence of an effect (Smith & Kutas, 2015a, 2015b).

We used the LIMO (hierarchical LInear MOdeling) toolbox (Pernet et al., 2011) working under MATLAB (R2017b, The MathWorks, Inc.). For each participant, we created a single-trial file from the EEG file, and a continuous variable containing surprisal of each word. In the first level analysis for each participant, a regression was performed for each data point in 0-924ms time window per electrode per word, with EEG voltage as the dependent variable and word surprisal as the independent variable, thus generating a matrix of beta values, which indicate whether and when surprisal has an effect for each participant. In the second level analysis across all participants, the averaged beta matrix was compared with 0 using a one-

sample t-test (bootstrap set at 1000, clustering corrected against spatial and temporal multiple comparison, Pernet et al., 2015).

EEG Analysis: Linear Mixed Effect Regression Analysis

After determining the time window where surprisal has an effect, we performed linear mixed effect analysis (LMER) on the resulting time window. LMER was selected due to the advantage in accommodating both categorical and continuous variables, thus increasing statistical power (MacCallum et al., 2002). Moreover, LMER can account for both by participant and item variance and can better accommodate unbalanced designs, suitable for EEG studies investigating naturalistic language processing (e.g. Frank et al., 2015; Alday et al., 2017). We excluded from analyses: (a) all function words, modal words, and proper names; (b) words without a surprisal value (26 words, due to the lack of occurrence of the combination between the word and its context in the corpus); (c) words without a mean F0 score (4 words, due to insufficient data points when calculating the average); (d) words associated with both beat and meaningful gestures (3 words); (e) words occurring without any gesture in the "with gesture" condition, and the corresponding words in without gesture videos. This was done to avoid data unbalance as there were three times more words with no gestures (combining the videos with and without gestures). Mean ERP in the 300-600ms time window (as determined in the prior hierarchical linear modeling step) was extracted from 32 scalp electrodes for each word and was used as the dependent variable. Mean ERP in the -100 to 0ms time window was extracted as the baseline. The independent variables included 1) predictors: log-transformed surprisal, mean F0, meaningful gestures, beat gestures, mouth informativeness, and all up to three-way interactions between them, excluding interactions containing meaningful gesture*beat gestures (as the three instances were removed from the

data), 2) control variables: baseline extracted between -100 to 0ms, word frequency, word length, word order in the sentence, sentence order in experiment, relative position of each electrode measured by its X, Y, and Z coordinate position (Winsler et al., 2018) acquired from BioSemi website (https://www.biosemi.com/download.htm). No main or interaction effects showed multicollinearity, with variance inflation factor (VIF) less than 2, kappa=4.871. All continuous variables, including ERP, surprisal, mean F0, mouth informativeness, baseline, frequency, word length, word order, sentence order, and X, Y, Z position of electrodes were standardized (centered and scaled) so that each coefficient represents the effect size of the variable. Surprisal and frequency were log transformed to normalize the data. All categorical variables were sum coded so that each coefficient represents the size of the contrast from the given predictor value compared with the grand mean (intercept) (Alday et al., 2017).

We further included word types and participant as random intercept in the random structure. We attempted to construct a maximal random structure by entering all main and interactions as a random slope of participants (Barr et al., 2013), but the model failed to converge. As a result, we included the highest interaction (three-way interactions) as random slope for participants (Barr, 2013), and surprisal was included as random slope for lemma. Our analysis included 31 participants, 381 word type lemmas and 480,212 data points.

Results

Behavioral Analysis

We found that accuracy decreased with an increase in surprisal (Mean=0.821, SD=0.384, β=-0.784, $p<.001$). Similarly, we found that sentences with higher averaged surprisal had slower reaction times (Mean=41298.82, SD=28813.35, β=0.089, $p=.024$).

EEG Analyses

Time Window Sensitive to Linguistic Context

We found that the beta values of central-parietal electrodes were significantly more negative compared with 0 in around 300-600ms time window across electrodes (Figure 2). Words with higher surprisal, elicited more negative signal or larger N400 amplitude (see Extended Data, Figure 2-1, for comparison between surprisal generated by using different lengths of previous context). No other time window was significantly sensitive to surprisal.

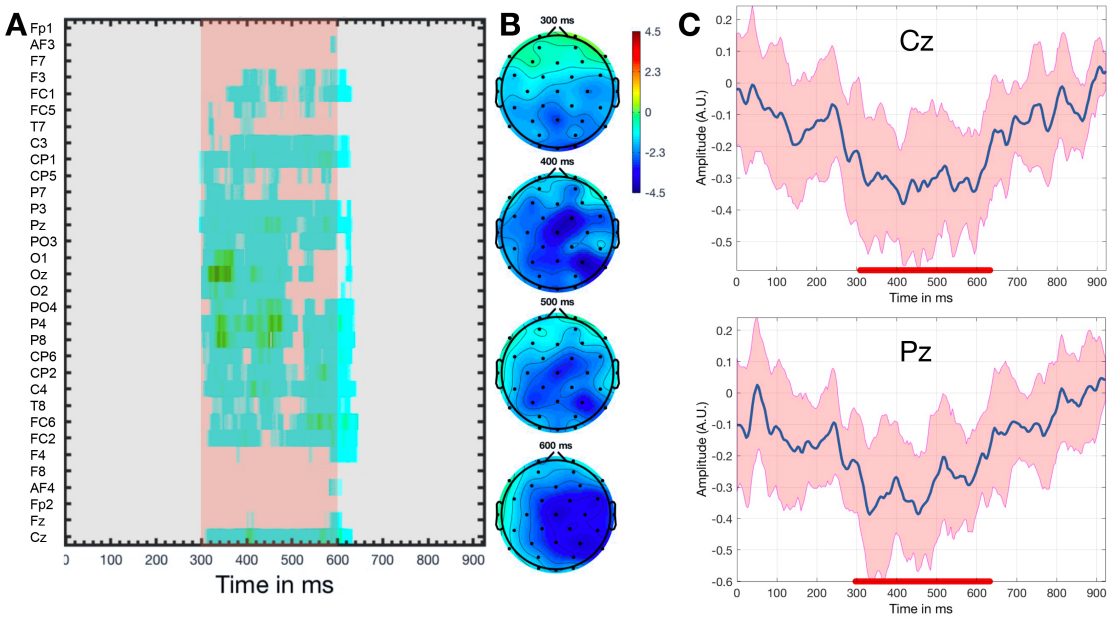

Figure 2. Hierarchical linear modelling results showing the ERP sensitive to surprisal (one-sample t-test P<0.05, cluster-corrected). (A) Time window (300-600ms) showing increased significant negativity associated with surprisal (in pink). Grey areas are not statistically significant. (B) Topographic maps illustrating the scalp distribution for the 300-600 time window. Deeper blue area indicated more negative beta values. (C) Averaged beta plot for electrode Cz and Pz illustrating that beta values for surprisal were significantly negative compared with 0 (flat waveform) in 300-600ms. The blue line indicates the average beta value, while red indicates the confidence interval. The red line underlying the figures indicated the significant time window. Cz and Pz are chosen here because they are most often used to depict N400 effects (that are maximal at central-parietal locations, Kutas & Federmeier, 2010)

