Cerebral Cortex, September 2021;31: 4300–4313

https://doi.org/10.1093/cercor/bhab087
Advance Access Publication Date: 24 April 2021
Original Article

ORIGINAL ARTICLE

Representational Pattern Similarity of Electrical Brain Activity Reveals Rapid and Specific Prediction during Language Comprehension

Ryan J. Hubbard[1] and Kara D. Federmeier[1,2,3]

[1]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA, [2]Department of Psychology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA and [3]Program in Neuroscience, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA

Address correspondence to Ryan J. Hubbard, 405 N Mathews Ave, Urbana, IL 61801, USA. Email: rjhubba2@illinois.edu

Abstract

Predicting upcoming events is a critical function of the brain, and language provides a fertile testing ground for studying prediction, as comprehenders use context to predict features of upcoming words. Many aspects of the mechanisms of prediction remain elusive, partly due to a lack of methodological tools to probe prediction formation in the moment. To elucidate what features are neurally preactivated and when, we used representational similarity analysis on previously collected sentence reading data. We compared EEG activity patterns elicited by expected and unexpected sentence final words to patterns from the preceding words of the sentence, in both strongly and weakly constraining sentences. Pattern similarity with the final word was increased in an early time window following the presentation of the pre-final word, and this increase was modulated by both expectancy and constraint. This was not seen at earlier words, suggesting that predictions were precisely timed. Additionally, pre-final word activity—the predicted representation—had negative similarity with later final word activity, but only for strongly expected words. These findings shed light on the mechanisms of prediction in the brain: rapid preactivation occurs following certain cues, but the predicted features may receive reduced processing upon confirmation.

Key words: comprehension, EEG, language, prediction, RSA

Introduction

Theories of cognition and neural functioning increasingly build in an important role for anticipatory processing—that is, prediction. Indeed, some have postulated that a core mechanism of neural coding involves higher level cortical systems in the brain attempting to predict and explain input at lower levels in a hierarchical fashion (predictive coding; Friston and Kiebel 2009). One area that has proven to be a particularly rich testing ground for understanding the import—and limitations—of predictive processing is language comprehension. When listening to or reading language, people can use contextual cues and prior knowledge to generate predictions about upcoming words in order to support rapid and efficient comprehension and communication (Federmeier 2007; Kutas et al. 2011; Kuperberg and Jaeger 2016; Pickering and Gambi 2018). The contents of these predictions can be multifaceted in nature, including orthographic (Laszlo and Federmeier 2009; Kim and Lai 2012), phonological (DeLong et al. 2005; Vissers et al. 2006), semantic (Federmeier and Kutas 1999; Lau et al. 2013), and morphosyntactic (Van Berkum et al. 2005; Dikker et al. 2010) features of words. However, such anticipatory processes are not always engaged (Wlotko and Federmeier 2015; Huettig and Guerra 2019), suggesting that the brain flexibly allocates resources to predict information to the extent that the environment allows it and as a function of the utility of those predictions for the task at hand.

Studies of prediction in language using behavioral (Schwanenflugel and LaCount 1988; Hess et al. 1995) and eyetracking (Ehrlich and Rayner 1981; Altmann and Kamide 1999; Staub and Clifton 2006) measures combine with a sizeable literature that has focused on neural responses to predictable and

unpredictable words as measured via electroencephalography (EEG), including event-related potentials (ERPs; Federmeier et al. 2007; Szewczyk and Schriefers 2018; Thornhill and Van Petten 2012) or magnetoencephalography (MEG; Dikker and Pylkkänen 2011; Wang, Hagoort, et al. 2018a; Wang, Kuperberg, et al. 2018b). This work has established that both the constraint of a context (i.e., how much it narrows expectations and allows a strong, consistent prediction) and the probability of the encountered word in its context modulate brain responses (Wlotko and Federmeier 2012; DeLong et al. 2014). Effects can even be seen on determiners or modifiers, when these have specific gender or phonological characteristics (e.g., "*a* kite") that are consistent or inconsistent with an anticipated noun ("*an* … kite"; DeLong et al. 2005; Szewczyk and Schriefers 2013; Wicha et al. 2003), as well as within 150 ms following word onsets that can differentiate between words with many possible continuations and words with few (Söderström et al. 2016; Roll et al. 2017).

Based on this work, the idea that comprehenders often engage probabilistic predictive mechanisms when language stimuli are encountered has become well-accepted (Kuperberg and Jaeger 2016). However, most extant work has assessed prediction by examining its consequences, measuring brain responses after a more or less predictable word (or a determiner/modifier carrying those features) has been encountered. Direct measurement of the prediction process itself—that is, the actual preactivation of the features—has been more elusive. As a consequence, there remain important open questions about when preactivation of various types of features might occur and what cues those preactivations (e.g., Huettig 2015).

