Behavioral/Cognitive

Delta-Band Neural Responses to Individual Words Are Modulated by Sentence Processing

Sophie Slaats,[1,2] Hugo Weissbart,[3] Jan-Mathijs Schoffelen,[3] Antje S. Meyer,[1,3] and Andrea E. Martin[1,3]
[1]Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, The Netherlands, [2]The International Max Planck Research School for Language Sciences, 6525 XD Nijmegen, The Netherlands, and [3]Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 EN Nijmegen, The Netherlands

To understand language, we need to recognize words and combine them into phrases and sentences. During this process, responses to the words themselves are changed. In a step toward understanding how the brain builds sentence structure, the present study concerns the neural readout of this adaptation. We ask whether low-frequency neural readouts associated with words change as a function of being in a sentence. To this end, we analyzed an MEG dataset by Schoffelen et al. (2019) of 102 human participants (51 women) listening to sentences and word lists, the latter lacking any syntactic structure and combinatorial meaning. Using temporal response functions and a cumulative model-fitting approach, we disentangled delta- and theta-band responses to lexical information (word frequency), from responses to sensory and distributional variables. The results suggest that delta-band responses to words are affected by sentence context in time and space, over and above entropy and surprisal. In both conditions, the word frequency response spanned left temporal and posterior frontal areas; however, the response appeared later in word lists than in sentences. In addition, sentence context determined whether inferior frontal areas were responsive to lexical information. In the theta band, the amplitude was larger in the word list condition ~100 milliseconds in right frontal areas. We conclude that low-frequency responses to words are changed by sentential context. The results of this study show how the neural representation of words is affected by structural context and as such provide insight into how the brain instantiates compositionality in language.

Key words: combinatorial processing; lexical processing; sentence comprehension; surprisal; temporal response functions; word frequency

Significance Statement

Human language is unprecedented in its combinatorial capacity: we are capable of producing and understanding sentences we have never heard before. Although the mechanisms underlying this capacity have been described in formal linguistics and cognitive science, how they are implemented in the brain remains to a large extent unknown. A large body of earlier work from the cognitive neuroscientific literature implies a role for delta-band neural activity in the representation of linguistic structure and meaning. In this work, we combine these insights and techniques with findings from psycholinguistics to show that meaning is more than the sum of its parts; the delta-band MEG signal differentially reflects lexical information inside and outside sentence structures.

Introduction

During language comprehension, listeners recognize words, retrieve stored information about them, and use this knowledge to combine the words into phrases and sentences. Psycholinguistic experiments have long shown that the behavioral responses to words change under the influence of the syntactic and sentential context that the words appear in (Marslen-Wilson and Welsh, 1978; Tyler and Wessels, 1983; Katz et al., 1987). In a step toward understanding how the brain builds sentence structure, the present study concerns the neural readout of this process. We ask (1) whether low-frequency neural readouts associated with words systematically change as a function of being or not being in a sentence context and (2) whether neural readouts are modulated by purely lexical properties over and above sensory and distributional

Received May 6, 2022; revised Apr. 17, 2023; accepted Apr. 27, 2023.

Author contributions: S.S., H.W., and A.E.M. designed research; S.S. and H.W. performed research; J.-M.S. contributed unpublished reagents/analytic tools; S.S. and H.W. analyzed data; S.S., A.E.M. and A.S.M. wrote the paper.

A.E.M. was supported by an Independent Max Planck Research Group and a Lise Meitner Research Group "Language and Computation in Neural Systems", by NWO Vidi grant 016.Vidi.188.029 to A.E.M., and by Big Question 5 (to Prof. Dr. Roshan Cools & Dr. Andrea E. Martin) of the Language in Interaction Consortium funded by NWO Gravitation Grant 024.001.006 to Prof. dr. Peter Hagoort. H.W. was supported by NWO Vidi grant 016.Vidi.188.029 to A.E.M. We thank Laurel Brehm for statistical advice; Inge Pasman, Esther de Kerf, Carlijn Herpt, and Dennis Joosen for research assistance; and the members of the Psychology of Language department at the Max Planck Institute for valuable input on earlier versions of this project.

The authors declare no competing financial interests.

Correspondence should be addressed to Sophie Slaats at sophie.slaats@mpi.nl.

https://doi.org/10.1523/JNEUROSCI.0964-22.2023

variables. We do this by contrasting MEG responses to words in sentences with word lists, the latter lacking any syntactic structure or coherent lexical and combinatorial meaning.

In psycholinguistic models, language comprehension is instantiated as a cascaded process in which information can flow bidirectionally (Marslen-Wilson and Welsh, 1978; Martin, 2016, 2020). Put simply, this means that speech sounds cue stored representations of words, and while the next words are being recognized, the retrieved information about words cues representations of phrase and sentence structure. At the same time, the already formed representations of sentences, phrases, and words cue lower-level representations: the information flows in two directions (Schoffelen et al., 2017).

As words are being combined into phrases and sentences, then, responses to words change as a consequence of the top-down information flow. Indeed, a long tradition of research in psycholinguistics has shown that words in sentences are recognized faster than those same words appearing in isolation (Marslen-Wilson and Welsh, 1978; Tyler and Wessels, 1983). This effect is so powerful that it reduces effects of properties of the words themselves, such as word frequency. In isolation, highly frequent words are recognized faster than low-frequency words. In sentence context, this effect tends to be reduced: low-frequency words are recognized faster in sentence context than in isolation, although there is little change in recognition times for the high-frequency words (Schuberth and Eimas, 1977; Simpson et al., 1989).

To gain a full understanding of human sentence comprehension, those in the field currently face the challenge of integrating these findings with knowledge of neural processing. Although previous studies provide insight into the neural correlates of sentence structure (Ding et al., 2016; Meyer et al., 2017; Nelson et al., 2017; Ding et al., 2018; Brennan and Martin, 2020; Kaufeld et al., 2020; Bai et al., 2022; Coopmans et al., 2022; Tavano et al., 2022; ten Oever et al., 2022), much about the process of building these structures remains unknown (ten Oever et al., 2022). Furthermore, although we know that the neural signal is sensitive to lexical information (Brodbeck et al., 2018a,b; Armeni et al., 2019; Weissbart et al., 2020; Heilbron et al., 2021) we do not know how neural responses to words are transformed in the process of comprehension.

In this study, therefore, we aim to add to our understanding of how the brain leverages linguistic information when building sentence structure by finding a neural readout of the context effect on responses to words above and beyond statistical predictability effects as quantified through entropy and surprisal. To this end, we analyzed a published MEG dataset by Schoffelen et al. (2019) of participants listening to sentences and word lists. Despite these conditions being the main experimental manipulation in this open dataset, they have not previously been directly compared. Using temporal response functions (TRFs), we disentangled delta- and theta-band responses to individual words from responses to the speech envelope and word onsets, as well as entropy and surprisal. This method allowed us to model any differences between the conditions that go beyond our difference of interest (structured/unstructured), and, as such, control for them. We compared the responses to individual words between word lists and sentences. Any differences between the lexical responses in these conditions reflect the effect of structure building on the processing of words.

The lexical response was modeled using word frequency. We chose this feature because word frequency is a proxy for the likely familiarity of the listener with the word and relatedly of ease of processing. Any modulation as a consequence of word frequency, therefore, captures the presence of word identity information in the signal. Furthermore, word frequency is unigram; in other words, it does not depend on the context. Therefore, the value corresponding to a given word is the same in a sentence and a word list. Differences between the neural readout of both conditions will therefore be because of the sentence context supplying structure and meaning and not the predictor itself.

We hypothesized that the delta-band responses to word frequency would be different in word lists and sentences as a consequence of the (in)availability of sentence context (Huizeling et al., 2022; Meyer, 2018; Meyer et al., 2020a,b). Studies that investigated the presence of lower-level features in the neural signal as a function of the availability of linguistic information suggest that lower-level features are represented by the delta-band neural signal more reliably when higher-level information is available. For example, mutual information between the speech signal and the neural signal is higher in the presence of structure and meaning (Kaufeld et al., 2020; Coopmans et al., 2022; ten Oever et al., 2022), and the strength of speech tracking is dependent on the listener's knowledge of the language (Molinaro and Lizarazu, 2018; Blanco-Elorrieta et al., 2020) and general comprehension (Keitel et al., 2018). Following these results, we expected a stronger presence of the word frequency response (the lower-level feature) in the sentence condition than in the word list condition (the higher-level information) in the delta band specifically. Theta-band effects tend to be found as a function of acoustic rather than abstract linguistic manipulations (Sohoglu et al., 2012; Molinaro and Lizarazu, 2018; Etard and Reichenbach, 2019; Blanco-Elorrieta et al., 2020). In this study, we expected to observe this distinction between delta- and theta-band activity through an absence of effects in the theta band.

Materials and Methods

To answer our research question, we analyzed a part of the open-access large multimodal MEG dataset (N = 204) Mother of all Unification Studies published by Schoffelen et al. (2019). In addition, we performed two types of control analyses, an analysis of a dataset published by ten Oever et al. (2022) and a set of simulations. Methods for all analyses are described below.

Participants

A total of 102 native speakers of Dutch (51 men, 51 women) with a mean age of 22 years (range, 18–33) were included in this analysis. In this half of the dataset, participants were presented with the stimuli auditorily (as opposed to the other half, where stimuli were presented visually). All participants were right-handed, reported normal hearing, had normal or corrected-to-normal vision, and had no history of neurologic, developmental, or linguistic deficits. All participants provided informed consent, and the study was approved by the local ethics committee (Committee on Research Involving Human Subjects in the Arnhem-Nijmegen region, The Netherlands) and followed guidelines of the Helsinki Declaration. Participants took part in an fMRI and an MEG session, during which they listened to sentences and word lists. Only the MEG data are included in the present study.

Materials

The complete set of stimuli consisted of 360 natural Dutch sentences of 9–15 words (mean, 11.6), with varying syntactic structures, and 360 word lists. To create the word lists, the words from the sentences were scrambled so that more than two consecutive words did not form a coherent fragment. The stimuli were recorded by a female native speaker of Dutch. The sentences were pronounced naturally. The word lists were pronounced with neutral prosody and with a clear pause between each word. The files were recorded in stereo at 44 100 Hz. The sentences had

an average duration of 4.27 s (SD 0.61), and the word lists of 7.67 s (SD 1.04). During the postprocessing, the audio files were low-pass filtered at 8500 Hz and normalized so that all the audio files had the same peak amplitude and peak intensity. In the word list condition, the individual words were spliced together with variable silence between them. This created conditions with different acoustic properties. We address this issue in the sections below, beginning with MEG preprocessing. In both conditions, the transition from silence to speech was ramped at the onset and offset with a rise/fall time of 10 ms. Word onsets and offsets were determined manually for each audio file using Praat software (Boersma and Weenink, 2018).

The stimuli were divided over two sets, A and B. During the MEG session, participants were presented with 120 sentences from set A and 120 word lists from set B (or the reverse). Across participants, all stimuli were presented the same number of times in the sentence and word list condition.

Procedure

Before the task, participants read written instructions and were allowed to ask clarification questions. The experimenter emphasized that the sentences and word lists should be attended to carefully and discouraged attempts to integrate the words in the word list condition. To familiarize the participants with the task, all participants performed a practice block with stimuli not included in the study. During the MEG measurement, the stimuli were presented in 24 blocks, alternating between sentence blocks (each containing five sentences) and word list blocks (each containing five word lists). The starting block type (either sentences or word lists) was randomized across participants. At the start of each block there was a 1500 ms presentation of the block type: zinnen (sentences in Dutch) or woorden (words in Dutch). The intertrial interval was jittered between 3200 and 4200 ms. During this period, an empty screen was presented, followed by a fixation cross.

To ensure participants paid attention to the stimuli, 20% of the trials were followed by a Yes/No question about the content of the preceding sentence/word list. Half the questions on the sentences addressed the content of the sentence (e.g., Did grandma give a cookie to the girl?), whereas the other half and all the questions about the word lists addressed one of the main content words (e.g., Was a grandma mentioned?). Participants answered the question by pressing a button for Yes/No with their left index and middle finger, respectively. Although the tasks were not identical between the conditions, the randomized order of appearance of question types ensured that participants could not approach the sentences any differently from the word lists; any sentence or list trial could be followed by the word monitoring task.

The stimuli were presented via plastic tubes and ear pieces in both ears. The hearing threshold was determined individually for each participant before the experiment, and the stimuli were presented at an intensity of 50 dB above the hearing threshold.

The experiment was run using Presentation software (version 16.0, Neurobehavioral Systems, www.neurobs.com). MEG was continuously recorded with a 275-channel axial gradiometer system (CTF) at a sampling frequency of 1200 Hz (cutoff frequency of the analog antialiasing low-pass filter was 300 Hz). Three head localizer coils were attached to the participant's head (nasion, left- and right ear canals) to determine the position of the head relative to the MEG sensors. The head position was monitored throughout the measurement. If needed, the participant was asked to reposition to correct for head position changes during breaks. The audio signal of the stimuli presented in the scanner was recorded along with the MEG data using an analog-to-digital converter channel.

Structural MRI images for source reconstruction were acquired using a T1-weighted magnetization-prepared rapid gradient echo pulse sequence with the following acquisition parameters: volume TR = 2300 ms, TE = 3.03 ms, flip angle = 8 degrees, 1 slab, slice matrix size = 256 × 256, slice thickness = 1 mm, field of view = 256 mm, isotropic voxel size = 1.0 × 1.0 × 1.0 mm. A vitamin E capsule was placed as a fiducial behind the right ear to allow visual confirmation of left–right consistency.

