Cortical Tracking of Surprisal during Continuous Speech Comprehension

Hugo Weissbart[1], Katerina D. Kandylaki[1,2], and Tobias Reichenbach[1]

Abstract

■ Speech comprehension requires rapid online processing of a continuous acoustic signal to extract structure and meaning. Previous studies on sentence comprehension have found neural correlates of the predictability of a word given its context, as well as of the precision of such a prediction. However, they have focused on single sentences and on particular words in those sentences. Moreover, they compared neural responses to words with low and high predictability, as well as with low and high precision. However, in speech comprehension, a listener hears many successive words whose predictability and precision vary over a large range. Here, we show that cortical activity in different frequency bands tracks word surprisal in continuous natural speech and that this tracking is modulated by precision. We obtain these results through quantifying surprisal and precision from naturalistic speech using a deep neural network and through relating these speech features to EEG responses of human volunteers acquired during auditory story comprehension. We find significant cortical tracking of surprisal at low frequencies, including the delta band as well as in the higher frequency beta and gamma bands, and observe that the tracking is modulated by the precision. Our results pave the way to further investigate the neurobiology of natural speech comprehension. ■

INTRODUCTION

To understand spoken language, a listener must rapidly process information that unfolds over several timescales, including the duration of syllables at around 150 msec, words of about 300 msec, and phrases of 1 sec (Giraud & Poeppel, 2012). Recent studies have shown that cortical activity in the delta, theta, and gamma frequency bands tracks acoustic features of speech such as the speech envelope as well as phonemic features (Ding et al., 2018; Di Liberto, O'Sullivan, & Lalor, 2015; Ding & Simon, 2014; Zion Golumbic et al., 2013; Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007). This cortical tracking of speech features has accordingly been proposed to reflect neural mechanisms of speech processing, for instance, an online segmentation of speech into acoustic speech tokens such as phonemes that occur on the timescale of a few hundreds of milliseconds (Hyafil, Fontolan, Kabdebon, Gutkin, & Giraud, 2015; Giraud & Poeppel, 2012).

The processing of higher level linguistic information in speech may employ cortical tracking as well. Recent findings showed that cortical activity in the delta and theta frequency bands synchronized to sequential cues such as the rhythm of phrases and sentences in continuous speech (Keitel, Gross, & Kayser, 2018; Ding, Melloni, Zhang, Tian, & Poeppel, 2016), to hierarchical cues such as context-free grammar structure (Brennan & Hale, 2019), as well as to the semantic dissimilarity between successive words (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018).

An important property of word sequences is that they can allow the prediction of an upcoming word, resulting in a word expectation. The degree to which a word can be predicted is referred to as precision and reflects the certainty with which a neural population generates its prediction. Predictions and precision are both closely related to putative implementations of predictive processing (Heilbron & Chait, 2018; Kanai, Komura, Shipp, & Friston, 2015; Feldman & Friston, 2010). Behavioral studies have indeed corroborated that the brain makes predictions about upcoming speech segments: Words can be better distinguished from noise when transition probabilities between words are high rather than low (Miller, Heise, & Lichten, 1951), and a highly expected word can be perceived as heard even when obscured by noise (Miller & Isard, 1963).

Neurophysiological research on ERPs elicited by a word in a sentence has shown that the brain response to a word reflects the word expectancy through modulation of the N400 response (Kutas & Hillyard, 1984). Although this response has not been found to be further modulated by the precision of the prediction (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007), precision can influence the neural power in the alpha and theta bands (Rommers, Dickson, Norton, Wlotko, & Federmeier, 2017). The power in the beta frequency band has been found to be reduced by semantic

[1]Imperial College London, [2]Maastricht University

Journal of Cognitive Neuroscience 32:1, pp. 155–166
https://doi.org/10.1162/jocn_a_01467

and syntactic violations and may therefore relate to word expectation as well (Kielar, Meltzer, Moreno, Alain, & Bialystok, 2014; Bastiaansen, Magyari, & Hagoort, 2010; Davidson & Indefrey, 2007). Gamma power has been observed to increase when a word is highly predictable but not when its predictability is low (Molinaro, Barraza, & Carreiras, 2013; Wang, Zhu, & Bastiaansen, 2012).

However, these prior studies on neural correlates of word expectancy and precision have focused on specific words in single sentences, contrasting words with high and low expectancy as well as with high and low precision. But natural speech often consists of many sentences, and the expectancy and the corresponding precision of successive words take a range of values that do not fall in only two classes of "high" and "low." It therefore remains unclear how neural responses to word expectancy and precision correlate with this graded variability.

Furthermore, assessing the cortical responses to the linguistic features of successive words in naturalistic stories allows to quantify the cortical tracking of these features. A recent investigation on word predictability and hierarchical structure in naturalistic speech used such an approach to show cortical tracking of word surprisal but did not investigate an influence of precision and did not investigate power modulation in higher frequency bands (Brennan & Hale, 2019; Frank & Willems, 2017).

Here, we therefore set out to investigate cortical tracking, including through power modulation in higher frequency bands, of word surprisal and the precision of word prediction in naturalistic stories. The surprisal of a word denotes the log-transformed conditional probability of a word based on the preceding context. The surprisal has been argued to relate to processing load (Levy, 2008) and predicts reading time (Frank, Otten, Galli, & Vigliocco, 2015; Smith & Levy, 2013). Precision is the inverse of the entropy of the conditional probability distribution over a close vocabulary set. We quantified word surprisal and precision from naturalistic stories using language modeling as estimated by a recurrent neural network and then related the obtained word features to EEG responses of volunteers who listened to the stories.

METHODS

Participants

Thirteen participants (aged 25 ± 3 years, six women) participated in the experiment. The volunteers were all right-handed native English speakers. They had no history of hearing or neurological impairment. All participants provided written informed consent. The experimental procedures were approved by the Imperial College Research Ethics Committee.

Experimental Design

We used naturalistic speech narratives in the participants' native language (English). The experiment consisted of one session in which we measured EEG responses to the short stories *Gilray's Flower Pot* and *My Brother Henry* by J. M. Barrie as well as *An Undergraduate's Aunt* by F. Anstey (Patten, 1910). The stimuli were sourced from the public domain librivox.org and were spoken by a male voice. The corresponding text was obtained from Project Gutenberg (www.gutenberg.org/ebooks/32846). The audio material was presented in 15 parts, each of which were 2.6 ± 0.43 min long. The total length of the stories was 40 min. After each part of a story, participants answered comprehension questions about what they just heard. These questions were presented as multiple-choice questions on a monitor. Participants were asked 30 questions in total.

Language Modeling

We used computational linguistics methods to quantify linguistic features in the stories. Specifically, we employed statistical language modeling to compute word frequency, entropy, and suprisal from the text of the stories.