Word predictability is Modulated by Multimodal Cues

We focus first on the main effects of the multimodal cues and their interactions with surprisal to establish whether multimodal cues mediated the effect of the predictability of linguistic context. Full model results are reported in Table 1.

Fixed Effects	β	Std Error	t	p
(Intercept)	0.004	0.010	0.416	0.677
Predictor Variables				
Surprisal	0.011	0.016	0.726	0.468
Mean F0	0.010	0.002	4.710	<.001***
Mouth Informativeness	-0.004	0.003	-1.477	0.140
Meaningful Gesture (Present)	0.006	0.001	4.867	<.001***
Beat Gesture (Present)	-0.005	0.001	-3.386	0.001**
Surprisal:Mean F0	0.017	0.002	7.996	<.001***
Surprisal:Mouth Informativeness	-0.002	0.002	-0.726	0.468
Surprisal:Meaningful Gesture (Present)	0.008	0.001	5.653	<.001***
Surprisal:Beat Gesture (Present)	-0.012	0.001	-9.248	<.001***
Mean F0:Mouth Informativeness	0.003	0.002	2.051	0.040*
Mean F0:Meaningful Gesture (Present)	0.004	0.001	3.710	<.001***
Mean F0:Beat Gesture (Present)	-0.006	0.002	-3.777	<.001***
Mouth Informativeness:Meaningful Gesture (Present)	-0.006	0.001	-5.757	<.001***
Mouth Informativeness:Beat Gesture (Present)	0.003	0.002	1.874	0.061
Surprisal:Mean F0:Mouth Informativeness	-0.013	0.005	-2.554	0.011*
Surprisal:Mean F0:Meaningful Gesture (Present)	0.007	0.006	1.278	0.201
Surprisal:Mean F0:Beat Gesture (Present)	0.002	0.006	0.453	0.650
Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	-0.010	0.005	-2.168	0.030*

Surprisal:Mouth Informativeness:Beat Gesture (Present)	0.003	0.004	0.683	0.495
Control Variables				
Word Order	-0.010	0.002	-4.782	<.001***
Word Length	-0.011	0.004	-2.361	0.018*
Sentence Order	-0.009	0.001	-9.247	<.001***
Baseline	0.788	0.001	876.10	<.001***
Frequency	0.036	0.009	3.834	<.001***
Electrode X	-0.006	0.001	-7.075	<.001***
Electrode Y	0.008	0.001	8.622	<.001***
Electrode Z	0.001	0.001	0.934	0.351

Random Effects		Variance	Std.Dev.
Lemma	(Intercept)	0.012	0.111
	Surprisal	0.045	0.213
Participant ID	(Intercept)	0.001	0.034
	Surprisal:Mean F0:Mouth Informativeness	0.001	0.028
	Surprisal:Mean F0:Meaningful Gesture (Present)	0.001	0.030
	Surprisal:Mean F0:Beat Gesture (Present)	0.001	0.030
	Surprisal:Mouth Informativeness:Meaningful Gesture (Present)	0.001	0.024
	Surprisal:Mouth Informativeness:Beat Gesture (Present)	0.001	0.023

Model				
AIC	BIC	logLik	deviance	df.resid
892158.7	892746	-446026.3	892052.7	480159

Notes. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 1: Full results of linear mixed effect regression model with N400 (300-600ms) amplitude as dependent variable

As shown in Figure 3 (panel A), we found a main effect of prosody (mean F0) (β=0.010, p<.001). Words produced with higher mean pitch showed less negative EEG, or smaller N400