MEG preprocessing

The MEG data were preprocessed with custom-written MATLAB scripts using the FieldTrip toolbox (Oostenveld et al., 2011; Donders Institute for Brain, Cognition and Behavior, Radboud University, The Netherlands; http://fieldtriptoolbox.org). Before filtering, the data were epoched from audio onset to audio offset. The epochs were baseline corrected and bandpass filtered into the designated frequency band using a windowed-sinc finite impulse response (FIR) filter (15 s data padded), after which they were resampled to 120 Hz for TRF estimation.

The frequency band of interest was defined on the basis of the rate of occurrence of words in the stimuli, the differences in speech–brain coherence between conditions, and the literature (Blanco-Elorrieta et al., 2020; Donhauser and Baillet, 2020; Molinaro and Lizarazu, 2018; Weissbart et al., 2020). The word rate in the word lists was 1.5 Hz (SD 0.1), and in the sentences 2.7 Hz (SD 0.3). To compute speech–brain coherence, we first computed the broadband speech envelope by taking the absolute value of the Hilbert transform of the speech signal, low passing it at 20 Hz, and scaling the output between zero and one. We computed the magnitude squared coherence estimate of the broadband speech envelope and the MEG signal using Welch's method. The differences between word lists and sentences were estimated using a cluster-based permutation test. This revealed three peaks in the low-frequency signal—one between 1 and 3 Hz, one between 4.5 and 7 Hz, and one between 9.5 and 12 Hz (Fig. 1; Lam et al., 2018). On the basis of these clusters and frequency bands analyzed in the literature (Donhauser and Baillet, 2020), we analyzed two frequency windows, delta (0.5–4 Hz) and theta (4–10 Hz). To account for differences in speech–brain coherence that were exclusively because of acoustic differences between the conditions, we included the speech envelope as a predictor in all the models of the data (Fig. 1*B*, modulation spectra). Details of the models are below in Temporal response functions and Stimulus representation.

Source reconstruction

MRI images were coregistered to the MEG headspace coordinate system by aligning the positions of the preauricular points and the nasion MEG coil to the MRI images using the MNE-Python coregistration GUI (Gramfort et al., 2013). For each participant, we reconstructed the cortical surface using the watershed algorithm from FreeSurfer. We created a surface-based source space with oct6 spacing, meaning ~5 mm was between the source points. This generates 4098 sources per hemisphere. We created a single-layer Boundary element model (BEM) model with surface ico downsampling of 5120, from which the lead field was computed. The sources were reconstructed using a scalar Linear-constraint minimum-variance (LCMV) beamformer approach with a unit-noise gain beamformer to deal with depth bias. The data covariance used for computing LCMV filters was whitened using the covariance matrix of resting-state data. The resting-state data were bandpass filtered into the appropriate frequency band (i.e., 0.5–4 Hz for the delta band, and 4–10 Hz for the theta band). After application of the LCMV beamformer filters to the epoched MEG data, the source-localized epochs were morphed to fsaverage for group statistics. These source-localized, morphed epochs were then entered into the pipeline for temporal response function estimation. Source localization failed for 11 participants because of convergence issues for the noise covariance matrix or missing resting-state data (N_{source} = 91).

Temporal response functions

To characterize the effect of linguistic structure and meaning on the neural response, we estimated TRFs with different acoustic and linguistic features. This approach has been used to determine responses to different linguistic features, ranging from the speech envelope and phonemic information (di Liberto et al., 2015; Donhauser and Baillet, 2020) to lexical information (Broderick et al., 2018; Weissbart et al., 2020) and even syntactic embedding (Nelson et al., 2017). The response function of interest here is the response to word frequency as this is a unigram feature and therefore has the same per-word values in both conditions.

The TRFs were estimated using linear regression. We modeled the neural response by convolving the TRF kernel with the stimulus representation signal. In summary, this method reduces to a multivariate multiple

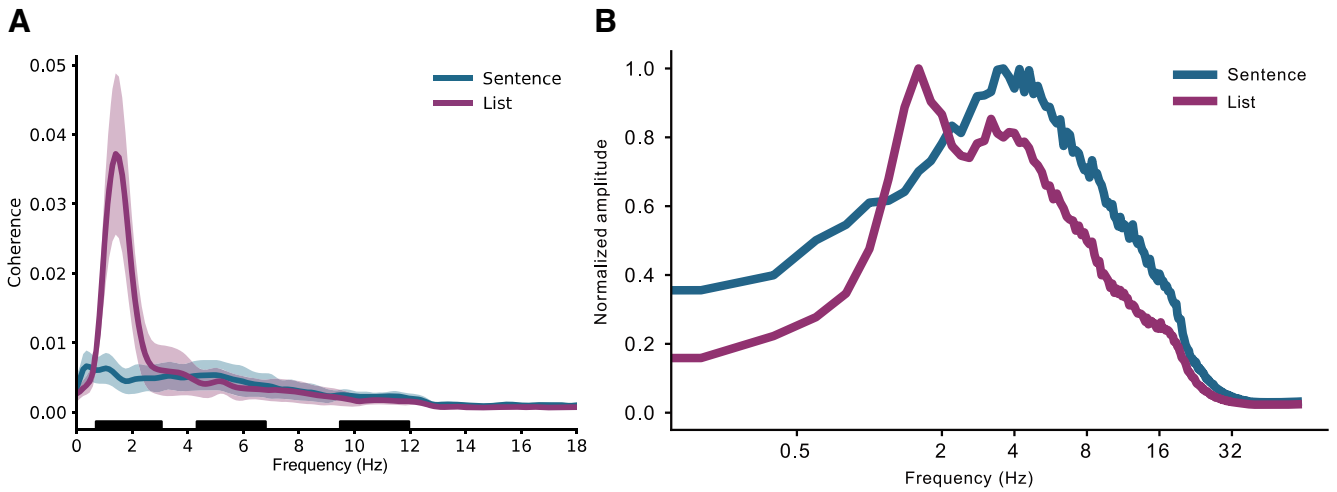

Figure 1. ***A***, Speech–brain coherence. Shaded area indicates SD. Black bars indicate frequencies that were part of clusters that contributed to the significant difference between sentence and word list coherence. ***B***, Modulation spectra of the broadband speech envelopes (part of the TRF base model). The modulation spectra were obtained by concatenating the stimuli per stimulus type and performing a fast Fourier transform on snippets of 5 s. The resulting spectra were averaged.

linear regression, where we used lagged time series of stimulus features as predictors. The model equation reads as follows:

$$y_c(t) = \sum_f \sum_k x_f(t)\beta_f(t-\tau_k) + \eta(t), \quad (1)$$

where $\{y_c\}_t, \{x_f\}_t, \{\beta_f\}_t$ represent the recorded MEG signal of channel c, the input feature f, and its temporal response function respectively; $\{\eta\}_t$ is a Gaussian noise process accounting for measurement noise. This linear model can be easily rewritten in its vectorized form and further concatenated such that we model at once all channel equations independently. We estimate the coefficients of the TRFs $\hat{\beta}_f$ by minimizing the squared error between the measured MEG signals and the reconstructed signal obtained from Equation 1 while keeping the norm of TRFs coefficients, $||\beta||_2$ low to avoid overfitting. This minimization problem is solved in a closed form by the following:

$$\hat{\beta} = (X^T X + \lambda I_d)^{-1} X^T Y, \quad (2)$$

where $Y \in \mathbb{R}^{N\times C}$ is the matrix representation of the measured MEG signal (for C channels arranged columnwise, each with N data samples); $\hat{\beta} \in \mathbb{R}^{(K.F)\times C}$ contains the estimated TRFs with K lags, F features for all C channels; $X \in \mathbb{R}^{N\times(K.F)}$ is a matrix containing all lagged feature time series of length N; λ is a regularization coefficient; and I_d is the identity matrix. The regularization coefficient is needed to avoid overfitting, which in this case translates to the square matrix $X^T X$ not being full rank. Numerically, small eigenvalues or simply ill-conditioned matrices suffice to make the inversion unstable and thus will require regularization. In our case, this happens when features present some amount of autocorrelation (as columns of X are a time-lagged version of other columns). Continuous regressors such as the acoustic envelope (see below, Stimulus representation) will present strong autocorrelation and thus call for regularization.

In Equation 1, the vector of weights $\beta_f(t)$ represents the coefficients parametrizing the temporal response functions. They form a time course reminiscent of an event-related potential that tells us at which point in time (and, potentially, where) a feature modulates the neural signal. Thus, an increase at a certain lag for a given feature reflects an increase in the associated brain response to this feature at that given sensor and at the given time lag after stimulus onset. The concept of stimulus onset, especially for a continuous regressor such as the envelope, here reduces to a situation where the brain would be stimulated by an impulse of sound. Eventually, we estimate, from a system identification perspective, the transfer function mapping input to output when the brain is considered as a linear time-invariant system.

To evaluate how our models perform at reconstructing the neural data, we computed the Pearson's correlation coefficient between the true data and data reconstructed using the estimated TRFs. The correlation between the reconstruction and the original MEG indicates how much of the variance in the neural signal is explained by the features. The TRFs were not estimated on the same portion of data used to score the model. As further explained (see below, Model fitting), we used a nested cross-validation procedure to tune the regularization parameter, estimate the TRF coefficients, and finally score the resulting model. Unless specified otherwise, all analyses described below were done with custom-made Python scripts using MNE-Python (Gramfort et al., 2013). The whole analysis was conducted both in sensor and in source space.

Stimulus representation

Its multivariate character makes the TRF especially suitable for the current analysis: it allows for controlling for differences between conditions that are not currently under discussion by modeling them. To characterize the speech signal and part of its linguistic content, we constructed the following five different features: word frequency (the feature of interest) and four control features consisting of the speech envelope, word onsets, entropy, and surprisal.

The speech envelope feature was computed for each stimulus by taking the absolute value of the Hilbert transform and downsampling it to 120 Hz to match the downsampled MEG sampling rate. The envelope feature was added to represent the acoustic response and as such captures the difference between conditions observed in the cerebro-acoustic coherence that was caused by differences in the acoustic input (Fig. 1*A*,*B*).

The word onset feature was added to capture broadly any time-locked response to word onset for which the variance is not already explained by other features. As such, this feature can also capture any effects of segmentation that were different between the conditions. The word onsets and offsets were transcribed manually for each stimulus. We used a train of unit impulses, where the feature signal is one at the word onset sample and zero otherwise as follows:

$$x(t) = \sum_{words} \delta(t - t_{onset}). \quad (3)$$

These impulse trains were convolved with a Gaussian kernel with an SD of 15 ms. Such temporal smoothing has the effect of inflating the autocorrelation of the signal. We designed the width of this smoothing so that the smoothed impulses end up with energy spanning a frequency band comparable to our continuous regressor (envelope). The Fourier transform of a Gaussian is also a Gaussian, and the 15 ms SD of the temporal smoothing kernel equates to a spectral SD of 21.22 Hz. This ensured

that all features required a similar degree of regularization in the regression analysis and made it possible to include impulse-like features such as word onsets and the envelope in the same regularized regression. Notably, this also translates into some uncertainty about or knowledge of the exact word onset timings.

Like the word onset feature, the word frequency feature was constructed as an impulse train of zeros everywhere but at word onset. Here, we used the respective word frequency value to modulate the height of the impulses. We used the log-transformed value of occurrence per million words, obtained from the SUBTLEX-NL corpus (Keuleers et al., 2010), as follows:

$$x_{wf}(t) = \sum_{words} -\log(P(w)) \times \delta(t - t_{onset}), \quad (4)$$

where $P(w)$ represents the unigram probability estimated from occurrence per million words.

If a word did not exist in the corpus, the fallback value of 0.301 (log/million) was used, corresponding to the lowest word frequency in the corpus. The values were z-scored across all stimuli. The resulting signal was convolved with the same Gaussian kernel as the word onset feature.

The entropy feature consists of lexical entropy, a weighted probability measure that quantifies the uncertainty about the upcoming word on the basis of the previous words. It provides a numeric answer to the question, Given the n previous words, with what degree of certainty can we predict the upcoming word? as follows:

$$H(w_i) = -\sum_k P(w_k|w_{i-1}...w_{i-n}) \times \log_2(P(w_k|w_{i-1}...w_{i-n})). \quad (5)$$

The value was derived from a trigram model trained on the NLCOW2012 corpus using WOPR (van den Bosch and Berck, 2009). If a value was missing, the average of all entropy values was used. Like the word frequency feature, the entropy values were z-scored relative to all stimuli and inserted in a stick function, after which the stick function was convolved with the same Gaussian window. This feature was added to ensure that any effects on the word frequency feature were of a compositional semantic and structural nature rather than a probabilistic one.

The surprisal feature reflects how surprising a given word is in its immediate context. From an information-theoretic perspective, this reflects the information content, or self-information, of a word. It was calculated as the log 10 transformation of the conditional probability of a word, which was taken from the same trigram model as the entropy values. This means that surprisal is always based on the two preceding words; given the two preceding words, how high was the chance that the observed word would indeed appear? If the chance was low, surprisal is high. The feature was constructed in the same way as the word frequency and entropy features; the values were z-scored across all stimuli, inserted in a stick function at word onsets, and convolved with the Gaussian window as follows:

$$I(w) = -\log_{10}(P(w_i|w_{i-1}...w_{i-n})). \quad (6)$$

Because the three numerical lexical features (frequency, entropy, surprisal) might be correlated to some extent, we need to assert that the degree of multicollinearity present in our stimulus representation will not hinder the TRF coefficient interpretation. We checked whether the variance inflation factor (VIF) was below five (considered a relatively conservative measure of multicollinearity; Sheather, 2009; Tomaschek et al., 2018). The VIF was computed by correlating the z-scored entropy, surprisal, and word frequency values and by taking the diagonal of the inverted correlation matrix. This was done for all the stimuli and for both conditions separately. The VIF was never higher than five; the highest VIF was for surprisal at 4.8 in the word list condition.