Word frequency is a property of each individual word out of context, which was computed from Google *N*-grams by using only the unigram values. This word feature is an estimate of the unconditional probability of the occurrence of a word w, $P(w)$. We use the negative logarithm of this probability such that all our information-theoretical word features are expressed in the same unit.

Both entropy and surprisal follow from conditional probabilities of a particular word given the preceding words. We denote by $P(w_m|w_1, \ldots, w_{m-1})$ the conditional probability of the mth word in the sequence, w_m, given the previous $m - 1$ words $w_1, w_2, \ldots w_{m-1}$. Taking the negative logarithm of this probability yields the "surprisal" value $S(w_m)$ for that word,

$$S(w_m) = -\log(P(w_m|w_1, \ \ldots, \ w_{m-1})) \qquad (1)$$

The surprisal, also referred to as self-information or information content, quantifies the information gain that an upcoming word generates with respect to the prior sequence of words. It can be related to how unexpected a word is given the previous words in the sentence. Inasmuch as surprisal informs about expected words, *precision* relates to the confidence about the predictions made (Koelsch, Vuust, & Friston, 2018). A high precision translates into a high confidence about a word expectation, meaning that the word is predictable.

The entropy $E(m)$ of the prediction of the mth word w_m, that is, the uncertainty for predicting the word w_m from the context $(w_1, \ldots, w_{m-1})$, is given by the sum of the conditional probabilities for each possible word w_k, weighted by the logarithm of this probability. In other words, the entropy is the expected surprisal,

$$E(m) = \sum_{w_k} p(w_k|w_1, \ \ldots, \ w_{m-1}) \ \log[P(w_k|w_1, \ \ldots w_{m-1})] \qquad (2)$$

The precision of the mth word w_m follows as the inverse of entropy $1/E(m)$. We note that the precision of the mth word is not a function of that word itself but of the probability distribution of the words at that position.

The conditional probabilities for the different words in the sequence, given the preceding words, were computed through a recurrent neural network language model (Graves, 2013; Bengio, Ducharme, Vincent, & Jauvin, 2003). The network had a hidden layer with recurrent connections to encode previous input. Such networks are particularly useful for processing sequences and have previously been successfully applied to language modeling (Graves, 2013; Bengio et al., 2003). In particular, a recurrent neural network can capture long-term dependencies, of variable length, by encoding preceding words through its recurrent connection into the state of the hidden neurons. This is enabled by a careful balance between short- and long-term memory and means that there is, in principle, no limit on the number of preceding words that such a network can take into account (Pascanu, Mikolov, & Bengio, 2013). This contrasts with N-gram language models, for instance, that are limited to a context window of $N - 1$ words (Brown, Desouza, Mercer, Pietra, & Lai, 1992).

The network was implemented using the feature-augmented recurrent neural network language modeling toolkit (Mikolov, Kombrink, Burget, Černocký, & Khudanpur, 2011). To decrease the computational time required for training, this toolbox assigns words to classes and factorizes the output layer into a part that describes the probability of each class given the previous words, as well as another part that describes the probability of each word within a class given the previous words. This factorization yields a significant decrease in training time at a small cost to accuracy; importantly, the network still computes the probability of individual words following the previous words (Mikolov et al., 2011). We employed 300 classes. As an embedding layer, we used the pretrained global vectors for word representation trained on the Wikipedia 2014 and the Gigaword 5 data sets (Pennington, Socher, & Manning, 2014). The recurrent layer encompassed 350 hidden units. The source code was customized to compute the entropy of each word, a feat that the original code did not allow. The neural network was then trained on the text8 data set that consists of 100 MB of data from Wikipedia (Mahoney, 2011), using back propagation through time, truncated to five words with a starting learning rate of 0.1. The data were cleaned to remove punctuation, html tags, capitalization, and numbers before training. Because the network can only train well on words that appear frequently enough in the training data to allow meaningful training, we limited the vocabulary to the 35,000 most common words in the training data set. The remaining words were mapped to an "unknown" token. Infrequent words in the stories, such as compound nouns used for style, that appeared repeatedly throughout the stories did therefore not obscure the results.

The output of the recurrent neural network was obtained from a softmax function and could therefore be interpreted as the probability distribution for an upcoming word given the preceding words in the input sequence. The network was therefore trained to predict the next word, that is, to compute an output that was as close as possible to a probability distribution that was one for the actual upcoming word and zero for all remaining ones. The trained network was then run on the stories that the participants heard. Precision and surprisal of each word were determined from the network's computed probability distribution at the corresponding word through Equations (1) and (2).

Speech Features

To relate surprisal and entropy to the EEG data, we constructed a time series for each linguistic feature. We first aligned each word of the speech to the acoustic signal through forced alignment using the Prosodylab-Aligner software (Gorman, Howell, & Wagner, 2011). We thereby obtained the time at which each word began. To construct features for surprisal and for precision that were aligned with the speech stimuli, we assigned each of the time points where a new word started a spike of a magnitude that corresponded to the surprisal and precision of that word (Figure 1A). A similar procedure has been employed recently for assessing neural responses to the semantic dissimilarity of consecutive words (Broderick et al., 2018).

Because surprisal and precision are high-level linguistic features of speech, we sought to ascertain that any putative cortical tracking of them could not be explained by lower level features. To this end, we added three low-level speech features. First, cortical activity can track the onset of words, which can partly be based on changes in the acoustics at word boundaries and partly result from the brain's parsing of the acoustic signal to form discrete linguistic units (Brodbeck, Presacco, & Simon, 2018; Ding & Simon, 2014). To account for this onset response, we constructed a word onset feature as a series of spikes, each of which had unit amplitude and was located at the onset of a word. Second, we computed the word position within a sentence. The latter can be correlated with precision, as the entropy tends to decrease across words within the sentence. The word position feature therefore served as a control to ensure that the neural response to precision is distinct from any incremental processing occurring throughout a sentence. Third, the frequency of a word in a given language, outside its context, is a linguistic feature that acts as a prior probability for computing the probability of a word in a sequence (Brodbeck et al., 2018). Word frequency can also interfere with surprisal: Less frequent words may indeed often be more surprising. To capture the share of the neural response that could be explained away by word frequency, we included the latter as a third linguistic feature. This feature was computed by