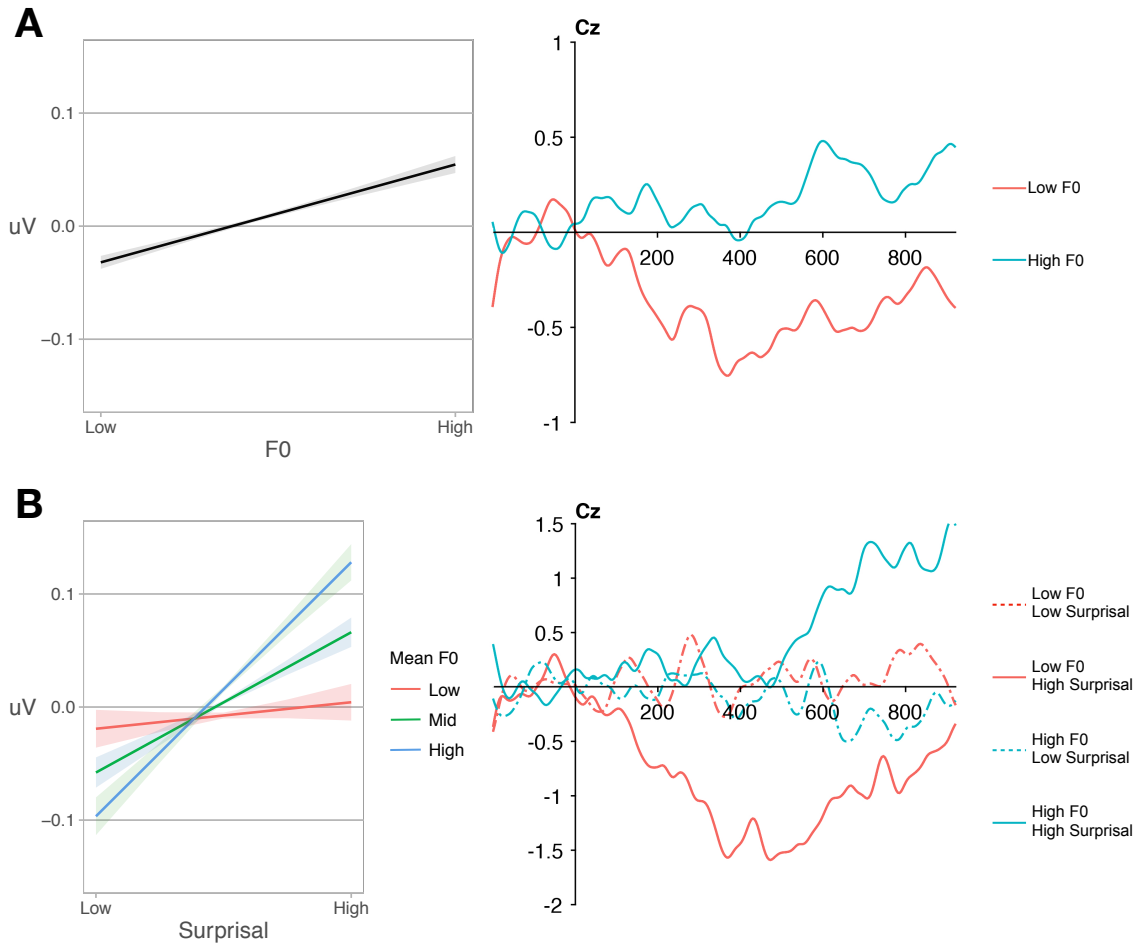

Figure 3. Prosodic stress (mean F0) modulation of N400 amplitude. (A) Main effect of Prosodic Stress. (B) Interaction between Prosodic Stress and Surprisal. Plots on the left depict the predicted value of expected value of the mean amplitude of the event-related potentials (ERPs) within 300-600ms (grey areas = confidence intervals). Plots on the right show the EEG waveform. For illustrative purposes, in the EEG plots all continuous variables are categorized and the EEG waveform was additionally filtered with 15Hz low pass filter.

amplitude in the 300-600 time window, compared with words produced with lower pitch. Figure 3, panel B reports the interaction between surprisal and mean F0 (β=0.017, p<.001). The larger N400 amplitude associated with high surprisal words was modulated by pitch. High surprisal words elicited a larger reduction of N400 amplitude when the pitch was higher, in comparison to low surprisal words.

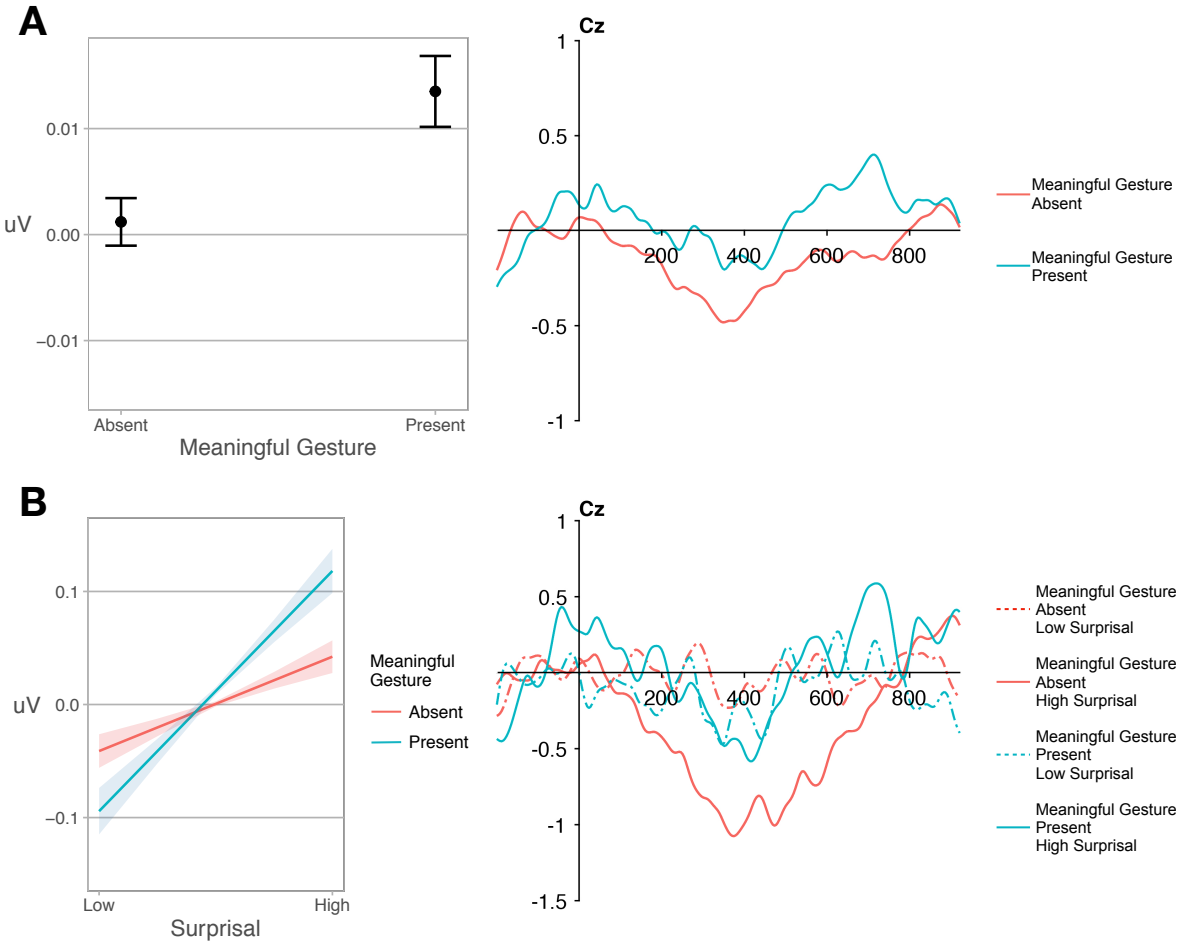

Figure 4. Meaningful gestures' modulation of N400 amplitude. (A) Main effect of meaningful gestures. (B) Interaction between Meaningful Gestures and Surprisal. Plots on the left depict the predicted value of expected value of the mean amplitude of the event-related potentials (ERPs) within 300-600ms (grey areas = confidence intervals). Plots on the right show the EEG waveform. For illustrative purposes, in the EEG plots all continuous variables are categorized and the EEG waveform was additionally filtered with 15Hz low pass filter.