Model fitting

The features were fitted in a cumulative manner to assess the contribution of each feature. This led to a total of seven models per frequency band, an Envelope model, consisting of only the speech envelope; an Onset model, consisting of the speech envelope and the word onset features; a Frequency model, consisting of the speech envelope, word onset, and word frequency features; an Entropy model, containing the speech envelope, word onset, and entropy features; a Surprisal model, consisting of the speech envelope, word onset, and surprisal features; and cross-combinations of those with and without the word frequency feature. An overview of all models and the corresponding features is provided in Table 1.

Table 1. The fitted encoding models

	Feature				
Model name	Envelope	Word onset	Entropy	Surprisal	Word frequency
Envelope	X				
Onset	X	X			
Entropy	X	X	X		
Surprisal	X	X		X	
Frequency	X	X			X
Entropy/Surprisal	X	X	X	X	
Entropy/Frequency	X	X	X		X
Surprisal/Frequency	X	X		X	X
Full	X	X	X	X	X

X indicates that a feature was included in the model.

Before model fitting, the data were split pseudorandomly into a training and testing set at a 80/20 ratio. Care was taken that the sentences and word lists were evenly divided across the training and test sets. The sentence and word list models were each trained on 96 of 120 trials. The regularization parameter was optimized individually per participant, frequency band, and model (but not per condition) using an eightfold cross-validation procedure with 20 log-spaced values around the eigenvalues of the covariance matrix of the lagged speech envelope ($\lambda = 60470.9$) ranging from $\lambda \times 10^{-3}$ to $\lambda \times 10^{3}$. The best regularization parameter was determined as the value for which the average (across sensors) reconstruction accuracies were highest. Occasionally, reconstruction accuracies would not increase with a higher degree of regularization; instead, increasing the regularization would leave the reconstruction accuracy at the same value until overregularization occurred and reconstruction accuracy went down. In this case, the highest lambda value before a drop in accuracy occurred was chosen to ensure some degree of regularization. Each model was fitted on the complete training set using the regularization parameter from the cross-validation procedure, yielding the TRFs.

In the analysis of the source-localized MEG data, the manipulations were simplified because of computational limitations. The two maximal models were fitted, with word frequency as the only difference as follows: the Entropy/Surprisal model, consisting of the speech envelope, word onsets, entropy, and surprisal features; and the full model, consisting of all features. The cross-validation procedure was brought down to five-fold with 10 log-spaced values around the eigenvalue of the stimuli (60470.9) ranging from $\lambda \times 10^{-2}$ to $\lambda \times 10^{2}$.

Model evaluation

Each model was evaluated by convolving the estimated TRFs with the unseen stimuli from the test dataset. This yields, in essence, a prediction of the neural signal according to the model. The predicted neural signal was then correlated with the original neural signal from the test set using the Pearson product-moment correlation on a sensor-by-sensor or source-by-source basis. For every individual participant, this yielded a set of sensor- or source-based reconstruction accuracies for each model.

Statistical analysis

The TRF analysis has two deliverables. First, the TRF (the development of the estimated coefficients across time) is an ERP-like waveform that captures how the neural signal changes as a function of, for example, word frequency, and, second, the reconstruction accuracy, which is a metric of model fit. Here, we wanted to know (1) whether the responses to word frequency differ between sentences and word lists in time and

space, so we compared the TRFs between conditions, and (2) whether the presence of the word frequency response differed between sentences and word lists, so we tested whether the word frequency predictor contributed differently to the reconstruction accuracy of a model in the two conditions.

Throughout, evaluation for statistical significance of the difference between TRFs was done using cluster-based permutation tests. Cluster-based permutation tests address the null hypothesis of exchangeability across conditions by a Monte Carlo estimate of the randomization distribution of a cluster-based test statistic, optimizing statistical sensitivity while controlling the false alarm rate. Here, we used the t statistic as the test statistic. In these tests, we create matrices of all sensors and samples. Then, we compute the difference between two conditions and express it as a *t* statistic for each of these data points. The *t* values are thresholded at an a priori threshold, and the thresholded *t* values are summed across clusters on the basis of spatial and temporal adjacency. The significance of the test statistic of the resulting largest cluster is compared with 1024 of similarly obtained test statistics, after random permutation of the condition labels. We used the function spatio_temporal_cluster_test from the MNE-Python library (Gramfort et al., 2013) with the *t* statistic as the test statistic and 1024 permutations.

To assess whether the responses to word frequency differed qualitatively between conditions in sensor space, the difference between the word frequency TRFs for the sentence and word list conditions was evaluated using a cluster-based permutation test. In addition, to characterize the response in each condition separately, we performed two cluster-based permutation tests with the same methods in which we contrasted the response against zero in each condition separately. In total, we performed three cluster-based permutation tests on the sensor TRFs, one on the difference between conditions and one on the TRF for each of the two conditions separately (against zero). In all cases, we calculated the threshold on the basis of the *t* distribution with a significance level of 5×10^{-8} with 101 (number of participants minus one) degrees of freedom. This equals three times the recommended threshold for the number of participants. The threshold was increased to yield the most informative results (i.e., to ensure not every sensor and time lag would be significant). Subsequent comparisons were done with a threshold calculated using a Bonferroni adjusted significance level (i.e., divided by two) to correct for multiple comparisons; everything else was the same.

In addition, we wanted to evaluate whether there was a latency difference between the responses in the two conditions. To this end, we compared the responses from the sentences and word list conditions in a cross-correlation. The cross-correlation was done on the grand-average TRF waveforms of overlapping sensors between conditions from the clusters resulting from the one-sample tests. We sequentially cross-correlated each sensor and normalized the values by dividing them by the maximal value from the cross-correlation for that sensor. We then obtained the peaks for every sensor. This number corresponds to the lag at which the two signals had the highest correlation and shows how different the responses are in time. Subsequently, we shifted the sentence response in time by the number of samples of the peak. We then correlated the shifted sentence response and the original word list response. To check for significance, we performed the same procedure for randomly selected channels and repeated this process 10,000 times.

In source space, we compared the TRFs for word lists and sentences using a cluster-based permutation test in two time windows on the basis of the results from the analysis in sensor space, 200–400 and 500–700 ms post stimulus onset (PSO), respectively. We did this to get a more reliable estimate of the spatial distribution of the effects, although cluster-based permutation tests account only for a difference between the distribution overall, therefore any spatial or temporal differences are approximations and inconclusive (Maris and Oostenveld, 2007; Sassenhagen and Draschkow, 2019). The threshold was set to the *t* distribution with an alpha of 0.025 (98.75th and 1.25th percentile) to correct for multiple comparisons, with 90 (number of participants minus one) degrees of freedom. Sources along the medial wall were excluded.

In the sensor space analysis, the reconstruction accuracies were averaged over sensors and submitted to a linear mixed model using lme4 in R software (Bates et al., 2015). The model had the factor *condition* (two levels, sentence and word list) and a random intercept for *participant*. In addition, the model contained three binomial factors, *frequency*, *entropy*, and *surprisal*, describing whether a feature was (1) or was not (0) in the model to calculate a slope for each feature separately as follows:

$$\text{accuracies} \sim \text{condition} * (\text{frequency} + \text{entropy} + \text{surprisal}) + (1|\text{participant}).$$

We used a stepwise variable selection to evaluate the contribution of each of these factors. To evaluate the contribution of a given factor (or interaction), a model with the factor was compared with a model without it, and the goodness-of-fit statistics were compared using a chi-square test. If the removal of a factor did not decrease goodness of fit, the next factor was removed. When the removal of a given feature or interaction significantly decreased model fit, the removal of features was stopped. The prefinal model should then describe the data best. As a final check, the Akaike information criterion (AIC) of the models was compared using the R package AICcmodavg (Mazerolle, 2020). *Post hoc t* tests were done between the Entropy/Surprisal and Full model to evaluate whether the effects held between the largest models.

In source space, a cluster-based permutation test was done to localize the interaction effect using the function permutation_cluster_test from the MNE-Python library. The test statistic was an *F* statistic from a two-way ANOVA with factors Condition (levels: word list, sentence) and Model (levels: Entropy/Surprisal, Full). The data were permuted 1024 times.

Control analysis I: data

The word lists were presented with variable silences between words. The sentences, on the other hand, were natural, with pauses occurring sparingly. This caused differences of word rate and signal length between the conditions that may affect our results. To examine potential effects of the pauses in the word list condition, we analyzed a second dataset of 16 participants listening to word lists and sentences using the same methods. Importantly, the word lists in this condition were naturally spoken, as were the sentences. This means that there were no pauses between the words in the word list condition, and there was coarticulation between words (Kaufeld et al., 2020). The data were supplied by ten Oever et al. (2022).

Control analysis I: Participants. A total of 20 native speakers of Dutch (4 men, 16 women with a mean age of 39.5 years) participated in the experiment. Four participants were excluded from this analysis for a variety of reasons (e.g., session was not completed). All participants were right-handed, reported normal hearing, had normal or corrected-to-normal vision, and had no history of neurologic, developmental, or linguistic deficits. All participants provided informed consent. The study was approved by the ethical Commission for Human Research Arnhem/Nijmegen (project number CMO2014/288). Participants were remunerated for their participation.

Control analysis I: Materials. The stimuli were identical to the stimuli used in Kaufeld et al. (2020). The experiment consisted of three conditions in total, sentences, jabberwocky, and word lists. Only the sentences and the word lists are analyzed here. The stimuli consisted of 10 words, which were all disyllabic except for de (the in Dutch) and en (and in Dutch). Sentences had a fixed syntactic structure of two coordinate clauses: [Adj N V N conj Det Adj N V N], for example, timid heroes pluck flowers and the brown birds gather branches. The word lists were scrambled versions of these sentences, and care was taken so there were no plausible internal combinations of words. The stimuli were recorded by a female native speaker of Dutch at a sampling rate of 44.1 kHz (monophonic). After recording, any pauses were normalized to ~150 ms in all stimuli, and the intensity was scaled to 70 dB using Praat voice analysis software (Boersma and Weenink, 2018).

Participants were asked to perform four different tasks on these stimuli—a passive listening task, a syllable recognition task, a word recognition task, and a word combination recognition task. In this analysis, we did not distinguish among tasks. ten Oever et al. (2022) describes the tasks performed.

Control analysis I: Procedure. At the beginning of each trial, participants were instructed to look at a fixation cross presented at the middle of the screen on a gray background. The audio was presented binaurally through tubes after an interval randomly jittered between 1.5 and 3 s. One second after audio offset, the task prompt (e.g., the syllables or words for recognition) was presented, which required participants to press a button on a button box. There were eight blocks of ~8 min. After each block, participants could take a break, during which the head position was corrected. MEG was recorded using a 275-channel axial gradiometer CTF MEG system at a sampling rate of 1200 Hz. After the session, head shape was collected using the Polhemus digitizer (using as fiducials the nasion and the entrance of the ear canals as positioned with ear molds).

Control analysis I: MEG preprocessing. The MEG data were processed with custom-written Python scripts using MNE-Python (Gramfort et al., 2013). As in the main analysis, the raw MEG data were filtered using a windowed-sinc FIR filter between 0.5 and 4 Hz for the delta band, and 4 and 10 Hz for the theta band, after which the data were epoched from audio onset to audio offset and resampled to 120 Hz for TRF estimation.

Control analysis I: Stimulus representation. In this analysis, we used the envelope, word onset, and word frequency representations from the main analysis (see above, Stimulus representation).

Control analysis I: Model fitting. We used the model-fitting approach described earlier (see above, Model fitting). We fit three models, Envelope (with only the envelope feature), Onset (envelope and word onset features), and Frequency (envelope, word onset, and word frequency features). The data were split pseudorandomly into a training and a testing set at an 80:20 ratio, ensuring that the sets contained 50% of items from each condition. The regularization parameter was optimized individually per participant and model, using an eightfold nested cross-validation procedure with 20 log-spaced values around 60,000 ($\lambda = 60{,}000$) ranging from $\lambda \times 10^{-2}$ to $\lambda \times 10^{2}$.

Control analysis I: Model evaluation. For model evaluation, we used the procedure described earlier (see above, Model evaluation).

Control analysis I: Statistical analysis. Like in the main analysis, we assessed whether the responses to word frequency qualitatively differed between conditions by evaluating the difference between the word frequency TRFs for the sentence and word list conditions using a cluster-based permutation test. In addition, to characterize the response in each of the conditions separately, we performed two additional cluster-based permutation tests with the same methods in which we contrasted the response against zero in each condition separately. In total, we performed three cluster-based permutation tests on the TRFs, one on the difference between conditions and one on the TRF for each condition separately (against zero). In all tests, we calculated the threshold on the basis of the t distribution with a significance level of 0.05 with 16 (number of participants minus one) degrees of freedom. Only clusters with a p value smaller than 0.01 were considered. Subsequent comparisons were done with a threshold calculated using a Bonferroni-adjusted significance level to correct for multiple comparisons; everything else was the same. For comparison to the main analysis, we also compared the word onset response between conditions with the methods described above.

To evaluate the effect of word frequency in each condition, we compared the reconstruction accuracies from the Onset and Frequency models in interaction with condition. The reconstruction accuracies were averaged over all sensors (conservative measure). After checking for normality and sphericity through (1) visual inspection of Q-Q plots and histograms; (2) statistical testing using the Shapiro–Wilk test, Anderson–Darling test, and D'Agostino's K^2 test for kurtosis and skewness as implemented in SciPy algorithms; and (3) the Mauchly test for sphericity as implemented in the Pingouin package (Vallat, 2018), the averaged reconstruction accuracy values were submitted to a repeated-measures ANOVA using the Statsmodels package.

Control analysis II: simulations

Using simulations, we evaluated whether the interword interval has an impact on TRF model evaluation. We did this by simulating raw MEG data consisting of a signal (different impulse responses) and a variable amount of noise.