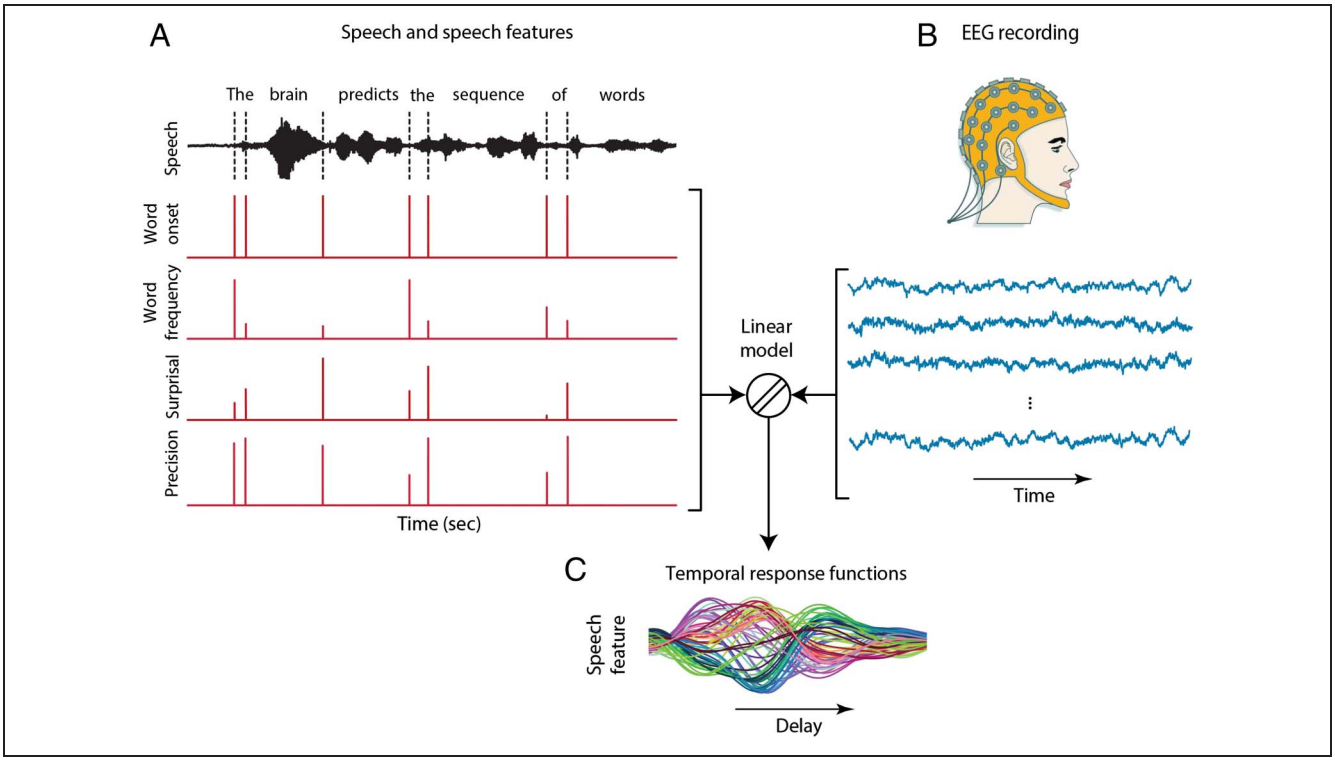

Figure 1. Experimental overview. (A) We employ continuous speech narratives and utilize speech processing as well as language modeling to extract acoustic and linguistic features, namely, word onset, word frequency, precision, and surprisal. (B) The participant's neural activity is recorded through EEG while they listen to the stories. (C) We extract TRFs for each of the four speech features through computing a linear model that estimates the EEG recordings from the speech features.

scaling the amplitude of the spike at each word onset by the negative logarithm of the frequency of the corresponding word. The logarithm was used such that word frequency and surprisal were expressed in the same units.

Finally, to investigate a possible modulating effect that precision may have on surprisal, we added an interaction term "Surprisal × Precision." This was computed by multiplying precision values with surprisal such that the interaction feature effectively stands as a confidence-weighted version of surprisal.

In summary, we computed five speech features: one acoustic feature, word onset, and four linguistic features, word position in its sentence, word frequency, precision, and surprisal. To those, we added the interaction term between surprisal and precision. Each feature was a time series of spikes, with each spike being located at the onset of a word. The amplitude of the spike was constant for the word onset feature. For each other feature, it was scaled to the corresponding value for each respective linguistic feature. All values of the different linguistic features were standardized to have unit variance and zero mean.

EEG Acquisition and Preprocessing

We recorded brain activity using 64 active electrodes (actiCAP, BrainProducts) and a multichannel EEG amplifier (actiCHamp, BrainProducts). The presented sound was recorded simultaneously through an acoustic adapter (Acoustical Stimulator Adapter and StimTrak, BrainProducts) and was used for aligning the EEG recordings to the audio signals. Both the EEG and the audio data were acquired at a sampling rate of 1 kHz. The left ear lobe was used as a reference for the EEG.

The EEG data were processed by first applying an anti-aliasing filter (Kaiser window, finite impulse response [FIR] filter, cutoff −6 dB at 125 Hz, transition bandwidth 50 Hz, order 130) and by downsampling the data to 250 Hz to reduce the computation time of subsequent operations. A high-pass filter (Hanning window, sinc Type I linear phase FIR filter, cutoff −6 dB at 0.3 Hz, transition bandwidth 0.15 Hz, order 5168) was then applied to every channel to remove nonstationary trends such as slow drifts and offsets. Bad channels were identified using the procedure clean_rawdata from the EEGLAB plugin ASR (Artifact Subspace Reconstruction); they were then removed and interpolated with spherical interpolation. All channels were then referenced to the channel average. We subsequently ran an independent component analysis (ICA) decomposition and removed artifacts from eye blink, eyes movement, as well as muscle motion by visual inspection of the ICA components. The cleaned data were low-pass filtered (Hamming window, linear phase FIR filter, cutoff −6 dB at 62 Hz, transition bandwidth 10 Hz, order 138) and further down-sampled to 125 Hz. The filtered EEG data therefore contained the broad frequency range from 0.3 to 62 Hz.

We computed temporal response functions (TRFs) from EEG data in several frequency bands. The TRFs followed from a linear forward model that expressed the EEG signal at each electrode as a linear combination of the speech features shifted by different latencies (Broderick et al., 2018; Ding & Simon, 2012). We used FIR Type I filters, designed with the synced windowed method, and employing a hamming window. We filtered the EEG data in several frequency bands of interest: delta band (low-pass filter, cutoff at 4.5 Hz, filter order 132), theta band (band-pass filter, cutoff frequencies at 4 Hz and 8 Hz, order 206), alpha band (band-pass filter, cutoff frequencies at 8 and 12 Hz, order 206), beta band (band-pass filter, cutoff 20 Hz and 30 Hz, order 82), and gamma band (cutoff at 30 and 60 Hz, order 164). For every frequency band other than delta, we computed the power modulation by taking the absolute value of the Hilbert transform of the band passed data and further band-pass filtered it between 0.5 and 20 Hz (filter order 824) to remove the DC offset and higher frequencies that do not occur in the speech features.