We found a similar main effect of meaningful gestures (Figure 4). Words accompanied by a meaningful gesture showed a significantly less negative N400 (β=0.006, p<.001). There was also a significant interaction between surprisal and meaningful gesture, indicating that for high surprisal words, the presence of a meaningful gesture makes the N400 less negative (β=0.008, p<.001).

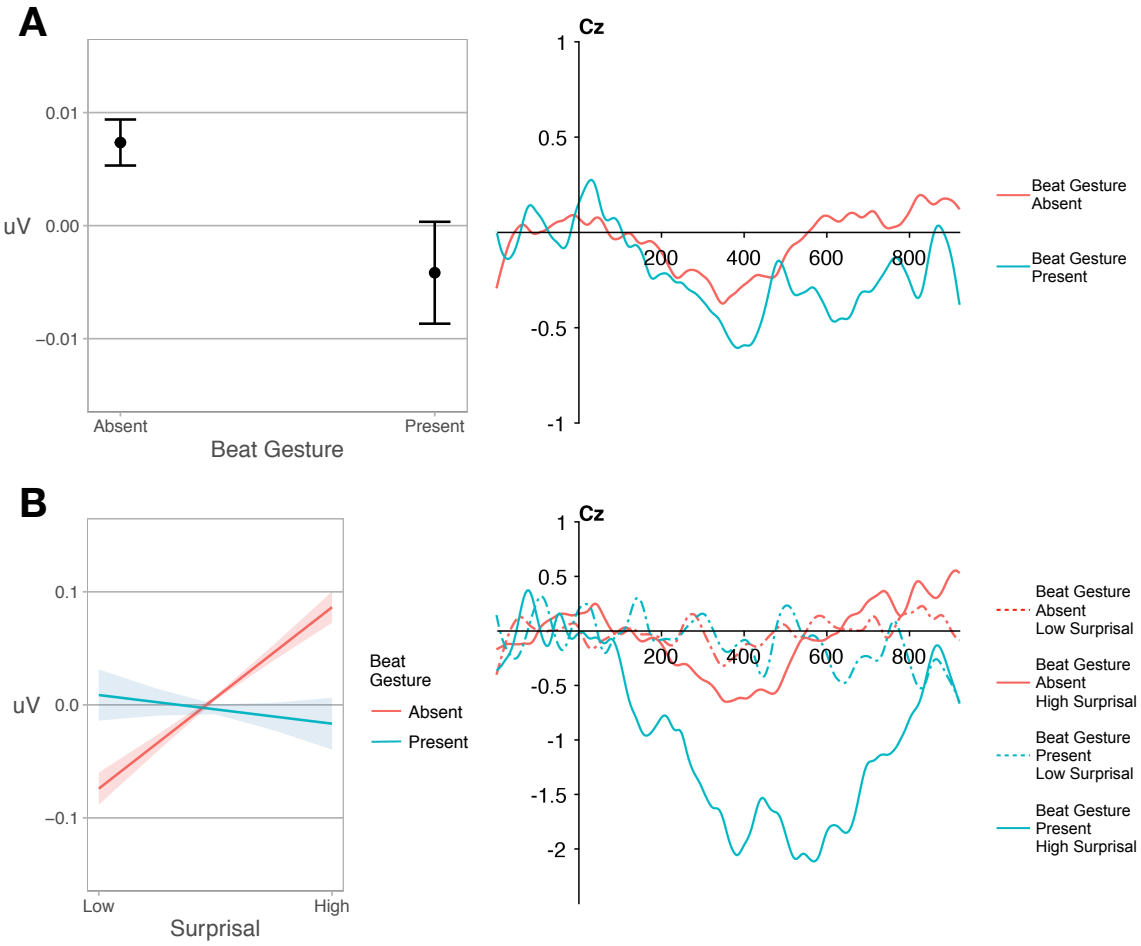

Figure 5. Beat gestures modulations of N400 amplitude. Main effect of beat gestures. (B)Interaction between Beat Gestures and Surprisal. Plots on the left depict the predicted value of expected value of the mean amplitude of the event-related potentials (ERPs) within 300-600ms (grey areas = confidence intervals). Plots on the right show the EEG waveform. For illustrative purposes, in the EEG plots all continuous variables are categorized and the EEG waveform was additionally filtered with 15Hz low pass filter.

In contrast to meaningful gestures, beat gestures showed a different pattern (Figure 5). We found a significant main effect of beat gestures (β=-0.005, p=.001). Words with beats gestures elicited a more negative N400. There was a significant interaction such that this increased negativity was especially present for high surprisal words accompanied by beat gestures (β=-0.012, p<.001).

Interactions among Multimodal Cues

We found a number of significant interactions between mean F0 and other non-linguistic cues (Figure 6). First, there was an interaction between mean F0 and meaningful gesture (β=0.004, p<.001). Words with meaningful gestures showed even less negative amplitude of N400 with increased mean F0. Second, there was an interaction between mean F0 and mouth informativeness (β=0.003, p=.040) in which words with higher mouth informativeness elicited less negative N400 when the pitch was high, but more negative N400 when the pitch was low. This interaction was further mediated by a three way interaction between mean F0, mouth informativeness and surprisal (β=-0.013, p=.011): for words with low F0, low mouth informativeness induced more negative N400 for high surprisal words, while for words with high F0, such effect was reversed. Finally, mean F0 and beat gestures also interacted (β=-0.006, p<.001), but the direction and significance of this effect varies with different measures of prosody; thus, we refrain from any further discussion of the effect.