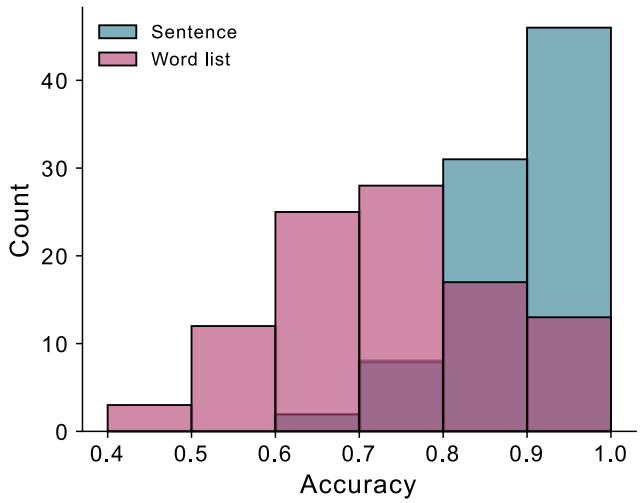

Figure 2. Accuracy scores for the behavioral task performed during the MEG recording. The accuracy scores include responses to word monitoring only. The word list accuracy scores are a random subset of the full set of responses to balance the number of trials ($n = 12$) in the word list and sentence conditions.

The simulated response was equivalent to the forward model, namely a noisy output of a convolution between a predefined kernel (the ground truth for the TRF estimate) and an impulse train (for the input signal). We generated those data with a variable amount of noise (i.e., explicitly manipulating the broadband signal-to-noise ratio) and with varying the interstimulus interval (ISI) while keeping the signal length the same and the number of impulses, or events, constant (in which case a shorter interstimulus interval results in the end portion of the output signal containing only noise).

We then scored the forward model by computing both the R^2 score and the Pearson's correlation coefficient between the reconstruction $\hat{y}$ and the true signal using a test portion of the data, not used to estimate the coefficients β. Importantly, we then computed the scores in two ways, (1) from the fixed signal length data described above, as we also used a fixed number of impulses, or events, this resulted in a portion of the stimulated output signal to contain only noise (2) or from a shortened signal, where we truncated all signals to the last stimulus event. This resulted in shorter signals for a shorter ISI.

Data availability

The code is available at https://osf.io/ky9bj/, with the exception of the preprocessing scripts. The preprocessed data are available on request. The raw data can be downloaded from the Donders Institute repository at https://data.donders.ru.nl/collections/di/dccn/DSC_3011020.09_236?0.

Results

Behavioral results

We compared participants' responses to the task that was present in both conditions, which targeted one of the main content words (e.g., Was a grandma mentioned?). To balance the number of trials included in the accuracy scores, we took a random subset of questions from the word lists (12 or 13 trials). The average proportion of correct responses was higher in the sentence condition ($\text{mean}_{\text{sent}} = 0.88$; $\text{sd}_{\text{sent}} = 0.08$) than in the word lists ($\text{mean}_{\text{list}} = 0.72$; $\text{sd}_{\text{list}} = 0.14$; $t = 10.08$, $p < 0.001$), meaning that participants remembered the words from the sentences better than the words from the word lists (Fig. 2).

Delta band

Sensor-level analysis

The cluster-based permutation test revealed differences between word lists and sentences in three clusters between 0 and 700 ms.

Figure 3*A* suggests that the peak of the response to word frequency was delayed by ~300 ms in the word list condition. To evaluate whether this was the case, we conducted one-sample cluster-based permutation tests and computed the cross-correlation between the two conditions for overlapping sensors from the clusters in both conditions. The one-sample cluster-based permutation test revealed a response in temporal areas in both conditions that peaks ~250 ms in the sentence condition, and ~600 ms in the word list condition (Fig. 3*B*,*C*).

The cross-correlation on overlapping sensors between the two conditions (time courses and sensors; Fig. 4*A*) revealed a high correlation between the word list and the sentence responses at a delay of 330 ms (mean r = 0.9). Random sampling of sensors and lags revealed the distribution shown in Figure 4*D*; the observed values are in the upper 0.05% percentile, indicating that the observed correlation is likely not caused by chance.

Because we wondered whether the delay could be because of the differences in the presentation rate, we examined differences between the TRFs for the other word-level feature that was numerically identical between conditions, word onsets (unit-spike-train in both conditions). We compared the word onset response from a model with only the envelope and word onset features. This model is equivalent to an ERP analysis that corrects for overlapping event windows (as is the case in the sentence condition) and controls for acoustic differences. A small delay of ~100 ms appears in this model. This delay is in accordance with findings of an ERP-analysis on high- versus low-constraining contexts (Liu et al., 2006; León-Cabrera et al., 2017). Importantly, this model collapses over variance caused by the lexical features included in the full model (word frequency, entropy, and surprisal). In other words, this underspecified model attributes variance that is in fact because of word frequency, entropy, or surprisal to the word onset predictor. When we include the other lexical predictors in the model and compared the conditions again, no such difference between the word onset responses is observed (Fig. 3*D*). In this response, there were some differences around time point zero before as well as slightly after; these differences may indicate differences in temporal expectancy of word onset between conditions.

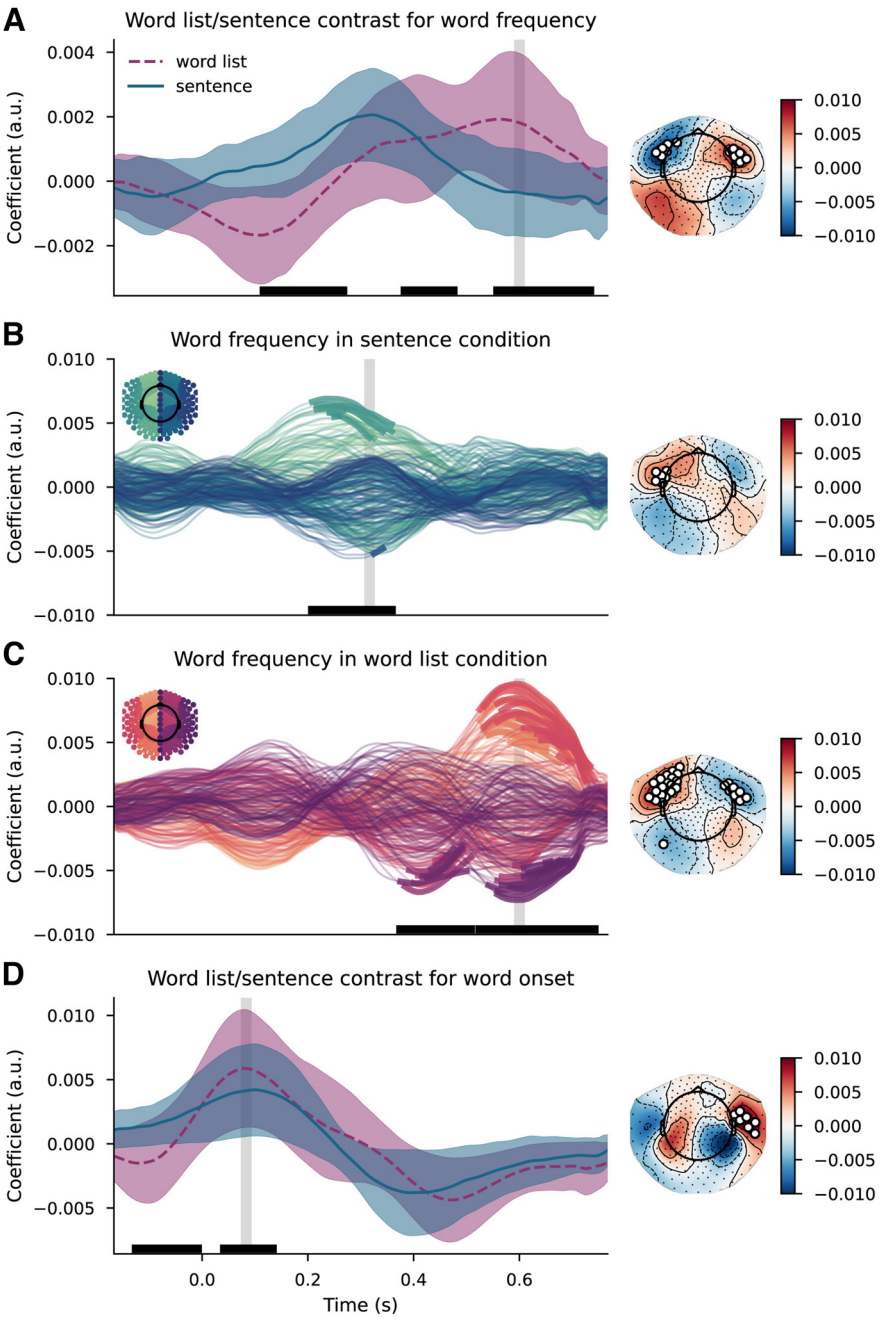

Figure 3. ***A***, The word frequency TRF in both conditions in the delta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points that contributed to clusters that allowed us to reject the null hypothesis. Shaded area indicates SD. ***B***, Word frequency TRF in the sentence condition. Individual lines represent sensors. Sensors in bold contributed to the clusters that allowed us to reject the null hypothesis. ***C***, Word frequency TRF in the list condition. Individual lines represent sensors. Sensors in bold contributed to the clusters that allowed us to reject the null hypothesis. ***D***, The word onset TRF in both conditions in the delta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points that contributed to clusters that allowed us to reject the null hypothesis. Shaded area indicates SD. Vertical gray lines indicate the time points of the scalp maps.

The reconstruction accuracies were evaluated with the model *accuracies* ~ *condition* * (*frequency* + *entropy* + *surprisal*) + (*1*/*participant*). The explanatory value of the interaction between condition and each of the lexical factors was evaluated; each interaction significantly improved model fit [frequency, $\chi^2(1) = 6.88$, $p < 0.01$; entropy, $\chi^2(1) = 4.48$, $p < 0.05$; surprisal, $\chi^2(1) = 7.24$, $p < 0.01$], so the full model was interpreted. The results of this model are summarized in Table 2.

Reconstruction accuracies were higher in the word list condition than in the sentence condition ($\beta = 1.67 * 10^{-2}$, SE = $9.43 * 10^{-4}$, $t_{(1530)} = 17.69$, $p < 0.01$). As can be seen in Figure 5*A*, each

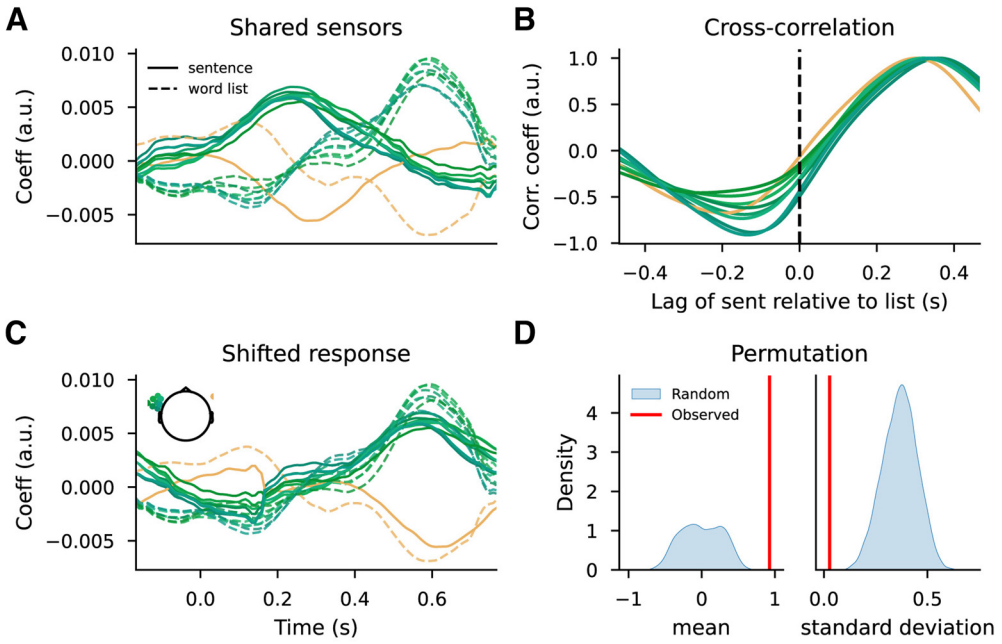

Figure 4. ***A***, TRF time courses for shared sensors between the sentence (solid lines) and word list (dashed lines). Colors indicate sensor position. ***B***, Cross-correlation between the sentence and word list responses for overlapping sensors between conditions from the clusters (scaled between −1 and 1). Colors indicate sensor position. ***C***, The shifted response from the sentence condition (solid lines) to overlap with the word list condition (dashed lines). Colors indicate sensor position. ***D***, Kernel density plots of means and SDs from correlations between randomly selected sensors at shifted randomly selected lags; the red bar indicates the values observed from the sensors selected after the cluster-based permutation test shifted at the lags from the cross-correlation. Coeff: coefficient.