EEG Data Analysis

To relate the speech features to the EEG data, we used a linear spatiotemporal forward model that reconstructed the EEG recordings from the acoustic feature and the linguistic features, shifted by different delays (Figure 1). Such an approach has recently been used successfully for assessing the cortical tracking of the speech envelope, phonemic information, as well as semantic dissimilarity of words in speech (Broderick et al., 2018; Di Liberto et al., 2015; Ding & Simon, 2012). The coefficients resulting from this regression constitute the TRFs that inform on the brain's response to each feature at different latencies.

In particular, the forward model sought to express the preprocessed EEG recordings $\{x_i(t_n)\}_{j=1}^{N}$ of the $N = 64$ channels at each time instance t_n through the time series $\{y_j(t_n-\tau_k)\}_{j=1}^{F}$ of the $F = 6$ speech features word onset, word frequency, word position, word precision, word surprisal, and the product of surprisal and entropy, shifted by T different delays $\{\tau_k\}_{k=1,}^{T}$

$$\grave{x}_i(t_n) = \sum_{j=1}^{6}\sum_{k=1}^{T}\beta_{ij}(\tau_k)y_j(t_n-\tau_k) \qquad (3)$$

We hereby considered equally spaced delays $\{\tau_k\}_{k=1}^{T}$ that ranged from -400 to 1100 msec. At the sampling rate of 125 Hz, this yielded a number of $T = 188$ lags. The obtained estimate for the EEG channel i is denoted by $\grave{x}_i$. The coefficient $\beta_{ij}(\tau_k)$ is the TRF for the ith EEG channel and speech feature j at the latency τ_k. The preprocessed EEG recording $\{x_i(t_n)\}_{i=1}^{N}$ was either the EEG signal in the delta band or the power of the EEG signal in the higher frequency bands. We computed the TRFs for each participant separately, leading to a set of TRFs on which we could apply group-level statistical analysis as described below. We then also computed the population average of the participant-specific TRFs; the population averages are shown in the figures.

The different speech features that we employed were partly correlated. The largest correlation emerged between surprisal and the interaction term "Surprisal × Precision," at a value of .61. We wondered if these correlations would hinder the EEG analysis, and in particular, if they would obscure the neural responses to the individual speech features through the linear regression analysis, an issue known as multicollinearity (Chatterjee & Hadi, 2015; Kumar, 1975). A high multicollinearity between features could result in higher variance or leakage between the coefficient $\beta_{ij}(\tau_k)$. However, the Frisch–Waugh–Lovell theorem from econometrics states that linear regression based on correlated features yields the same results as when the features are first orthogonalized, that is, decorrelated (Lovell, 2008; Frisch & Waugh, 1933). In addition, in our implementation of the multiple linear regression, we used a singular value decomposition of the design matrix of time-lagged features, resulting in transformed features that were mutually uncorrelated (Klema & Laub, 1980). The correlation of the features was therefore not problematic. The only issue that multicolinearity can cause is significantly increased variance for each $\beta_{ij}(\tau_k)$ estimate, which typically emerges when the variance inflation factor is above 5. For our speech features, we obtained variance inflation factors between 1.22 and 2.25, indicating that increased noise due to correlated features is not an issue.

As an additional control that our TRFs did not contain leakage from responses to different features, we developed a null model that was employed to assess the statistical significance of the actual TRFs (see below). The null model was constructed such that a potential leakage between features would appear similarly both in the actual model and in the null model and therefore would not result in statistically significant results. It follows that any statistically significant part in the TRFs that we obtained did not result from leakage between the features.

Statistical Significance

To determine the statistical significance of the estimated TRFs, we determined chance-level TRFs as a null model. The chance-level TRFs were computed by constructing unrelated speech features and by relating these to the EEG recordings in the same way as for the computation of the actual TRFs. To establish chance-level linguistic TRFs, only the linguistic information of interest contained in the spike amplitude of the speech features but not the acoustic information in the spike timing needed to be unrelated to the EEG. We therefore constructed unrelated speech features by keeping the timing of the spikes identical to those in the true model. The speech feature that described word onsets was therefore not altered.

However, we changed the amplitude of the spikes for the other linguistic speech features by taking their values from an unrelated story, that is, a story that was not aligned with the EEG data. To obtain a large number of null models, we considered permutations of our 15 story parts. Through permutating entire story parts and not the order of individual words, the statistical relationship between the linguistic features of successive words was conserved. Because we kept the timing of the spikes in the null model as in the actual stories, the obtained null model could only be used to determine the significance of the neural responses to the linguistic features, but not for those to the acoustic word onset.

The actual TRFs were then analyzed for statistical significance through comparison to 1000 null models. The comparison was obtained from a permutation test together with cluster-based correction for multiple comparison (Oostenveld, Fries, Maris, & Schoffelen, 2011), where only clusters of at least four electrodes were kept. Specifically, we used the function *spatio_temporal_cluster_test* from the MNE python library. The statistic for each model coefficient, at each electrode and each lag, was computed using the empirical distribution formed by values from the null models, setting the threshold at the 99th percentile of the null distribution. The cluster-level p values were computed, and we considered only clusters with a p value greater than .05/10. We hereby used the Bonferroni correction to account for the 10 different tests that reflected the different frequency bands and the different linguistic features.

Data Availability

The EEG data from all participants, together with the corresponding speech features, are available on figshare.com (https://doi.org/10.6084/m9.figshare.9033983.v1). An exemplary script for computing TRFs can be obtained from figshare as well (https://doi.org/10.6084/m9.figshare.9034481.v1).

RESULTS

Behavioral Assessment

We first assessed to what degree the participants understood the stories through asking them comprehension questions. These questions were answered with an average of 96% accuracy, evidencing that the volunteers consistently understood the speech and paid attention.

Cortical Tracking of Acoustic and Linguistic Speech Features

The cortical tracking of the speech features can be found in different frequency bands. First, because all four features relate to words, the frequency range of the features is similar to the rate of words in speech. The latter is about 1–4 Hz and corresponds to the delta frequency range. Cortical activity at low frequencies, including the delta frequency band, can therefore be evoked by or entrain to the rhythm set by the acoustic and linguistic word features. Second, the amplitude of the neural activity in higher frequency bands can be modulated by the speech features. This may, in particular, occur for the theta band (4–8 Hz), the alpha band (8–12 Hz), the beta frequency band (20–30 Hz), and the gamma frequency band (30–100 Hz), the power of which can be modulated by prediction in sentence comprehension (Wang et al., 2012; Weiss & Mueller, 2012; Bastiaansen et al., 2010; Bastiaansen & Hagoort, 2006).