Figure 7 (panel A) shows the interaction between mouth informativeness and meaningful gesture (β=-0.06, p<.001). Words with meaningful gestures elicited more negative N400 when mouth informativeness was high. This two-way interaction was further mediated by a significant three-way interaction between surprisal, mouth informativeness and meaningful gesture (β=-0.010, p=.030). When words were not accompanied by meaningful gestures, high

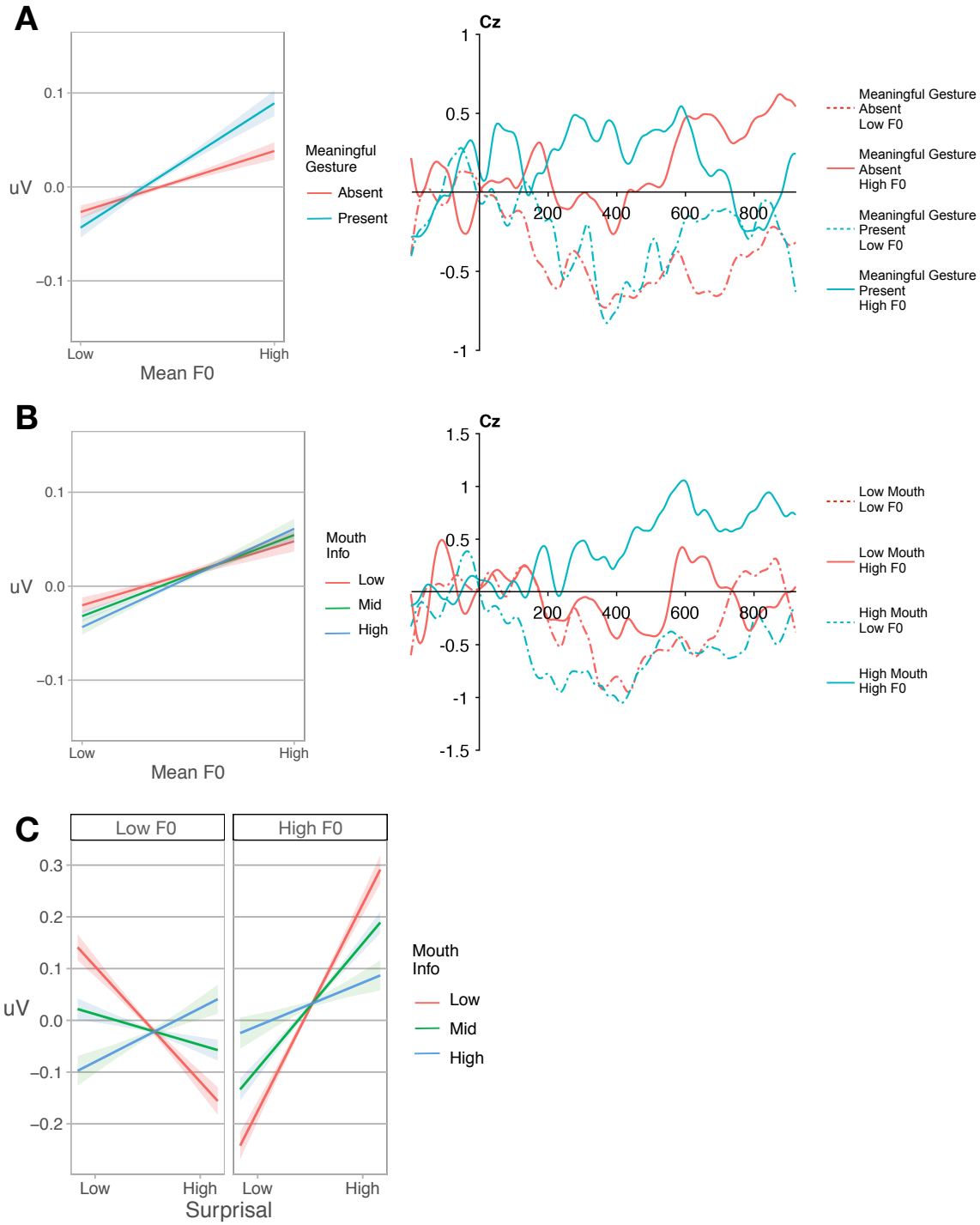

Figure 6. Interactions between Prosodic Stress and other Cues. (A) Interaction between Prosodic Stress and Meaningful Gestures. (B) Interaction between Prosodic Stress and Mouth Informativeness. (C) Interaction between Prosodic Stress, Surprisal and Mouth Informativeness. Plots on the left depict the predicted value of expected value of the mean amplitude of the event-related potentials (ERPs) within 300-600ms (grey areas = confidence intervals). Plots on the right show the EEG waveform. For illustrative purposes, in the EEG plots all continuous variables are categorized and the EEG waveform was additionally filtered with 15Hz low pass filter.

surprisal words with low mouth informativeness elicited more negative N400. However, when words were accompanied by meaningful gestures, high surprisal words with low mouth informativeness elicited more positive N400 (Figure 7, Panel B).