Table 2. Results of the LME on the reconstruction accuracies in the delta band

Factor	β coefficient	SE	df	*t* value	*p* value
(Intercept)	8.61×10^{-2}	1.82×10^{-3}	1306	47.22	***
Word frequency	3.61×10^{-4}	6.66×10^{-4}	1530	0.54	n.s.
Surprisal	6.24×10^{-4}	6.66×10^{-4}	1530	0.94	n.s.
Entropy	-3.88×10^{-4}	6.66×10^{-4}	1530	−0.58	n.s.
Condition	-1.67×10^{-2}	9.43×10^{-4}	1530	−17.69	***
Word frequency * condition	2.47×10^{-3}	9.43×10^{-4}	1530	2.63	**
Surprisal * condition	2.54×10^{-3}	9.43×10^{-4}	1530	2.69	**
Entropy * condition	2.00×10^{-3}	9.43×10^{-4}	1530	2.12	*

SE: standard error, df: degrees of freedom, n.s. not significant, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 3. Results of the LME on the reconstruction accuracies in the theta band

Factor	β coefficient	SE	df	*t* value	*p* value
(Intercept)	4.02×10^{-2}	1.29×10^{-3}	1382	31.04	***
Word frequency	1.17×10^{-3}	5.64×10^{-4}	1530	2.07	*
Surprisal	7.26×10^{-4}	5.64×10^{-4}	1530	1.29	n.s.
Entropy	2.43×10^{-3}	3.99×10^{-4}	1530	6.10	***
Condition	-2.09×10^{-3}	6.90×10^{-4}	1530	−3.02	**
Word frequency * condition	1.55×10^{-3}	7.97×10^{-4}	1530	1.95	n.s.
Surprisal * condition	1.59×10^{-3}	7.97×10^{-4}	1530	1.99	*
Entropy * condition					

The factor Entropy * condition is not included in the model that was interpreted; SE: standard error, df: degrees of freedom, n.s. not significant, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

feature contributed positively to the reconstruction of the neural signal in the sentence condition, less so in the word list condition, hinting at an interaction effect. Indeed, the factor *frequency* interacted with *condition* ($\beta = 2.47 * 10^{-3}$, SE = $9.43 * 10^{-4}$, $t_{(1530)} = 2.63$, $p < 0.01$), showing that reconstruction accuracies improved more from the addition of the word frequency predictor in the sentence condition than in the list condition (Fig. 5*B*). Further, although we do not discuss these effects, *entropy* and *surprisal* interacted with *condition* as well (entropy, $\beta = 2.00 * 10^{-3}$, SE = $9.43 * 10^{-4}$, $t_{(1530)} = 2.12$, $p < 0.05$; surprisal, $\beta = 2.54 * 10^{-3}$, SE = $9.43 * 10^{-4}$, $t_{(1530)} = 2.69$, $p < 0.01$).

To gain more insight into the effect of *frequency*, we performed a *post hoc t* test comparing the two largest models (Entropy/Surprisal and Full). These tests confirmed that the word frequency predictor enhanced reconstruction accuracy in the sentence condition ($t_{(101)} = 5.35$; $p < 0.01$), but not in the word list condition ($t_{(101)} = -0.15$, $p = 1$; Bonferroni corrected).

Finally, we hypothesized that the higher reconstruction accuracy in the word list condition was because of the salience of isolated words, possibly evoking a larger auditory response. If this is true, a model with only the envelope predictor, and no word-level feature, should also fit the list condition better. To evaluate this hypothesis, we compared the reconstruction accuracies (averaged over all sensors) for the Envelope model between conditions. This model was not included in the analyses of the word frequency effect. And indeed, this was the case; reconstruction accuracies were higher for word lists than sentences using only the envelope as predictor ($t_{(101)} = 13.40$, $p < 0.01$).

In sum, the response to word frequency differed between word lists and sentences. The TRFs in sensor space revealed a left-lateralized frontotemporal response to the feature that peaked ~250 ms after word onset in the sentence condition, and ~600 ms in the word list condition. The sentence effect is in line with other studies that used word frequency as a feature in TRF models of natural language comprehension (Brennan and Hale, 2019; Weissbart et al., 2020). A cross-correlation analysis between a set of left (and one right) temporal and frontal sensors that were involved in the response in both conditions suggested that the word list response peaks ~300 ms later. The reconstruction accuracies in sensor space suggests that the word frequency predictor explains more variance over and above acoustics, entropy, and surprisal in the sentence condition, but not in the word list condition.

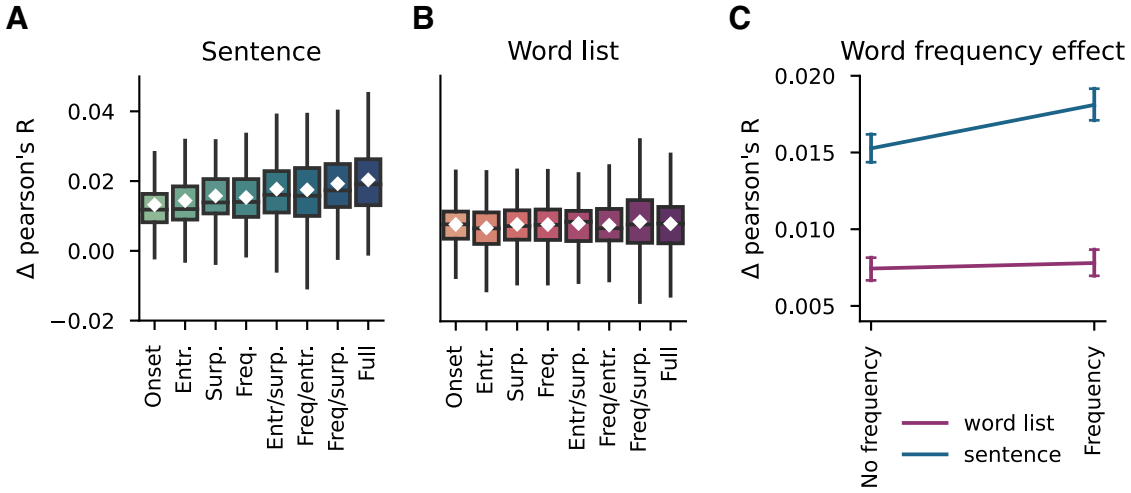

Figure 5. Reconstruction accuracies in the delta band. ***A***, Reconstruction accuracy difference with the envelope model for each model in the sentence condition. Middle line indicates the median, the white diamond indicates the mean. ***B***, Reconstruction accuracy difference with the envelope model for each model in the word list condition. Middle line indicates the median, the white diamond indicates the mean. ***C***, The interaction between condition and frequency on the reconstruction accuracies. Values on the *y*-axis are the difference with the envelope (as in ***A***, ***B***). Error bars represent the 95% confidence interval. Entr.: entropy, surp: surprisal, freq: frequency.

Source reconstruction

In source space, we compared the TRFs for word lists and sentences using a cluster-based permutation test in two time windows on the basis of the results from the analysis in sensor space, 200–400 and 500–700 ms post stimulus onset, respectively. The cluster-based permutation test on the TRFs from the source reconstructed MEG revealed two clusters in the early time bin and four clusters in the late time bin. In line with the analysis in sensor space, coefficients were higher in the sentence condition than in the word list condition in the early time-bin (200–400 ms PSO). These differences appeared bilaterally in the posterior superior and middle frontal gyrus (dorsolateral and dorsomedial prefrontal cortex) and cingulate gyrus (Fig. 6*A*). In the right hemisphere, the cluster extended to the inferior frontal gyrus (Fig. 6*A*).

In the late time bin (500–700 ms PSO; Fig. 6*B*), coefficients were higher in the word list condition than in the sentence condition in three of four clusters. Those clusters appeared in the left hemisphere in the posterior temporal lobe across the superior, middle, and inferior gyri/sulci, the temporal pole, and the parahippocampal gyrus. In the right hemisphere, the effects appeared in superior temporal, inferior parietal, and caudal frontal areas, as well as cingulate gyrus. In a final cluster in the late time bin, the coefficients were higher in the sentence than in the word list condition. This cluster spanned left inferior frontal areas, orbital cortex, as well as a small portion of the middle frontal gyrus.

In addition, we observed a difference between the responses in left orbitofrontal and ventrolateral prefrontal cortex, including the inferior frontal gyrus. In this area, the response peaked in the late time bin in the sentence condition only. That this area is where we found a difference in late time lags is not surprising given the large literature implicating the left inferior frontal cortex, or Broca's area, in syntactic processes (Friederici, 2011, 2012, 2015; Hagoort, 2013, 2016; Matchin and Hickok, 2020).

Given our finding that the word list response appeared delayed in comparison to the response in the sentence condition, we also considered responses in the sentence and word list conditions separately through one-sample cluster-based permutation tests. Here, we observed a widespread response in both conditions; and indeed, this response appears in the early time window in the sentence condition (Fig. 6*C*) and in the late time window in the word list condition (Fig. 6*F*).

As we already observed in the contrast, in the late time window, the response to word-internal information encompasses the left posterior superior, middle, and inferior temporal gyri (including parahippocampal gyrus) and the temporal poles, as well as bilateral somatosensory areas in both conditions. These areas are traditionally associated with lexical and semantic memory (Binder and Desai, 2011; Hagoort, 2013, 2016). Furthermore, as we observed in the early time window, this response includes the bilateral dorsolateral prefrontal cortex. These areas are part of the dorsal attention network and have been implied to control activation and selection of information stored in temporoparietal cortices (Binder and Desai, 2011). In addition, like we observed in the contrast between conditions, in the sentence condition a late response appears in the left inferior frontal gyrus (Fig. 6*E*). This response was absent in the word list condition. We compared the reconstruction accuracies using a cluster-based two-way ANOVA with factors Condition (levels: word list, sentence) and Model (levels: Entropy/Surprisal, Full). There were no significant differences (all p values > 0.1).

Together, these findings indicate that (1) much, but not all, of the response to word internal information is shared between conditions in space; (2) the response develops differently in time, with a delay in the word list condition; and (3) word internal information modulates activity in the left inferior frontal gyrus only in the presence of a coherent context.

Theta band

Sensor-level analysis

In the theta band, the cluster-based permutation test revealed no differences between the word list and sentence TRFs for the word frequency feature (Fig. 7). The one-sample tests indicated, however, a response between 100 and 200 ms in the word list condition that was absent in the sentence condition.

Like in the delta band, the full model was *accuracies* ~ *condition* * (*frequency* + *entropy* + *surprisal*) + (*1*/*participant*). Removing the interaction between *frequency* and *condition*, or the interaction between *surprisal* and *condition*, decreased model fit [marginally; frequency, $\chi^2(1)$ =

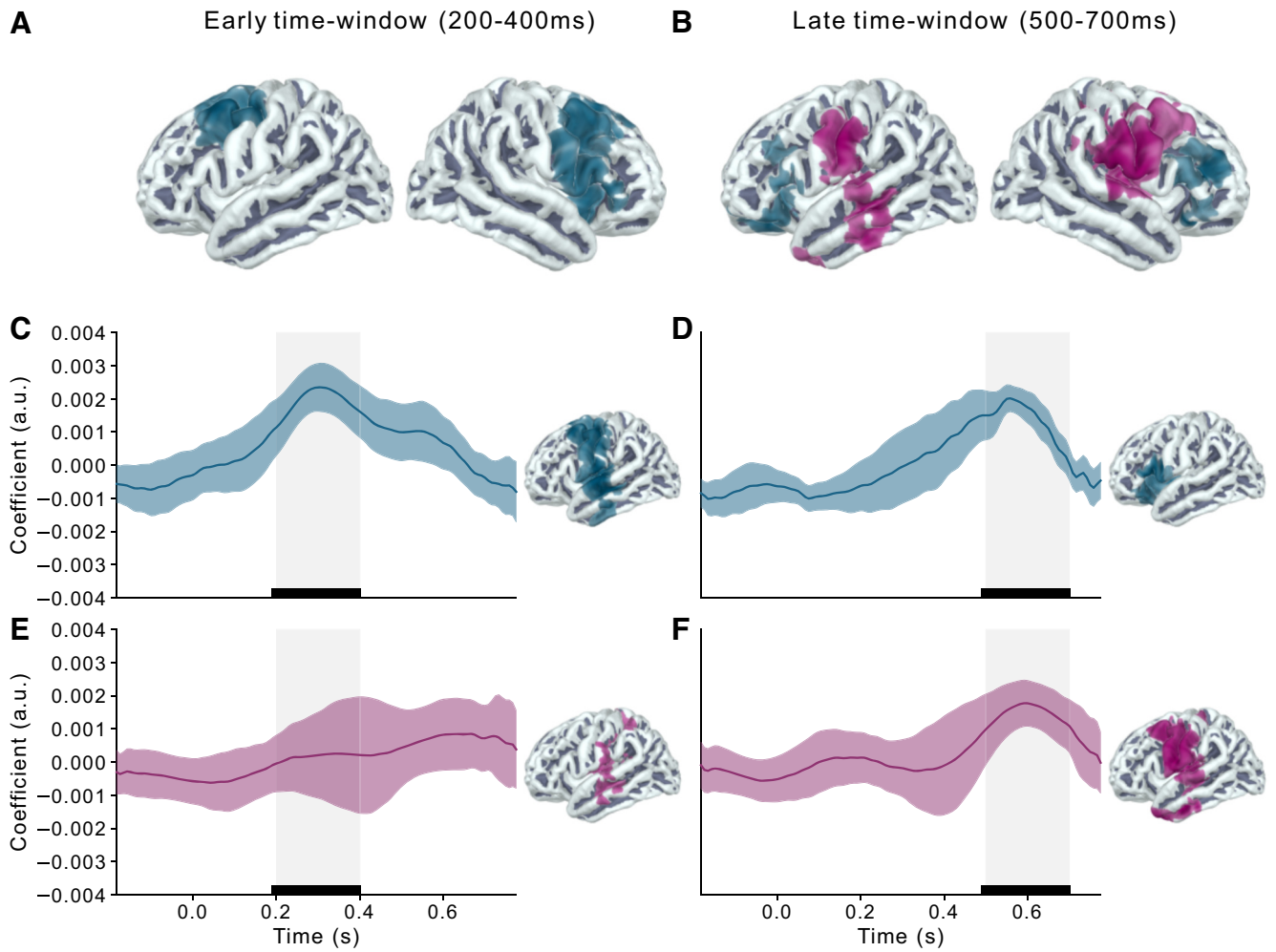

Figure 6. Clusters from the cluster-based permutation test. Left column, Early time window (200–400 ms). Right column, Late time window (500–700 ms). ***A***, Differences between the word list and sentence responses to word frequency in the early time window. Blue indicates that the coefficients in the sentence condition are higher than in the word list condition; pink indicates the coefficients in the word list condition are higher than in the sentence condition. ***B***, Differences between the word list and sentence responses to word frequency in the late time window. Blue indicates that the coefficients in the sentence condition are higher than in the word list condition; pink indicates the coefficients in the word list condition are higher than in the sentence condition. ***C***, Sentence condition. TRF and spatial distribution of one-sample cluster in early time-window. Time-window is indicated in gray. ***D***, Sentence condition. TRF and spatial distribution of one-sample cluster in late time window. Time window is indicated in gray. ***E***, Word list condition. TRF and spatial distribution of one-sample cluster in early time window. Time window is indicated in gray. ***F***, Word list condition. TRF and spatial distribution of one-sample cluster in late time window. Time window is indicated in gray. Shaded areas in blue and pink indicate SD.