We started by quantifying the neural tracking of the word features at low frequencies. We found neural responses to word frequency between delays of 300 and 610 msec (Figure 2). The topographic plots of the responses show large differences between the temporal scalp areas on the one hand and the parietal and occipital areas on the other hand.

Importantly, we found significant responses to the word surprisal around a delay of 450 msec (Figure 2). These responses emerged predominantly in the EEG channels on the temporal and occipital scalp areas and were lateralized on the left hemisphere. Precision was tracked by cortical activity at delays of around 100 msec and around 500 msec. Moreover, we observed a significant neural response to the interaction of surprisal and precision, at an earlier latency of around 400 msec and at a longer latency of around 1000 msec.

We also computed the modulation of the power in the theta band, the alpha band, the beta band, as well as the gamma band by the acoustic and linguistic features (Figures 4 and 5). Although the power in the alpha band was not significantly related to the linguistic features, the power in the theta band was shaped by word frequency at delays of around 300 msec and around 1000 msec (Figure 3). Furthermore, the power in the theta band was significantly decreased by precision at delays of about 700 msec.

The power in the beta band correlated positively with surprisal at delays of around 700 and 1000 msec (Figure 4). At the latter delay, the influence of surprisal was strongest at the left temporal channels. Moreover, the power in the beta band was modulated by precision at a delay of about 700 msec, with the main contributions coming from the occipital channels.

The power in the gamma band was increased by words with higher surprisal at long latency of around 1000 msec, mainly for the left temporal channels (Figure 5). The interaction of surprisal and precision shaped the gamma power as well, at the early delay of about 0 msec.

DISCUSSION

We have shown that cortical activity tracks the surprisal of words in speech comprehension. Such cortical tracking

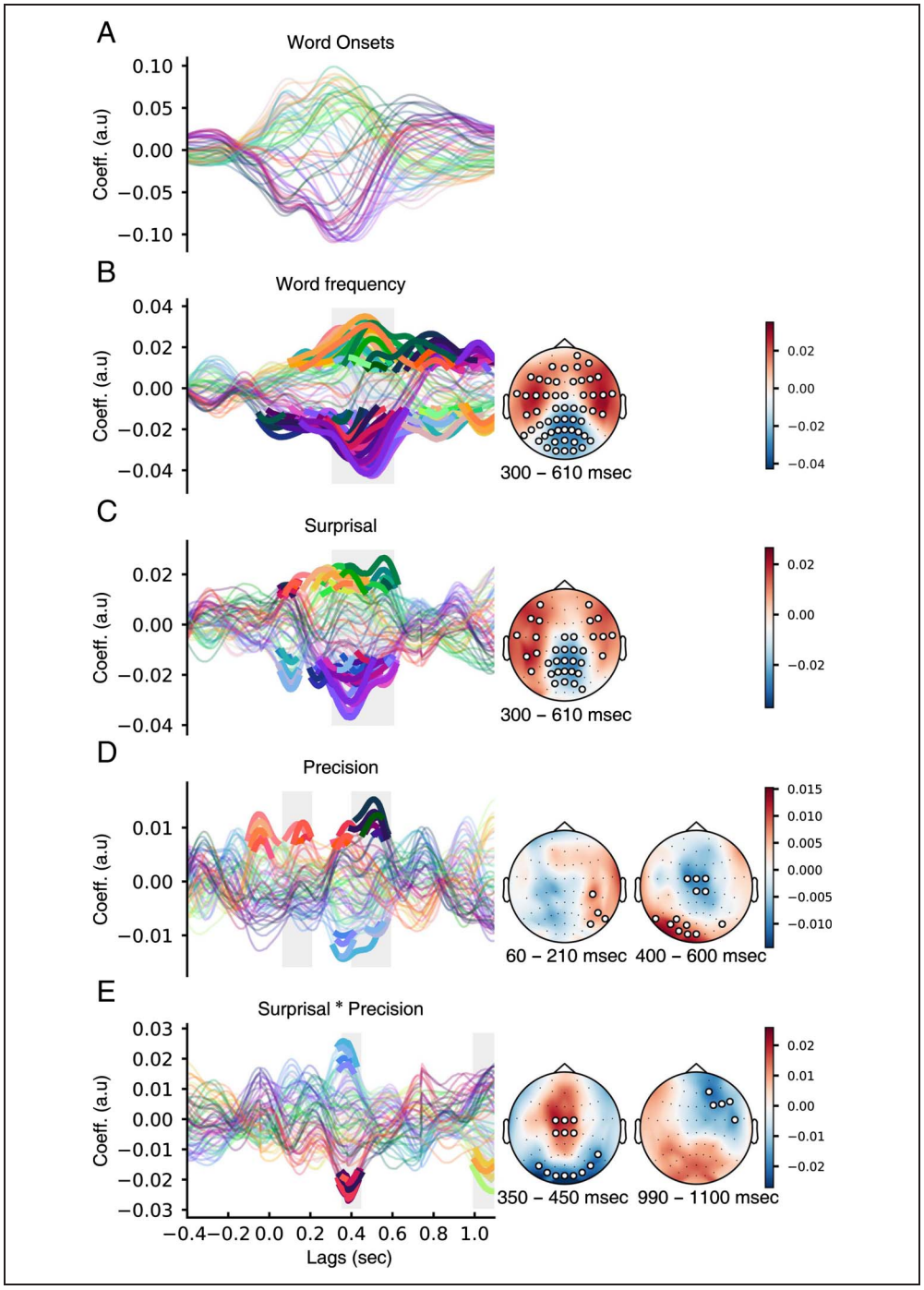

Figure 2. TRFs for acoustic and linguistic speech features. The TRFs for each electrode are shown in bold at time instances where they are significant compared with a null model that is based on shuffled data. EEG channels that yield a significant response within a particular range of delays, highlighted in gray, are indicated in white in the topographic plots. (A) The responses to the word onset appear as insignificant due to the construction of the null model. (B, C) We obtain significant neural responses to word frequency as well as surprisal for delays around 400 msec (D) Significant neural responses to precision arise around delays of 100 msec as well as around 500 msec. (E) The interaction between surprisal and precision leads to a neural response at a delay of 400 msec as well as at a long delay of 1000 msec.

has emerged at low frequencies, that is, within the delta band that encompasses a similar frequency range as the rate of words in speech. Importantly, we found that the neural activity in the faster theta, beta, and gamma frequency bands tracks surprisal as well. These frequency bands have previously been suggested to be involved in the bottom–up and top–down propagation of predictions and prediction errors (Lewis & Bastiaansen, 2015).