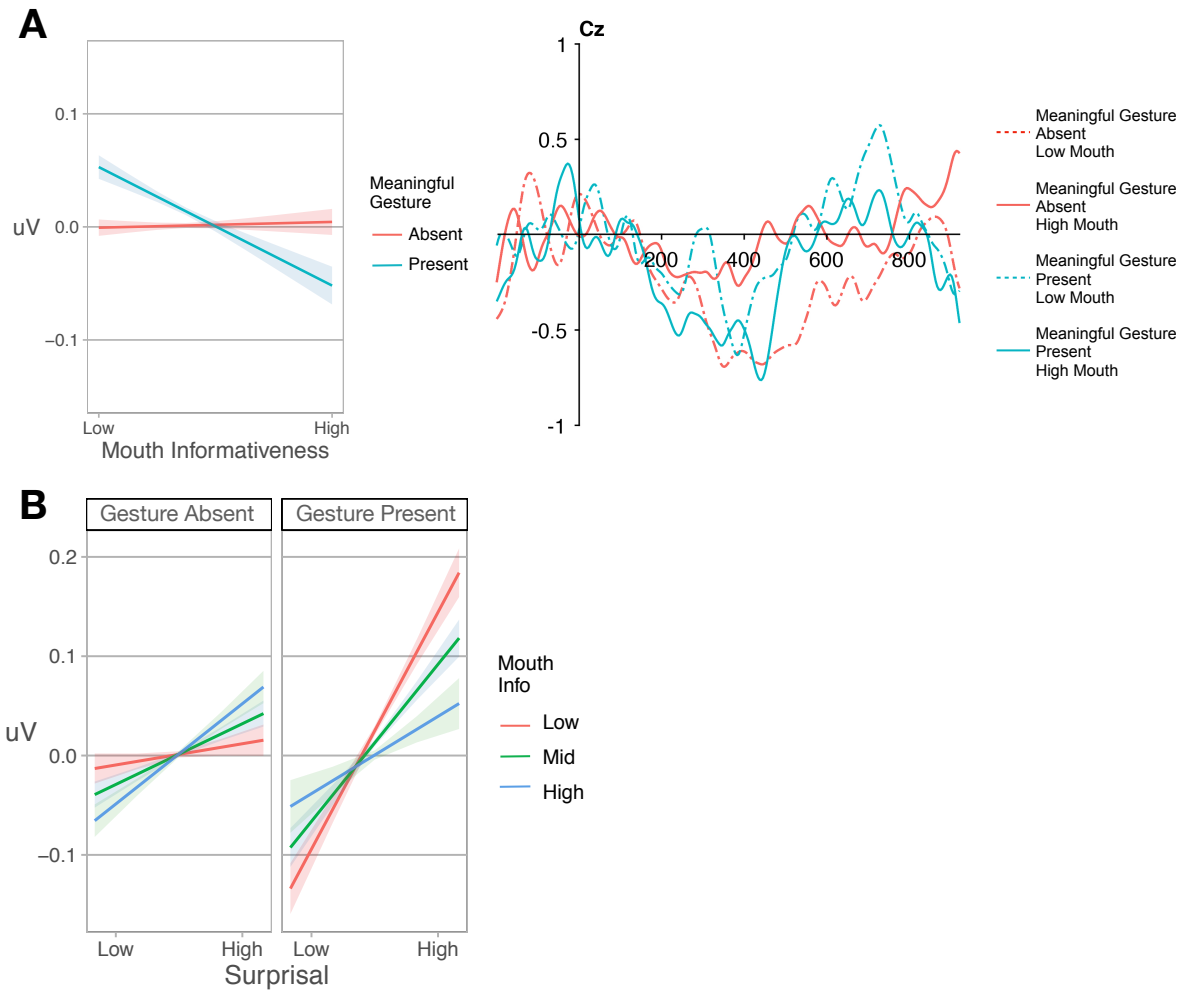

Figure 7. Interactions between Mouth Informativeness and other Cues. (A) Interaction between Mouth Informativeness and Meaningful Gestures. (B) Interaction between Mouth Informativeness, Surprisal and Meaningful Gestures. Plots on the left depict the predicted value of expected value of the mean amplitude of the event-related potentials (ERPs) within 300-600ms (grey areas = confidence intervals). Plots on the right show the EEG waveform. For illustrative purposes, in the EEG plots all continuous variables are categorized and the EEG waveform was additionally filtered with 15Hz low pass filter.

Discussion

In line with our predictions, we confirmed the N400 as a biomarker of prediction during naturalistic language comprehension. High surprisal words elicited longer RTs and lower accuracy behaviourally and a more negative N400 between 300 and 600ms post stimulus, strongest in posterior electrodes. Crucially, this effect was modulated by prosodic stress and meaningful gestures (which reduced the N400 amplitude overall, especially for high surprisal words). Finally, the facilitatory effect of meaningful gesture was reduced for words with high mouth informativeness, especially for high surprisal words.

We also found a number of interactions among the cues themselves. First, the presence of beat gestures increased the N400 amplitude overall, but especially for high surprisal words. Second, the facilitatory effect of meaningful gestures was greater for stressed words. Similarly, a facilitatory effect of mouth informativeness was only observed when words were stressed and only for high surprisal words.

Overall, these results show that the predictability of words based on linguistic context is *always* modulated by multimodal cues, and that the weight given to each cue depends on which other informative cues are present.

Multimodal Cues and Word Predictability

Prosodic stress is considered to mark 'newness' (Cutler, Dahan & Van Donselaar, 1997), as speakers are more likely to stress a word if it conveys new information (Cruttenden, 2006). Previous studies have shown that unstressed new words elicit more negative N400 (Bögels et al., 2011; Heim & Alter, 2006; Li et al., 2008; Magne et al, 2005; Schumacher and Baumann, 2010). Our findings complement previous work in showing that presence of prosodic stress

for less predictable words reduces the amplitude of the N400 suggesting that prosodic stress can enhance expectation for lower probability continuations.

Prosodic stress had a pervasive effect in our study: it interacted with gestures and mouth informativeness in addition to surprisal. The N400 reduction observed for meaningful gestures and mouth informativeness was enhanced for words carrying prosodic stress; moreover, an N400 reduction for high surprisal words with high mouth informativeness was only observed for words carrying prosodic stress. This global effect is consistent with the claims by Kristensen, Wang, Petersson and Hagoort (2013) according to whom, prosodic stress engages a domain general attention network. Thus, stress may draw attention to other cues which consequently would be weighted more heavily. Alternatively, or additionally, as argued by Holler & Levinson (2019), listeners are attuned to natural correlations among the cues (e.g., high pitch correlates to larger mouth movements and increased gesture size) and would use cue-bundles for prediction.

We found that meaningful gestures also support processing, especially for high surprisal words. This result is in line with studies that showed N400 reduction for the subordinate meaning of ambiguous words (e.g. “ball” meaning dancing party) in the presence of a corresponding gesture (Holle and Gunter, 2007; Obermeier et al., 2011; Obermeier et al., 2012), and previous work that showed that words produced with incongruent gestures induce larger N400 (see review in Özyürek, 2014). Our results show that gestures play a general role thus not only when semantic processing is difficult (due to incongruence or ambiguity). They provide additional semantic information about upcoming words therefore increasing their predictability.

Crucially, *meaningful* gestures increase the predictability of upcoming words, but not beat gestures. High surprisal words accompanied by beat gestures elicited a larger N400. This effect might be accounted for in terms of beats enhancing the saliency of a specific word (Krahmer and Swerts, 2007), and highlighting its lack of fit into the previous context. Alternatively, it is possible that listeners try to extract meaning from all gestures and integrate it with speech by default, and since beats are not meaningful, integration fails, inducing processing difficulties.

Wang and Chu (2013) failed to find effects of beat gestures within the N400 time window. The reasons for the different results are unclear, however, the lack of any meaningful gestures in their study could have discouraged listeners from paying attention to gestures. Shifts in the weight attributed to different multimodal cues depending upon the specific task used are documented in the literature (e.g., Gunter and Weinbrenner, 2017; Holle and Gunter, 2007; Obermeier et al., 2015).