3.80, $p = 0.051$; surprisal, $\chi^2(1) = 3.95$, $p < 0.05$], but removing the interaction between *entropy* and *condition* did not [$\chi^2(1) = 0.47$, $p = 0.49$]. We continued with the model *accuracies* ~ *condition* * (*frequency* + *surprisal*) + *entropy* + (*1/participant*). The AIC comparison confirmed that this model was the best descriptor of the data. The results of this model are summarized in Table 3.

In theta, too, there was a main effect of condition ($\beta = 2.09 * 10^{-3}$, SE = $6.90 * 10^{-4}$, $t_{(1530)} = 3.02$, $p < 0.01$), with reconstruction accuracies being higher in the word list condition than in the sentence condition; see Figure 8. In addition, there was a main effect of *frequency* ($\beta = 1.17 * 10^{-3}$, SE = $5.64 * 10^{-4}$, $t_{(1530)} = 2.07$, $p < 0.05$) indicating that generally the addition of word frequency improved reconstruction accuracy. The interaction between *frequency* and *condition* approached but did not reach significance ($\beta = 1.56 * 10^{-3}$, SE = $7.97 * 10^{-4}$, $t_{(1530)} = 1.95$, $p = 0.051$), indicating a potential trend for the frequency effect to be larger in the sentence condition than in the word list condition (Fig. 7).

With respect to the other predictors, there was a positive effect of *entropy* ($\beta = 2.43 * 10^{-3}$, SE = $3.99 * 10^{-4}$, $t_{(1530)} = 1.95$, $p < 0.01$) and an interaction between *condition* and *surprisal* ($\beta = 1.55 * 10^{-3}$, SE = $7.92 * 10^{-4}$, $t_{(1530)} = 1.99$, $p < 0.05$), indicating that *surprisal* enhanced reconstruction accuracies more in the sentence condition than in the word list condition.

Again, we performed *post hoc t* tests comparing the two largest models (Entropy/Surprisal and Full) to gain more insight in the effect of word frequency on the reconstruction accuracies. These showed that the word frequency predictor enhanced reconstruction accuracies in the sentence condition ($t_{(101)} = 5.67$; $p < 0.01$), but not in the word list condition ($t_{(101)} = 1.48$; $p = 0.57$). There were no effects of condition for these two models (all p values = 1).

Source reconstruction

Given that the permutation test in the sensor-based analysis did not reveal any effects in the theta band, and we could not select time bins a priori, we performed a cluster-based permutation test on the full TRF. This revealed two clusters in the right hemisphere between 100 and 250 ms. Both of these clusters reflect a larger amplitude across right frontal and temporal areas for the TRF in the word list condition than the sentence condition, as can be seen in the plots of

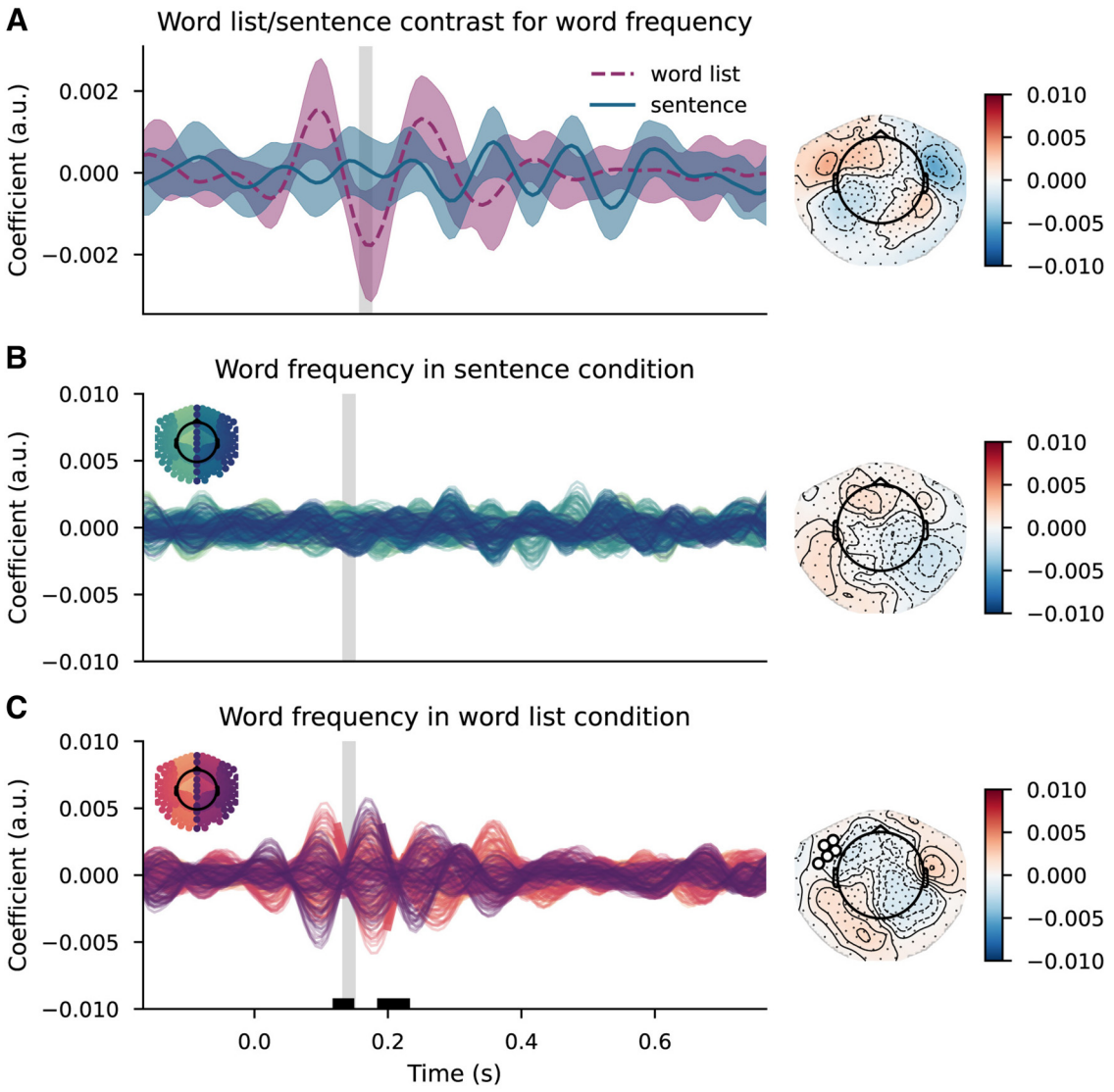

Figure 7. ***A***, The word frequency TRF in both conditions in the theta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points of those significant clusters. Shaded area indicates SD. ***B***, Word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. ***C***, Word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Vertical gray lines indicate the time points of the scalp maps.

the time courses of the clusters in Figure 9. These effects, although visible in Figure 7, *A–C*, did not reach significance in the sensor analysis, potentially because of the stringent threshold (recommended value multiplied by three) chosen there.

Control analysis I: data from ten Oever et al. (2022)

In the delta band, the cluster-based permutation test revealed no significant differences between the word frequency response in the word lists and sentences. To evaluate whether this was because there were no detectable responses or no difference between conditions, we performed one-sample cluster-based permutation tests. Here we observed a response in the sentence condition over a large array of left-posterior sensors that was significant from word onset to ~400 ms. The peak appears ~200 ms (Fig. 10*A*). Although Figure 10*B* suggests a potential response of ~400 ms in the word list condition, there were no significant clusters. As in the main analysis, there were no significant differences between conditions in the responses to word onset.

The absence of a difference between the conditions and the lack of a detectable response in the word list condition alone make the results from this analysis difficult to interpret in relation to the main analysis. The large difference between the sample sizes ($N = 102$ vs $N = 16$, respectively) may play a role in this difference. We performed a power analysis on the difference between the conditions in the control analysis using the average t values from the time points and sensors taken from the significant clusters from the same contrast in the main analysis. This showed that power would increase on average by 30.7% when taking a sample of 102 participants, with three clusters reaching a power of above 96%. This suggests that the control analysis did not have enough power to reject or confirm the hypothesis that the delay in the response in the word list condition is caused by the different temporal dynamics in the original analysis. We therefore refrain from drawing conclusions on the basis of this finding.

Nevertheless, the ANOVA on the reconstruction accuracies revealed a main effect of model ($F_{(1,15)} = 38.01$; $p < 0.01$), indicating that the word frequency predictor enhanced reconstruction accuracy, and an interaction between condition and model ($F_{(1,15)} = 6.79$; $p < 0.05$), suggesting that this effect was larger for

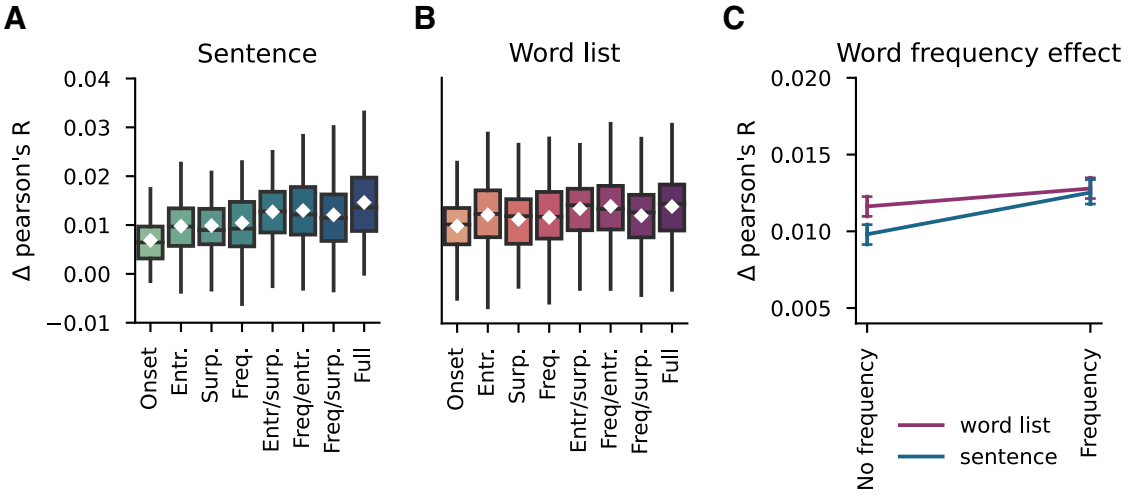

Figure 8. Reconstruction accuracies in the theta band. ***A***, Reconstruction accuracy difference with the envelope model for each model in the sentence condition. Middle line indicates the median, the white diamond indicates the mean. ***B***, Reconstruction accuracy difference with the envelope model for each model in the word list condition. Middle line indicates the median, the white diamond indicates the mean. ***C***, The interaction between condition and frequency on the reconstruction accuracies ($p = 0.051$, see above, Theta band). Values on the *y*-axis are the difference with the envelope (as in ***A***, ***B***). Error bars represent the 95% confidence interval. Entr.: entropy, surp: surprisal, freq: frequency.

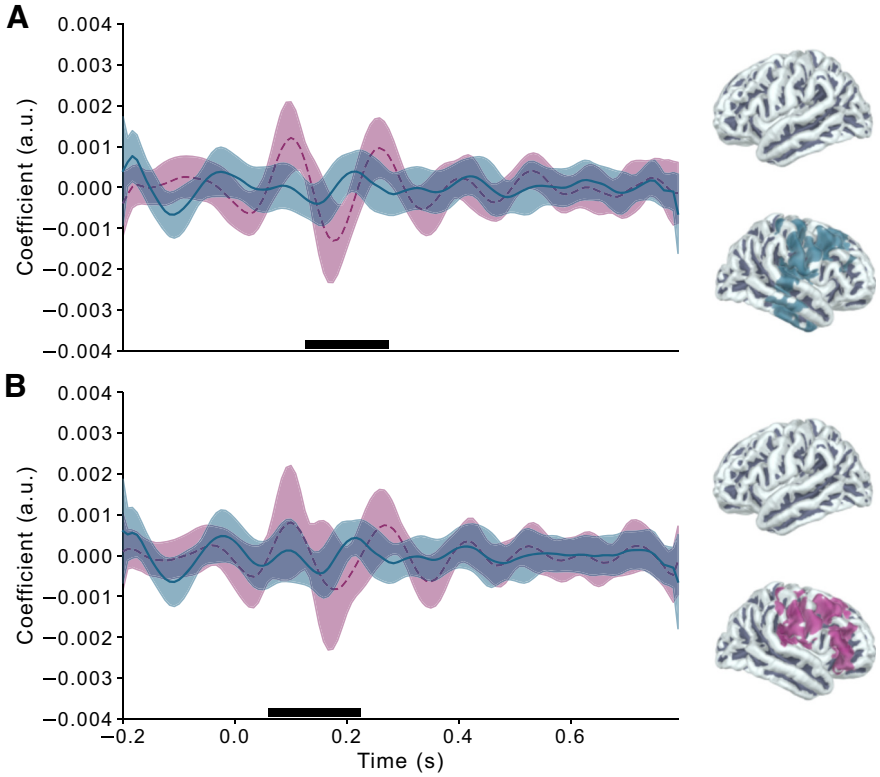

Figure 9. Clusters from the theta-band TRFs in source space. Blue indicates that coefficients sentence is greater than word list; pink indicates word list is greater than sentence. ***A***, Right-lateralized cluster where TRF sentence is greater than word list. Shaded area indicates SD. ***B***, Right-lateralized cluster where TRF word list is greater than sentence. Shaded area indicates SD.

the sentence condition than for the word list condition (Fig. 10*C*). There was no main effect of condition ($p = 0.16$). In the theta band, there were no significant effects on the TRF waveforms nor on the accuracy values (Fig. 11).