We have further demonstrated that the cortical tracking of word surprisal is modulated by precision: The interaction between surprisal and precision leads to responses both in the slow delta band as well as in the power of the faster gamma band. In particular, word predictions that are made with high precision but then lead to large surprisal cause an increased gamma power at zero lag. However, as opposed to a previous study on ERPs, we did not observe a significant effect in the theta or alpha bands (Rommers et al., 2017). This difference may be due to our use of naturalistic stimuli and the inclusion of all words in the analysis, whereas the previous study used specialized sentences with final words that had either high or low surprisal and either high or low precision.

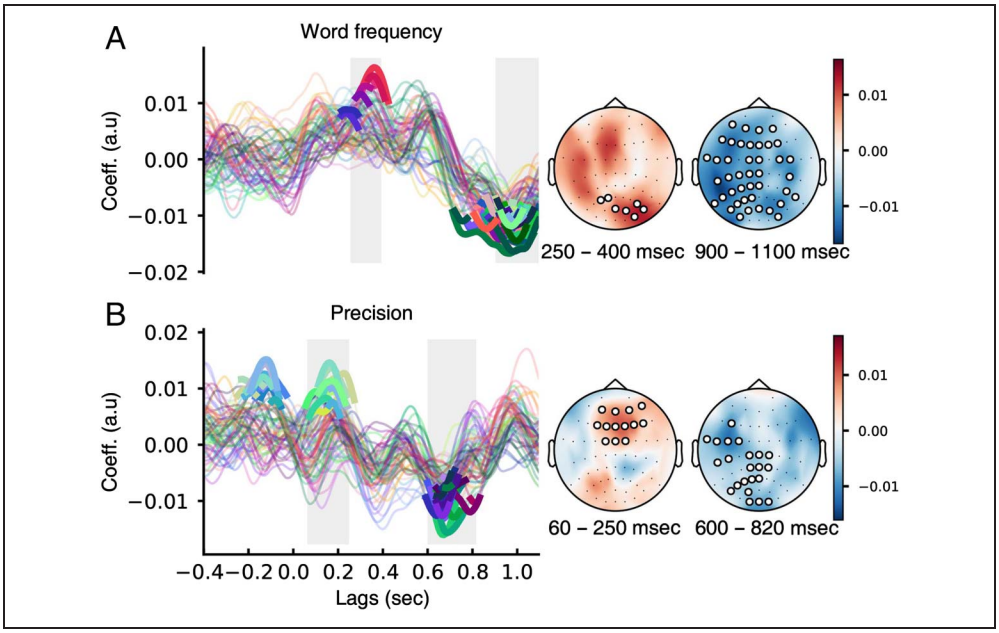

Figure 3. Neural responses in the theta frequency band. (A) Word frequency is positively correlated to theta power at a delay of 300 msec and is negatively correlated at a delay of 1000 msec. (B) Words that can be predicted with higher precision lead to an increased theta power at 150 msec and a decreased theta power at a latency of 700 msec.

The cortical tracking of surprisal may indicate predictive processing by the brain. Predictive processing is a framework for perception in which it is assumed that the brain infers hypotheses about a sensory input by generating predictions of its neural representations and that the hypotheses are constantly updated as new sensory information becomes available (Kanai et al., 2015; Bendixen, SanMiguel, & Schröger, 2012; Friston, 2010; Friston & Kiebel, 2009). In particular, the surprisal of a word reflects a prediction error, a key quantity in the framework of predictive coding (Friston, 2010). However, the expectancy of a word based on previous words also correlates with the plausibility of a word in a particular context (Nieuwland et al., 2019; DeLong, Quante, & Kutas, 2014). Further studies are therefore required to disentangle neural correlates of actual word prediction from those that do not require predictive processing, such as word plausibility.

The surprisal of a word can reflect both its semantic as well as syntactic information, and previous investigations into the neurobiological mechanisms of language comprehension have manipulated both independently (Henderson, Choi, Lowder, & Ferreira, 2016; Humphries, Binder, Medler, & Liebenthal, 2006). In contrast, our

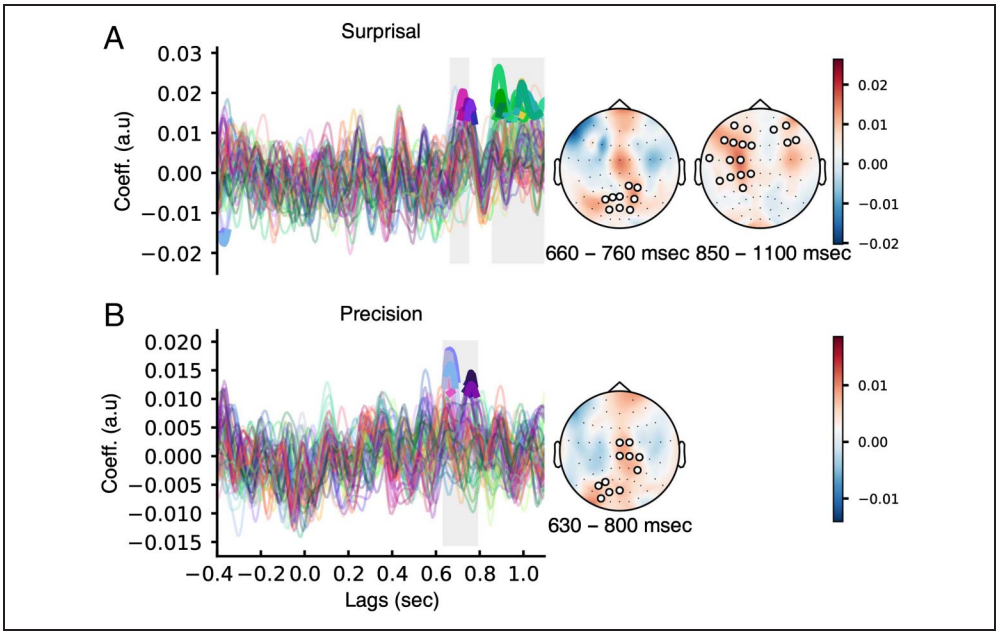

Figure 4. Neural responses in the beta frequency band. (A) There are significant neural responses to surprisal, emerging at delays of 700 and 1000 msec. (B) Precision causes an increased power in the beta band activity around a delay of 700 msec.