Importantly, the dissociation between meaningful and beat gestures further allows us to exclude the possibility that the N400 reduction observed (for meaningful gestures but also for prosody) comes about because these multimodal cues share processing resources with speech processing, increasing the probability that less predictable words go unnoticed.

Beats and prosodic stress have been argued to serve the same function in communication, namely, to make specific words more prominent and therefore attract attention to them (Krahmer and Swerts, 2007). Our results provide evidence against such a claim as their electrophysiological correlates dissociated: beat gestures elicited more negative N400 especially for high surprisal words, while prosodic stress elicited less negative N400, especially for high surprisal words (see also Wang and Chu, 2013). Thus, our work supports

the view that only prosodic stress is used as a marker of information status supporting prediction for new words.

We did not find any significant main effect of mouth informativeness or any interactions between mouth and surprisal in the N400 time-window. Mouth movements have long been recognized to facilitate speech perception especially in noise (Sumby and Pollack, 1954), and synchronized audiovisual compared with audio only speech showed reduction of N1/P2 amplitude, indicating easier sensory-level processing (e.g. Pilling, 2009; van Wassenhove, Grant & Poeppel, 2005). However, in our study we focused on 300-600ms after word onset in order to capture the effect of surprisal and we did not consider earlier (100-300ms) time windows. Two previous studies have investigated the impact of mouth movements within the N400 time window. Hernández-Gutiérrez and colleagues (2018) did not find any N400 difference between audiovisual and audio-only speech; while Brunellière and colleagues (2013) found an increase in N400 amplitude for more informative mouth movements. Further research is necessary to clarify these discrepancies, however, our results suggest that mouth informativeness can affect processing in the N400 time window in combination with other cues.

Finally, our results further extend the previous literature by showing that when multiple cues from the same (visual) channel - such as meaningful gestures and informative mouth movements - are present, they can compete for attentional resources such that the facilitatory effect of representational gestures for high surprisal words is reduced when mouth movements are informative.

Toward a Neurobiological Model of Natural Language Use

In probabilistic-based predictive accounts, the N400 is taken as an index of the processing demands associated with low predictability (e.g. Kuperberg and Jager, 2015). Prior to the bottom-up information, a comprehender holds a distribution of probabilistic hypotheses of the upcoming input constructed by combining his/her probabilistic knowledge of events with contextual information. This distribution is updated with new information, and consequently becomes the new prior distribution for the next event (Levy, 2008). Thus, the N400 is linked to the process of updating the distribution of hypotheses: smaller N400 is associated to more accurate prior distributions/predictions (Kuperberg and Jager, 2015). Our work shows that these mechanisms do not operate only on linguistic information but crucially, they weight 'non-linguistic' multimodal cues. Prosodic stress marks low predictability of the upcoming words, thus more attention and larger weights would be assigned to other cues at both semantic (meaningful gestures) and sensory (mouth movement) levels. Meaningful gestures, could directly impact the prior distribution for the next word (see also discussion in Holler & Levinson, 2019).

Conventional dual-stream neurobiological models do not typically concern themselves with face-to-face language and are mostly localised to perisylvian and inferior frontal regions (e.g., Hickok and Poeppel, 2007). As such, they cannot easily accommodate the results here. A better fit are those models in which language comprehension is considered in context and associated with many interconnected networks distributed throughout the whole brain (Hasson et al., 2018; Skipper, 2015). For example, in the Natural Organization of Language and Brain (NOLB) model, each multimodal cue is proposed to be processed in different but partially overlapping sub-networks (Skipper, 2015). Indeed, different sub-networks have been

associated with gestures and mouth movements, with a 'gesture network' being weighted more strongly than a 'mouth network' when gestures are present (Skipper, 2007, 2009). These distributed sub-networks are assumed to actively predict and provide constraints on possible interpretations of the acoustic signal, thus enabling fast and accurate comprehension (e.g., Skipper, van Wassenhove, Nusbaum, and Small, 2007). Our finding of multiple interactions between cues is compatible with this view, thus suggesting that multimodal prediction processes are dynamic, re-weighting each cue based on the status of other cues.

Conclusions

To conclude, our study assessed language processing in the naturalistic multimodal environment for the first time, and provided evidence that multimodal cues constantly and dynamically interact to construct predictions. Our study provides a new, more ecologically valid, way to understand the neurobiology of language, in which multimodal cues are dynamically orchestrated.

References

Alday PM (2019). M/EEG analysis of naturalistic stories: a review from speech to language processing. *Language, Cogniiton and Neuroscience,* 34, 457-473.

Alday PM, Schlesewsky M, Bornkessel-Sclesewsky I (2017). Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: event-related potentials reflect continuous model updates. *eNeuro,* doi: 10.1523/ENEURO.0311-16.2017

Armeni K, Willems R, Frank, SL (2017). Probabilistic language models in cognitive neuroscience: promises and pittfalls. *Neuroscience & Biobehavioural Reviews,* 83, 579-588.

Arnal LH, Wyart V, Giraud A (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience,* 14, 797-801.

Barr DJ, Levy R, Scheepers C, Tily HJ (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, 255-278.

Bates D, Maechler M, Bolker B, Walker S (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014.

Boersma P, Weenink D (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved 8 September 2018 from http://www.praat.org/

Bögels S, Schriefers H, Vonk W, Chwilla D J. (2011). Pitch accents in context: How listeners process accentuation in referential communication. *Neuropsychologia*, *49*(7), 2022-2036

Broderick M P, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803-809.

Brunellière A, Sánchez-García C, Ikumi N, Soto-Faraco S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *International Journal of Psychophysiology*, 89, 136-147.

Chaumon M, Bishop DV, Busch NA (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of neuroscience methods*, *250*, 47-63.

Clark A (2013). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Behavioural and Brain Sciences,* 36, 181-253.

Cruttenden A (2006). The de-accenting of given information: A cognitive universal? In G Bernini, M Schwatrz (Eds) *Pragmatic Organization of Discourse in the Languages of Europe*. De Gruyter.

Cutler A, Dahan D, Van Donselaar W (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, Pt 2/.

Dambacher M, Kliegl R, Hofmann M, Jacobs AM (2006). Frequency and predictability effects on event-related potentials during reading. *Brain research*, *1084*(1), 89-103.