Control analysis II: simulations

To evaluate the effect of differences in interstimulus intervals (i.e., pauses), we simulated raw MEG data consisting of a signal (different impulse responses) and optional noise. Strikingly, the interstimulus interval has no direct influence on the reconstruction score, although the length of the segment on which we estimate the score does (Fig. 12). In this case, the difference in interstimulus interval, which eventually leads to a difference in data length, shows how the bias in the score observed between conditions is solely because of the difference in duration. The bias, however, is constant, and should be controlled for when directly comparing models within conditions. Moreover, we actually observe the opposite effect in our MEG analysis; the absolute scores for the longer segment of data (the word lists) are higher than the shorter segment of data (the sentences). This means that our score differences exist above and beyond any bias generated from the stimulus difference.

Discussion

In this study, we asked whether low-frequency neural readouts associated with words systematically changed as a function of being in a sentence context and whether neural readouts were modulated by purely lexical properties over and above sensory and contextual distributional variables. We contrasted responses to word frequency for words in sentences with word lists, the latter lacking any syntactic structure and combinatorial lexical meaning. We hypothesized that the delta-band but not theta-band responses to word frequency would be different in word lists and sentences as a consequence of the (in)availability of sentence context. Specifically, following findings from speech tracking, we expected a stronger presence of the word frequency response in the sentence condition.

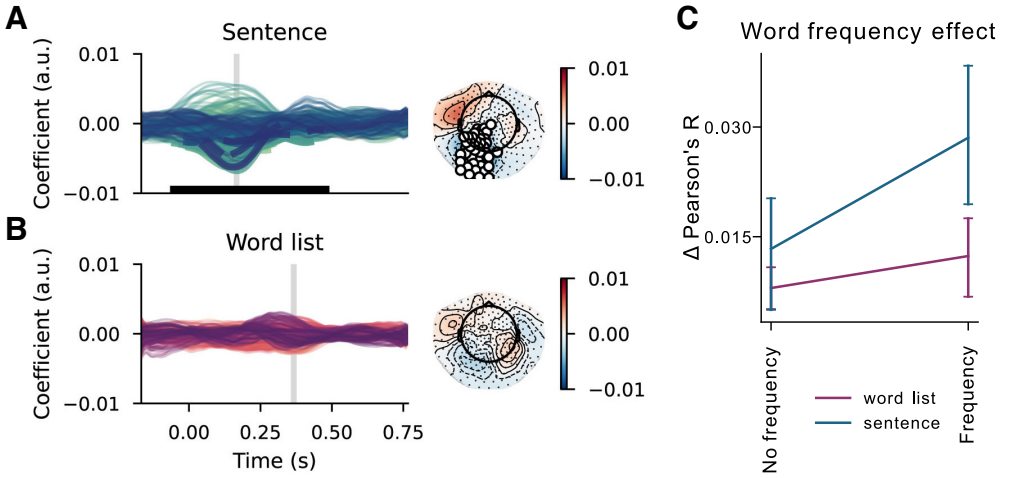

Figure 10. Delta-band effects in the extra data. ***A***, Word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters. ***B***, Word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters (none). Vertical gray lines indicate the time points of the scalp maps. ***C***, The interaction between condition and frequency on the reconstruction accuracies. Values on the *y*-axis are the difference with the envelope (as in ***A***, ***B***). Error bars represent the 95% confidence interval.

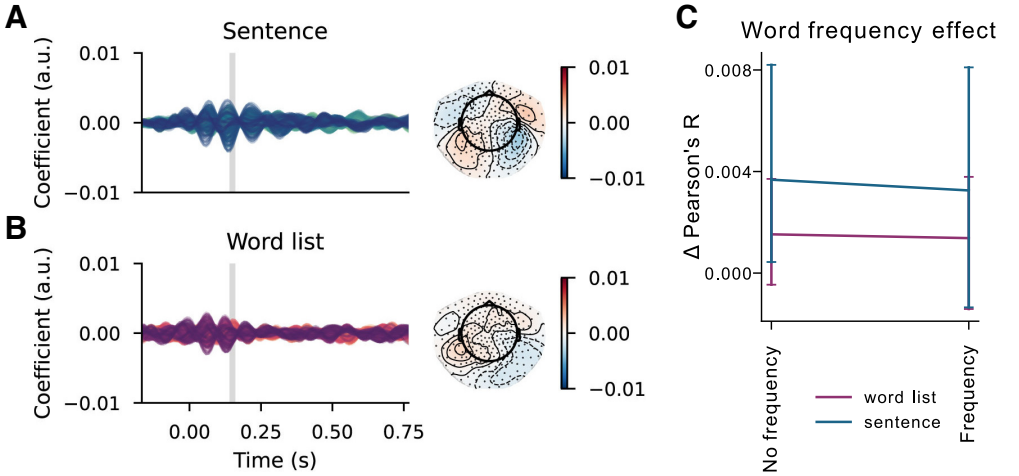

Figure 11. Theta-band effects in the extra data. ***A***, Word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters. ***B***, Word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters (none). Entr.: entropy, surp: surprisal, freq: frequency. ***C***, The (lack of an) interaction between condition and frequency on the reconstruction accuracies. Values on the *y*-axis are the difference with the envelope (as in ***A***, ***B***). Error bars represent the 95% confidence interval.

Our findings showed that the delta-band response to word frequency differs between word lists and sentences in time and, albeit minimally, in space. In both conditions, word internal information modulates a response across the left temporal lobe and the frontal cortex. However, this response occurred ~300 ms earlier in the presence of a coherent sentence context. In addition, in a sentence context, word internal information could be seen to modulate activity in the left inferior frontal gyrus at ~600 ms after word onset, a response that is absent when a word is not embedded in a sentence. Furthermore, the word frequency feature explains more variance over and above the other features in the sentence condition than in the word list condition. In the theta band, there were only minimal differences between the conditions. We discuss our results in more detail below.

In psycholinguistic theories of word recognition, word frequency is often modeled as the baseline of activation or the prior probability of a word, for example, the Logogen model (Morton, 1969), Cohort model (Marslen-Wilson, 1987), and Shortlist A and B (Norris, 1994; Norris and McQueen, 2008). We assume therefore that the neural readout associated with word frequency represents neural activity during the process of word recognition. Our results provide direct evidence that this process happens differently depending on whether the structure building of sentence comprehension is also occurring. We know that words are recognized faster when they are embedded in a coherent sentence context (Marslen-Wilson and Welsh, 1978; Tyler and Wessels, 1983); this is reflected in the delayed word list response to word frequency (Lam et al., 2016).

Furthermore, the reconstruction accuracies in sensor space suggest that the response to word frequency explains more variance in the sentence condition than in the word list condition. This may seem contradictory to findings from psycholinguistics. Indeed, the behavioral effect of word frequency, when assessed with reaction time measures, diminishes in the sentence context (Schuberth and Eimas, 1977; Tyler and Wessels, 1983; Simpson

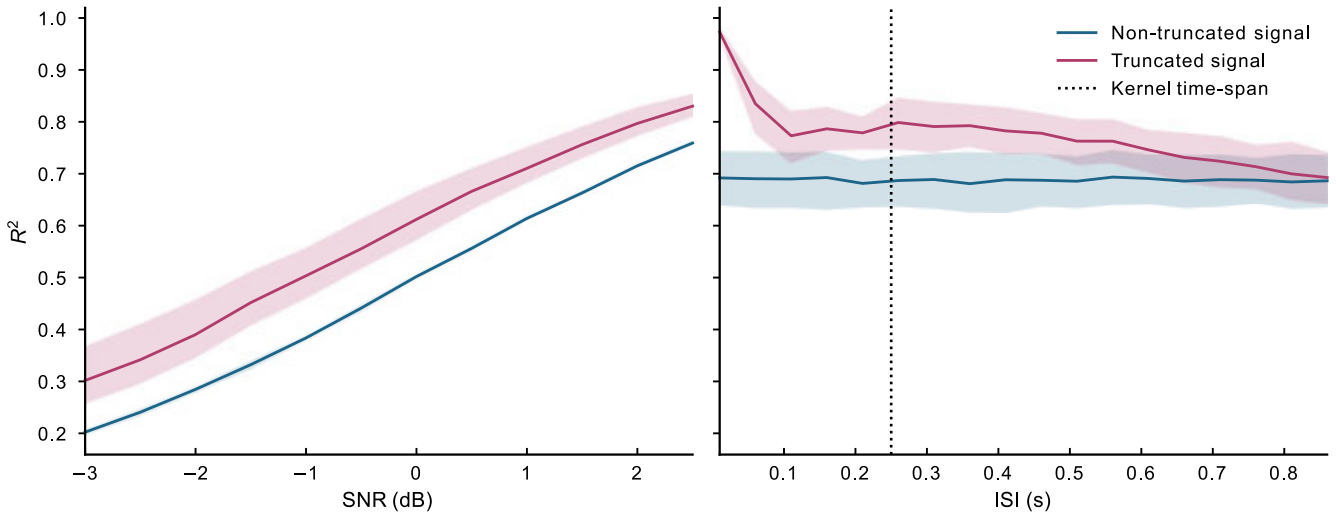

Figure 12. Influence of interstimulus interval (ISI), data length, and noise on the score (reconstruction accuracy; R_2). Left, The (proportional) influence of broadband signal-to-noise ratio (SNR) on the score. Right, For every interstimulus interval value, the same score is measured if the data length is kept constant, and the score deflated for longer signals as more noise is being evaluated in the scoring.

et al., 1989). Put differently, words with a low frequency are recognized more slowly than words with a high frequency. This does not necessarily mean that lexical information explains less variance in the neural signal. In fact, studies that consider metrics like mutual information between the brain and the speech signal find that the brain represents aspects of the speech signal more reliably when more linguistic information is present (Kaufeld et al., 2020; ten Oever et al., 2022), whereas the acoustic information in speech matters less for word recognition when the word is embedded in a sentence (Boothroyd and Nittrouer, 1988; Mattys et al., 2012). In general terms, these findings suggest that the brain represents lower-level features more reliably when higher-level information can be inferred, whereas the lower-level information itself becomes less important for the outcome of the task. Indeed, that words are represented more robustly when sentence context is provided is reflected in the accuracy scores on the word monitoring task performed in this study; participants were more likely to correctly remember whether a word was mentioned when they had been presented with a sentence than when they heard a word list.

There are two causes for this finding. First, the perceptual salience of the words in the word list condition leads to a large response to the speech envelope; the response to lexical features then are of relatively lower power and explain less of the variance in the signal relative to the lower-level features. Second, as a consequence of words being embedded in larger structures, phrases and sentences, word frequency is likely present in a larger neural network in the sentence condition than in the word list condition (Martin, 2020). The signal is therefore reconstructed better in a wider array of sensors, leading to an overall larger increase in reconstruction accuracies. As discussed below, the presence of the effect in the control analysis favors the latter interpretation. The propagation of lexical information to a wider network is additionally reflected in the differences between conditions in the inferior frontal gyrus at ~600 ms. This interpretation is consistent with findings that show that sentence structure influences the dynamics and distribution of neural signals (Blank et al., 2016; Schell et al., 2017; Matchin et al., 2019a,b; Grodzinsky et al., 2021; Bai et al., 2022; Coopmans et al., 2022; ten Oever et al., 2022).

Importantly, both the TRF and the reconstruction accuracy effects of sentence context on the representation of word-internal information are independent of (1) the contextual probability predictors surprisal and entropy and (2) sensory information in the speech envelope. Each of these predictors is undoubtedly important for how the neural signal represents lexical information (e.g., sensory, Doelling et al., 2014; and probability, Weissbart et al., 2020). Given that these influences were accounted for by the encoding model, the differences that remain imply a role for abstract structure and meaning on the transformation of low-frequency neural readouts associated with words (or more minimally, associated with purely lexical features). These conclusions are in line with findings on the visual part of the dataset, not analyzed here (Huizeling et al., 2022).

Also striking is the difference between the effects in the delta and theta bands. In the theta band, the responses to word frequency differed between conditions only slightly; the amplitude of the response was larger in the word list condition than in the sentences in the right frontal and temporal hemisphere ~100 ms, possibly indicating that word frequency in interaction with contextual information tunes sensory sampling. The addition of the word frequency predictor had a small effect on the reconstruction accuracies, which was present only in the *post hoc* analysis. In general, theta-band activity appears to be more sensitive to perceptual aspects of the stimulus than to linguistic aspects. For example, tracking of sound by theta-band activity persists even in the absence of linguistic information (Molinaro and Lizarazu, 2018), whereas it is affected when acoustic edges in the stimulus are experimentally manipulated (Doelling et al., 2014). However, in line with the differences that we do see, Donhauser and Baillet (2020) showed that the gain of early theta responses varies according to the contextual uncertainty of speech. The results from the present analysis are consistent with an account in which the theta band is important for speech processing but not as central for the representation of higher-level features such as lexical-internal information. At the same time, the process reflected by theta modulations during language comprehension is likely to be influenced by linguistic context.