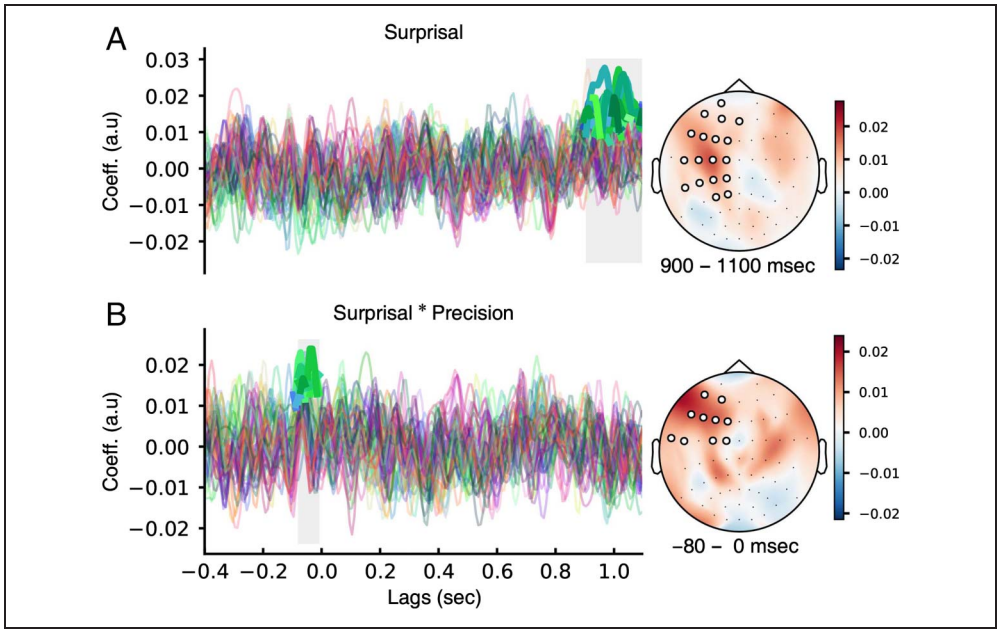

Figure 5. Tracking of surprisal by gamma band activity. (A) The gamma activity is decreased at around 1000 msec, mostly in the left temporal and frontal scalp areas. (B) The interaction between precision and surprisal leads to a modulation of the gamma power at the latency of around 0 msec. This modulation occurs predominantly for left temporal and frontal channels as well.

approach has taken a naturalistic and holistic approach to surprisal; we employed natural speech without manipulations combined with statistical learning of a rich variety of natural language cues through a recurrent neural network. Because the neural network infers both syntactic rules as well as semantic information from the training of the speech material, the reported neural response to word surprisal can reflect both semantic as well as syntactic information (Collobert et al., 2011).

It is instructive to compare the reported neural responses to surprisal to the well-characterized event-related responses that can be elicited by violations of semantics, syntax, or morphology in sentences. In particular, semantic violations can cause the N400 response, a negativity at 200–500 msec at the central and parietal scalp area (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980). Syntactic anomalies due to ungrammaticality or temporary misanalysis elicit the P600, a broad positive potential that is located at the posterior scalp area and arises around 600 msec after the anomaly (Hagoort & Brown, 2000; Friederici, Pfeifer, & Hahne, 1993). More specific syntactic anomalies can lead to negative potentials that occur anteriorly and that can be left lateralized, either occurring at 300–500 msec ((L)AN) or earlier, at 125–150 msec (ELAN; Steinhauer & Drury, 2012; Friederici, 2002; Van Den Brink, Brown, & Hagoort, 2001; Rösler, Pechmann, Streb, Röder, & Hennighausen, 1998).

These ERPs do presumably not reflect the activation of single static neural sources, but rather waves of neural activity that propagate in time across different brain areas (Kutas & Federmeier, 2011; Tse et al., 2007; Maess, Herrmann, Hahne, Nakamura, & Friederici, 2006). In the case of the N400, for instance, this wave of activity starts at about 250 msec in the left superior temporal gyrus and then propagates to the left temporal lobe by 365 msec as well as to both frontal lobes by 500 msec (Van Petten & Luka, 2006; Halgren et al., 2002; Helenius, Salmelin, Service, & Connolly, 1998). A recent theory suggests that this wave of activity reflects reverberating activity within the inferior, middle, and superior temporal gyri that corresponds to the activation of lexical information, the formation of context and the unification of an upcoming word with the context (Baggio & Hagoort, 2011).

The spatiotemporal characteristics of the responses to surprisal that we have measured here share certain similarities with these ERPs. In particular, we have found neural responses to surprisal at latencies between 300 and 600 msec. These responses show a central-parietal negativity that is reminiscent of the N400. However, other features of the neural responses that we describe here appear distinct from these ERPs. The neural response to surprisal in the delta band at the latency of 600 msec does, for instance, not display the posterior positivity of the P600. Moreover, we have identified late responses around 700 and 1000 msec. We have also shown that neural responses to surprisal arise in various frequency bands, beyond the delta band that matters for the ERPs. However, a further comparison of the neural response to surprisal to the related ERPs is hindered by the lack of spatial resolution offered by EEG recordings. Future neuroimaging studies using intracranial recordings or magnetoencephalography (MEG) may localize the sources of the neural response to surprisal that we have measured here and quantify potential shared sources with the ERPs.

The difference of the cortical tracking of surprisal to the well-known neural correlates of semantic, syntactic, or morphological anomalies and, in particular, the late responses at a delay of around 1 sec may come as a result

of our use of natural speech that differs from the artificially constructed and tightly controlled stimuli used to measure ERPs. First, in our experiment, the participants encountered no violations of semantics, syntax, and morphology but instead heard naturalistic speech, within which the words occurred in context. Second, our stimuli did not contain artificial manipulations of word surprisal or precision. Instead of altering the stimuli, we focused on quantifying surprisal and precision as they varied naturally in the presented stories. Third, we assessed the responses to surprisal and precision at each word in the story and hence for words in every sentence position, rather than for words at a particular position within each sentence. Because we accounted for word position through a corresponding control feature, we avoided the possibility of sentence position having an effect on the results (Bastiaansen et al., 2010). Fourth, we did not employ isolated sentences but continuous stories so that information of integration occurred over timescales exceeding a few seconds.

Although our EEG recordings showed the cortical tracking of surprisal in different frequency bands, they did not allow us to precisely localize the sources of the activity in the cortex. Pairing EEG with fMRI or employing MEG may allow to add spatial information to the temporal tracking that we have assessed here. A recent fMRI study, for instance, found that the left inferior temporal sulcus, the bilateral posterior superior temporal gyri, and the right amygdala responded to surprisal during natural language comprehension, whereas the left ventral premotor cortex and the left inferior parietal lobule responded to entropy (Willems, Frank, Nijhof, Hagoort, & van den Bosch, 2015). Another recent MEG measurement of the brain's natural speech processing found that entropy and surprisal play a role in the assembly of phonemes into words and involve brain areas such as core auditory cortex and the STS (Brodbeck et al., 2018). Combining the temporal precision of EEG with the spatial precision of fMRI or harnessing the ability of MEG to locate neural sources temporally and spatially will allow to further clarify the spatiotemporal mechanisms of natural language comprehension in the brain.