Drijvers L, Ozyurek A (2018). Native language status of the listener modulates the neural integration of speech and iconic gestures in clear and adverse listening conditions. *Brain and language*, *177*, 7-17.

Frank SL, Otten, LJ, Galli G, Vigliocco G (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, 140, 1-11.

Friston, K, Kiebel S (2009). Predictive coding under the free energy principle. *Philosophical Transactions of the Royal Society B,* 364, 1211–1221.

Gunter TC, Weinbrenner, JED (2017). When to take a gesture seriously: on how we use and prioritize communicative cues. *Journal of Cognitive Neuroscience,* 29, 1355-1367.

Hagoort P, Brown CM (2000). ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia,* 38(11), 1518-30.

Hale J (2001). A probabilistic early 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.

Hasson U, Egidi G, Marelli M, Willems R (2018). Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. *Cognition,* 180, 135-157.

Heim S, Alter K (2006). Prosodic pitch accents in language comprehension and production: ERP data and acoustic analyses. *Acta Neurobiologiae Experimentalis*, *66*, 55.

Hernández-Gutiérrez D, Rahman RA, Martín-Loeches, M, Muñoz F, Schacht A, & Sommer W (2018). Does dynamic information about the speaker's face contribute to semantic speech processing? ERP evidence. *Cortex*, *104*, 12-25.

Hickok G, Poeppel D (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, 8, 393–402.

Holle H, Gunter TC (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*, 1175-1192.

Holler J, Levinson SC (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences,* 23, 639-652.

Hubbard AL, Wilson SM, Callan DE, Dapretto M (2009). Giving speech a hand: 807 gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, 30, 1028–37.

Kakouros S, Salminen N, Räsänen O (2018). Making predictable unpredictable with style–Behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech. *Neuropsychologia*, *109*, 181-199.

Kendon A (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kristensen LB, Wang L, Petersson, KM, Hagoort, P (2012). The interface between language and attention: Prosodic focus marking recruits a general attention network in spoken language comprehension. *Cerebral Cortex,* 23, 1836-1848.

Krahmer E, Swerts M (2007). The effects of visual beats on prosodic prominence: acoustic analysis, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396–414.

Kuperberg GR, Jager TF (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31, 32-59,

Kutas M, Federmeier KD (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.

Kutas M, Hillyard SA (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*, 99-116.

Kutas M, Hillyard SA (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161-163.

Li X, Hagoort P, Yang Y (2008). Event-related potential evidence on the influence of accentuation in spoken discourse comprehension in Chinese. *Journal of Cognitive Neuroscience*, 20, 906-915.

Levy R (2008). Expectation-based syntactic comprehension. *Cognition,* 3, 1126-1177.

Lopez-Calderon J, Luck SJ (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, 8, 213.

MacCallum RC, Zhang S, Preacher CJ, Rucker, DD (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods,* 7, 19-40.

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Sciences*, *8*, 204-210.

Magne CC, Astesano A, Lacheret-Dujour M, Morel KA, Besson M (2005). On-line processing of "pop-out" words in spoken French dialogues. *Journal of Cognitive Neuroscience*. 17, 740-756.

Massaro DW, Jesse A (2007). Audiovisual speech perception and word recognition. In G. Gaskell (Ed.) *Oxford Handbook of Psycholinguistics,* 19-35.

McNeill D (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Obermeier C, Dolk T, Gunter, TC (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *cortex*, *48*, 857-870.

Obermeier C, Holle H, Gunter TC (2011). What iconic gesture fragments reveal about gesture–speech integration: When synchrony is lost, memory can help. *Journal of Cognitive Neuroscience*, *23*, 1648-1663.

Özyürek A (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Phil. Trans. R. Soc. B*, *369*, 20130296.

Pernet CR, Chauveau N, Gaspar C, Rousselet GA (2011). LIMO EEG: a toolbox for hierarchical LInear MOdeling of ElectroEncephaloGraphic data. *Computational intelligence and neuroscience*, *2011*, 3.

Pernet CR, Latinus M, Nichols TE, Rousselet GA (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85-93.

Pilling M (2009). Auditory event related potentials (ERPs) in audiovisual speech perception. *Journal of speech, language and hearing research,* 52, 1073-1081.

Rapp S (1995). Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models. An Aligner for German.

Schäfer R (2015). FYI: COW: Free, Large Web Corpora in European Languages. *LINGUIST List*, *26*.

Schumacher PB, Baumann S (2010). Pitch accent type affects the N400 during referential processing. *NeuroReport*, *21*(9), 618-622.

Shannon CE (1949). Communication in the presence of noise. *Proceeding of the IRE*, 37, 10-21.

Skipper, JI (2014). Echoes of the spoken past: how auditory cortex hears context during speech perception. *Philosophical Transactions of the Royal Society B,* 369, 20130297.

Skipper JI (2015). The NOLB model: a model of the natural organization of language in the brain. In RM Willems (Ed) *Cognitive Neuroscience of Natural Language Use*. Cambridge, Cambridge University Press. (p.101-162).

Skipper JI, Goldin-Meadow S, Nusbaum HC, Small SL (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and Language*, 101, 260–277.

Skipper JI, Goldin-Meadow S, Nusbaum HC, Small SL (2009). Gestures orchestrate brain networks for language understanding. *Current Biology*, 19, 661–667.

Skipper JI, van Wassenhove V, Nusbaum HC, Small SL (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17, 2387–2399.

Sloetjes H, Wittenburg P (2008). Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

Smith NJ, Kutas M (2015). Regression-based estimation of ERP waveforms: the rERP framework. *Psychophysiology,* 51, 157-168.

Smith NJ, Kutas, M (2015). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction and practical considerations. *Psychophysiology*, 52, 169-181.

Smith NJ, Levy R (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319.

Studio R (2012). RStudio: integrated development environment for R. *RStudio Inc, Boston, Massachusetts*, 74.

Sumby WH, Pollack, I (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustic Society of America*, 26, 212-215.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

van Wassenhove V, Grant KW, Poeppel D (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America,* 102, 1181-1186.

Wang L, Chu M (2013). The role of beat gesture and pitch accent in semantic processing: an ERP study. Neuropsychologia, 51(13), 2847-2855.

Winsler K, Midgley, KJ, Grainger J, Holcomm PJ (2018). An electrophysiological megastudy of spoken word recognition. *Language, Cognition and Neuroscience*, 12, 1063-1082.