In addition to the linguistic differences, there was a variable pause between the words in the word list condition only. To examine the potential effect of this additional difference between the conditions on our results, we ran several simulations. The simulations showed that the ISI between events modeling word-

like responses has no effect on model evaluation and TRF estimation. However, there will be a constant bias in the model score that is proportional to the broadband signal-to-noise ratio (where the noise is the additive contribution beyond variance explained by the linear model). This bias is not directly because of the differences in ISI but rather the fact that we are integrating a larger portion of data in the list condition, thus more noise to contribute to the score. As such, any model comparison contrasting scores within condition will eliminate the constant bias. Furthermore, this bias leans toward deflating the score of the model evaluated on the longest segment of data (the word list condition). We found that with the envelope alone, the scores in the list condition were higher than the scores of the sentence condition; this is in direct contrast with the expectations from the simulations. From these simulations we conclude therefore that the delay in the TRF waveform and the interaction effect in the reconstruction of the neural signal are not just because of difference in signal length between the word list and sentence condition.

The next question is, then, What are the potential cognitive effects of silence between the words? There are three potential effects, (1) higher perceptual saliency of each word, already mentioned above; (2) decreased word rate; and (3) absence of phonological cues between words, such as prosody and coarticulation. (A reviewer suggested we add a prosody predictor. We constructed a prosody predictor by extracting the prosody contour using Parselmouth, a Praat wrapper for Python. Running the analysis with this extra predictor did not qualitatively change the results.) We consider phonological cues to be consequences of as well as cues to the sentence context; they would be different between word lists and sentences in naturalistic conditions as well. The first two, however, need some consideration.

As mentioned above, the perceptual difference between two consecutive words is much smaller than the difference between silence and a word. This effect was visible in the speech–brain coherence for both conditions (Fig. 1; coherence was much higher in the word list condition in the delta band) and caused overall higher reconstruction accuracy in the word list condition. Importantly, in the analysis on a second dataset in which this difference between conditions did not exist, the interaction effect between word lists and sentences was replicated. The word frequency feature explained more variance over and above the envelope and word onset predictors in the sentence condition than in the word list condition. Furthermore, we stipulated that a general delaying effect on word processing generated by the decreased word rate in the word list condition would be visible with other features as well. Nevertheless, the word onset feature, the only feature in addition to word frequency that was numerically identical between conditions, did not show such a difference. These findings indicated that it was only the response to word-internal information that was delayed and suggests that the brain processes lexical information later in the absence of a coherent sentence context. Taken together, this indicates that the effects described in this work are unlikely to be driven by silence.

In summary, this study suggests that delta-band, and to a lesser extent, theta-band responses to word-internal information are affected by sentence context in time and in space. Given that a difference in encoding of a strictly lexical feature persists when context-driven lexical features like entropy and surprisal are added, we conclude that low-frequency responses to word internal information are changed by sentential structure and meaning and not by probabilistic differences alone. In the delta band, a lexical response across the posterior and anterior left temporal lobe and the bilateral parietal lobe is delayed in the absence of sentence context. In addition, a word embedded in a sentence context determines whether inferior frontal areas are responsive to lexical information. In the theta band, a larger amplitude in the word lists at ~100 ms across the right frontal and parietal areas suggests that linguistic information can tune sensory sampling. In addition, this study shows that the TRF can be used to model acoustic differences between stimuli when measuring higher-level linguistic effects (Bai et al., 2022). The results of this study show how the neural representation of words is affected by the linguistic structure of sentence context and as such provide beginning insight into how the brain instantiates compositionality in language processing.

References

Armeni K, Willems RM, van den Bosch A, Schoffelen JM (2019) Frequency-specific brain dynamics related to prediction during language comprehension. Neuroimage 198:283–295.

Bai F, Meyer AS, Martin AE (2022) Neural dynamics differentially encode phrases and sentences during spoken language comprehension. PLOS Biol 20:e3001713.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1–48.

Binder JR, Desai RH (2011) The neurobiology of semantic memory. Trends Cogn Sci 15:527–536.

Blanco-Elorrieta E, Ding N, Pylkkänen L, Poeppel D (2020) Understanding requires tracking: noise and knowledge interact in bilingual comprehension. J Cogn Neurosci 32:1975–1983.

Blank I, Balewski Z, Mahowald K, Fedorenko E (2016) Syntactic processing is distributed across the language system. Neuroimage 127:307–323.

Boersma P, Weenink D (2018) Praat: doing phonetics by computer (6.0.40). Available at: http://www.praat.org/.

Boothroyd A, Nittrouer S (1988) Mathematical treatment of context effects in phoneme and word recognition. J Acoust Soc Am 84:101–114.

Brennan JR, Hale JT (2019) Hierarchical structure guides rapid linguistic predictions during naturalistic listening. PLoS ONE 14:e0207741.

Brennan JR, Martin AE (2020) Phase synchronization varies systematically with linguistic structure composition. Phil Trans R Soc B 375:20190305–20190383.

Brodbeck C, Presacco A, Simon JZ (2018a) Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. Neuroimage 172:162–174.

Brodbeck C, Hong LE, Simon JZ (2018b) Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol 28:3976–3983.e5.

Broderick MP, Anderson AJ, di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 28:803–809.e3.

Coopmans CW, de Hoop H, Hagoort P, Martin AE (2022) Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. Neurobiol Lang 3:386–412.

di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164.

Ding N, Pan X, Luo C, Su N, Zhang W, Zhang J (2018) Attention is required for knowledge-based sequential grouping: insights from the integration of syllables into words. J Neurosci 38:1178–1188.

Doelling KB, Arnal LH, Ghitza O, Poeppel D (2014) Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage 85:761–768.

Donhauser PW, Baillet S (2020) Two distinct neural timescales for predictive speech processing. Neuron 105:385–393.e9.

Etard O, Reichenbach T (2019) Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. J Neurosci 39:5750–5759.

Friederici AD (2011) The brain basis of language processing: from structure to function. Physiol Rev 91:1357–1392.

Friederici AD (2012) The cortical language circuit: from auditory perception to sentence comprehension. Trends Cogn Sci 16:262–268.
Friederici AD (2015) White-matter pathways for speech and language processing. In Handbook of clinical neurology (Aminoff MJ, Boller F, Swaab DF, eds), pp 177–186. Elsevier:Amsterdam. [10.1016/B978-0-444-62630-1.00010-X]
Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen MS (2013) MEG and EEG data analysis with MNE-Python. Front Neurosci 7:267.
Grodzinsky Y, Pieperhoff P, Thompson C (2021) Stable brain loci for the processing of complex syntax: a review of the current neuroimaging evidence. Cortex 142:252–271.
Hagoort P (2013) MUC (memory, unification, control) and beyond. Front Psychol 4:416.
Hagoort P (2016) MUC (memory, unification, control): a model on the neurobiology of language beyond single word processing. In: Neurobiology of language. pp 339–347. Elsevier:Cambridge, Massachusetts. https://doi.org/10.1016/B978-0-12-407794-2.00028-6 .
Heilbron M, Armeni K, Schoffelen J-M, Hagoort P, de Lange FP (2021) A hierarchy of linguistic predictions during natural language comprehension. Proc Natl Acad Sci U S A 119:e2201968119.
Huizeling E, Arana S, Hagoort P, Schoffelen JM (2022) Lexical frequency and sentence context influence the brain's response to single words. Neurobiol Lang 3:149–179.
Katz L, Boyce S, Goldstein L, Lukatela G (1987) Grammatical information effects in auditory word recognition. Cognition 25:235–263.]
Kaufeld G, Bosker HR, ten Oever S, Alday PM, Meyer AS, Martin AE (2020) Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. J Neurosci 40:9467–9475.
Keitel A, Gross J, Kayser C (2018) Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLOS Biol 16:e2004473.
Keuleers E, Brysbaert M, New B (2010) SUBTLEX-NL: a new measure for Dutch word frequency based on film subtitles. Behav Res Methods 42:643–650.
Lam NHL, Schoffelen JM, Uddén J, Hultén A, Hagoort P (2016) Neural activity during sentence processing as reflected in theta, alpha, beta, and gamma oscillations. Neuroimage 142:43–54.
Lam NHL, Hultén A, Hagoort P, Schoffelen JM (2018) Robust neuronal oscillatory entrainment to speech displays individual variation in lateralisation. Lang Cogn Neurosci 33:943–954.
León-Cabrera P, Rodríguez-Fornells A, Morís J (2017) Electrophysiological correlates of semantic anticipation during speech comprehension. Neuropsychologia 99:326–334.
Liu Y, Shu H, Wei J (2006) Spoken word recognition in context: evidence from Chinese ERP analyses. Brain Lang 96:37–48.
Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177–190.
Marslen-Wilson WD (1987) Functional parallelism in spoken word-recognition. Cognition 25:71–102.
Marslen-Wilson WD, Welsh A (1978) Processing interactions and lexical access during word recognition in continuous speech. Cogn Psychol 10:29–63.
Martin AE (2016) Language processing as cue integration: grounding the psychology of language in perception and neurophysiology. Front Psychol 7:120–17.
Martin AE (2020) A compositional neural architecture for language. J Cogn Neurosci 32:1407–1427.
Matchin W, Hickok G (2020) The cortical organization of syntax. Cereb Cortex 30:1481–1498.
Matchin W, Brodbeck C, Hammerly C, Lau E (2019a) The temporal dynamics of structure and content in sentence comprehension: evidence from fMRI-constrained MEG. Hum Brain Mapp 40:663–678.
Matchin W, Liao CH, Gaston P, Lau E (2019b) Same words, different structures: an fMRI investigation of argument relations and the angular gyrus. Neuropsychologia 125:116–128.
Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: a review. Language and Cognitive Processes 27:953–978.
Mazerolle MJ (2020) AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). Available at: https://cran.r-project.org/package=AICcmodavg .
Meyer L, Henry MJ, Gaston P, Schmuck N, Friederici AD (2017) Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cereb Cortex 27:4293–4302.
Meyer L (2018) The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. European Journal of Neuroscience 48:2609–2621.
Meyer L, Sun Y, Martin AE (2020a) "Entraining" to speech, generating language? Language, Cognition and Neuroscience 35:1138–1148.
Meyer L, Sun Y, Martin AE (2020b) Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. Language, Cognition and Neuroscience 35:1089–1099.
Molinaro N, Lizarazu M (2018) Delta(but not theta)-band cortical entrainment involves speech-specific processing. Eur J Neurosci 48:2642–2650.
Morton J (1969) Interaction of information in word recognition. Psychol Rev 76:165–178.
Nelson MJ, el Karoui I, Giber K, Yang X, Cohen L, Koopman H, Cash SS, Naccache L, Hale JT, Pallier C, Dehaene S (2017) Neurophysiological dynamics of phrase-structure building during sentence processing. Proc Natl Acad Sci U S A 114:E3669–E3678.
Norris D (1994) Shortlist: a connectionist model of continuous speech recognition. Cognition 52:189–234.
Norris D, McQueen JM (2008) Shortlist B: a Bayesian model of continuous speech recognition. Psychol Rev 115:357–395.
Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: open source software for advanced analysis of MEG, EEG and invasive electrophysiological data. Comput Intell Neurosci 2011:1–9.
Sassenhagen J, Draschkow D (2019) Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. Psychophysiology 56:e13335.
Schell M, Zaccarella E, Friederici AD (2017) Differential cortical contribution of syntax and semantics: an fMRI study on two-word phrasal processing. Cortex 96:105–120.
Schoffelen JM, Hultén A, Lam N, Marquand AF, Uddén J, Hagoort P (2017) Frequency-specific directed interactions in the human brain network for language. Proc Natl Acad Sci U S A 114:8083–8088.
Schoffelen JM, Oostenveld R, Lam NHL, Uddén J, Hultén A, Hagoort P (2019) A 204-subject multimodal neuroimaging dataset to study language processing. Sci Data 6:17.
Schuberth RE, Eimas PD (1977) Effects of context on the classification of words and nonwords. J Exp Psychol Hum Percept Perform 3:27–36.
Sheather S (2009) Diagnostics and Transformations for Multiple Linear Regression. In: A Modern Approach to Regression with R. Springer Texts in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-09608-7_6.
Simpson GB, Peterson RR, Casteel MA, Burgess C (1989) Lexical and sentence context effects in word recognition. J Exp Psychol Learn Mem Cogn 15:88–97.
Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. J Neurosci 32:8443–8453.
Tavano A, Blohm S, Knoop CA, Muralikrishnan R, Fink L, Scharinger M, Wagner V, Thiele D, Ghitza O, Ding N, Menninghaus W, Poeppel D (2022) Neural harmonics of syntactic structure. bioRxiv 031575. https://doi.org/10.1101/2020.04.08.031575.
ten Oever S, Carta S, Kaufeld G, Martin AE (2022) Neural tracking of phrases in spoken language comprehension is automatic and task-dependent. Elife 11:e77468.
Tomaschek F, Hendrix P, Baayen RH (2018) Strategies for addressing collinearity in multivariate linguistic data. J Phon 71:249–267.
Tyler LK, Wessels J (1983) Quantifying contextual contributions to word-recognition processes. Percept Psychophys 34:409–420.
Vallat R (2018) Pingouin: statistics in Python. JOSS 3:1026.
van den Bosch A, Berck P (2009) Memory-based machine translation and language modeling. Prague Bull Math Linguist 91:17–26.
Weissbart H, Kandylaki KD, Reichenbach T (2020) Cortical tracking of surprisal during continuous speech comprehension. J Cogn Neurosci 32:155–166.