In summary, we showed that neural responses to word surprisal can be measured from EEG responses to naturalistic stories. Our results demonstrate that both the slow delta band as well as the power in higher frequency bands, in particular the beta and gamma bands, are shaped by surprisal. Moreover, we also showed that the neural response to surprisal is modulated by the precision of a prediction. In particular, predictions made with high precision, which lead to high surprisal modulate gamma power in the left temporal and frontal scalp areas. In addition, we also demonstrated that neural activity in the delta, theta, and beta frequency bands is shaped by the precision of word prediction directly. These responses arise at different latencies and at different scalp areas, suggesting a rich spatiotemporal dynamics of neural activity related to word prediction.

Acknowledgments

This research was supported by Wellcome Trust grant 108295/Z/15/Z, by EPSRC grants EP/M026728/1 and EP/R032602/1, as well as in part by the National Science Foundation under grant no. NSF PHY-1125915.

Reprint requests should be sent to Tobias Reichenbach, Department of Bioengineering and Centre for Neurotechnology, Imperial College London, South, Kensington Campus, SW7 2AZ, London, United Kingdom, or via e-mail: reichenbach@imperial.ac.uk.

REFERENCES

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*, 1338–1367.

Bastiaansen, M., & Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. *Progress in Brain Research*, *159*, 179–196.

Bastiaansen, M., Magyari, L., & Hagoort, P. (2010). Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. *Journal of Cognitive Neuroscience*, *22*, 1333–1347.

Bendixen, A., SanMiguel, I., & Schröger, E. (2012). Early electrophysiological indicators for predictive processing in audition: A review. *International Journal of Psychophysiology*, *83*, 120–131.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS One*, *14*, e0207741.

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *Neuroimage*, *172*, 162–174.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*, 803–809.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based *n*-gram models of natural language. *Computational Linguistics*, *18*, 467–479.

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. Hoboken, NJ: Wiley.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Davidson, D. J., & Indefrey, P. (2007). An inverse relation between event-related and time–frequency violation responses in sentence processing. *Brain Reseach*, *1158*, 81–92.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*, 2457–2465.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *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. *Journal of Neuroscience*, *38*, 1178–1188.
Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, U.S.A.*, *109*, 11854–11859.
Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311.
Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84.
Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.
Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.
Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*, 1192–1203.
Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*, 78–84.
Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, *1*, 183–192.
Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, *1*, 387–401.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.
Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London: Series B: Biological Sciences*, *364*, 121–1221.
Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*, 511–517.
Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Journal of the Canadian Acoustical Association*, *39*, 192–193.
Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.
Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia*, *38*, 1531–1549.
Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., et al. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *Neuroimage*, *17*, 1101–1116.
Heilbron, M., & Chait, M. (2018). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*, *389*, 54–73.
Helenius, P., Salmelin, R., Service, E., & Connolly, J. F. (1998). Distinct time courses of word and context comprehension in the left temporal cortex. *Brain*, *121*, 1133–1142.
Henderson, J. M., Choi, W., Lowder, M. W., & Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage*, *132*, 293–300.
Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience*, *18*, 665–679.
Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *eLife*, *4*, e06213.
Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philophical Transancations of the Royal Society of London: Series B: Biological Science*, *370*, 20140169.
Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, *16*, e2004473.
Kielar, A., Meltzer, J. A., Moreno, S., Alain, C., & Bialystok, E. (2014). Oscillatory responses to semantic and syntactic violations. *Journal of Cognitive Neuroscience*, *26*, 2840–2862.
Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, *25*, 164–176.
Koelsch, S., Vuust, P., & Friston, K. (2018). Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences*, *23*, 63–77.
Kumar, T. K. (1975). Multicollinearity in regression analysis. *Review of Economics and Statistics*, *57*, 365–366.
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.
Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*, 279–292.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.
Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155–168.
Lovell, M. C. (2008). A simple proof of the FWL theorem. *Journal of Economic Education*, *39*, 88–91.
Maess, B., Herrmann, C. S., Hahne, A., Nakamura, A., & Friederici, A. D. (2006). Localizing the distributed language network responsible for the N400 measured by MEG during auditory sentence processing. *Brain Research*, *1096*, 163–172.
Mahoney, M. (2011). *About the test data*. Retrieved from mattmahoney.net/dc/textdata.html.
Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. Paper presented at the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*, 329–335.
Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, *2*, 217–228.
Molinaro, N., Barraza, P., & Carreiras, M. (2013). Long-range neural synchronization supports fast and efficient reading: EEG correlates of processing expected words in sentences. *Neuroimage*, *72*, 120–132.
Nieuwland, M., Barr, D., Bartolozzi, F., Busch-Moreno, S., Donaldson, D., Ferguson, H. J., et al. (2019). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. https://www.biorxiv.org/content/10.1101/267815v4.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. Paper presented at the 30th International Conference on International Conference on Machine Learning, Atlanta, GA.

Patten, W. (1910). *International short stories* (Vol. 2). Aurora, IL: P.F. Collier & Son.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. Paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.

Rommers, J., Dickson, D. S., Norton, J. J., Wlotko, E. W., & Federmeier, K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience*, *32*, 576–589.

Rösler, F., Pechmann, T., Streb, J., Röder, B., & Hennighausen, E. (1998). Parsing of sentences in a language with varying word order: Word-by-word variations of processing demands are revealed by event-related brain potentials. *Journal of Memory and Language*, *38*, 150–176.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, *120*, 135–162.

Tse, C.-Y., Lee, C.-L., Sullivan, J., Garnsey, S. M., Dell, G. S., Fabiani, M., et al. (2007). Imaging cortical dynamics of language processing with the event-related optical signal. *Proceedings of the National Academy of Sciences, U.S.A.*, *104*, 17157–17162.

Van Den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience*, *13*, 967–985.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language*, *97*, 279–293.

Wang, L., Jensen, O., Van den Brink, D., Weder, N., Schoffelen, J. M., Magyari, L., et al. (2012). Beta oscillations relate to the N400m during language comprehension. *Human Brain Mapping*, *33*, 2898–2912.

Wang, L., Zhu, Z., & Bastiaansen, M. (2012). Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension. *Frontiers in Psychology*, *3*, 187.

Weiss, S., & Mueller, H. M. (2012). "Too many betas do not spoil the broth": The role of beta brain oscillations in language processing. *Frontiers in Psychology*, *3*, 201.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, *26*, 2506–2516.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*, 980–991.