To try to instead capture processing at the time that predictions are being made, some work has examined event-related activity differences elicited by a verb or adverb, as a function of whether that word does or does not afford a prediction for a target, sentence-final word (Maess et al. 2016; Freunberger and Roehm 2017). These studies report more negative N400s for words that carry more information about the target and, correspondingly, afford stronger predictions. However, it is unclear if these responses do or do not reflect preactivation of specific features of upcoming information. Other studies have employed novel paradigm manipulations in order to examine anticipatory processing. For instance, León-Cabrera et al. (2017) presented participants with sentences that varied in contextual constraint. By imposing a 1000-ms delay before the target word, they were able to observe slow negative potentials that were sensitive to constraint. Dikker and Pylkkänen (2013) implemented a picture-noun matching task to examine the preactivation of lexical features, in which more or less predictive pictures preceded related nouns; they found MEG activity differences based on predictability 400 ms prior to noun onset. However, it is unclear if the effects in these studies arise due to the unique demands of the task, and thus if the same processes would be observed in more natural language comprehension settings—that is, during sentence reading.

Examining ERP and MEG responses to prior words or time windows in isolation presents difficulties in separating predictive processing for the upcoming information from reactive processing of the preceding information, or more general effects of constraint. An optimal method for examining preactivation would consider both neural activity prior to and following the target stimulus to determine if there is similarity in neural processing, and if that similarity varies with predictability. If the brain preactivates features of upcoming words, then patterns of neural activity specifically related to processing that word should be present in advance, and comparing these patterns should reveal similarity graded by constraint and match to expectation. Representational similarity analysis (RSA) presents a promising solution (Kriegeskorte et al. 2008; Cichy and Pantazis 2017). With this technique, multivariate patterns of neural activity are compared with a correlational approach, which can be performed across the neural time-series or across electrodes. This method allows not only for identification of both temporal and spatial patterns of similarity but also for detection of more subtle but statistically separable neural states than may be found with more conventional ERP analyses (Cichy et al. 2015).

A recent study used RSA of MEG data to investigate the preactivation of semantic features in a language comprehension paradigm (Wang, Kuperberg, et al. 2018b). Specifically, participants read strongly constraining sentences that were constructed in pairs, such that within-pair sentences predicted the same sentence-final critical word (e.g., "In the crib there is a sleeping …" and "In the hospital there is a newborn …" both predict the word "baby") and between-pair sentences did not. A greater increase in neural similarity was found for within-pair sentences, suggesting greater similarity of neural patterns across sentences wherein the same word was predicted. However, it is unclear whether this was entirely due to prediction, or at least partly reflected that the brain was in a more similar state due to the shared semantics of within-pair sentences. Additionally, even "pseudo-repetitions" of words that were expected but never presented can lead to a lingering representation in the brain, despite intervening sentences (Rommers and Federmeier 2018a), which may have influenced pattern similarity.

In this paper, we specifically target preactivation by comparing pre-final word activity with postfinal word activity in sentences that varied in constraint and had expected or unexpected endings. We thus circumvent the issue of semantic similarity across sentences that yield similar predictions, because, in this case, the pattern comparisons for expected and unexpected words are within the same sentence context. If the input of the pre-final word cues the preactivation of features of the final word, then some aspects of the neural representation of the final word should appear during the processing of the pre-final word, which will be detected with RSA. Critically, if this correlation is, indeed, due to prediction-related preactivations, then similarity for expected words should be greater when sentential constraint and the corresponding cloze probability of the sentence-final word are higher, as more strongly constraining sentences allow for stronger and/or more specific predictions. Moreover, similarity should be reduced or potentially abolished when the ending is unexpected, as the neural representation of the final word may no longer match with the preactivated representation. Finally, the timing of prediction generation can be examined by assessing the similarity of final word activity patterns with patterns elicited by words preceding the pre-final word.

We also employ a time generalization analysis in order to examine the time-course or development of predictions over time (King and Dehaene 2014; Heikel et al. 2018). For instance, pattern similarity may increase gradually across time as the onset of the target word approaches or may come on and offline more rapidly. Additionally, this analysis method allows us to probe the fate of predicted representations after encountering the predicted word itself. Representations of words that were previously predicted have been found, downstream, to be impoverished compared with unpredicted words, suggesting later processing of word representations differs based on predictability (Rommers and Federmeier 2018b; Hubbard et al. 2019).

This difference in processing may be observable by analyzing representational similarity of pre-final word activity with later time windows of postfinal word activity.

Materials and Methods

This paper uses novel techniques to reanalyze data from the study reported in Federmeier et al. (2007). Further specifics of the methodology can be found in that publication.

Participants

Thirty-two right-handed individuals participated in the experiment in exchange for course credit or cash. One individual was dropped due to technical issues with importing the data, resulting in a total of 31 participants in the analysis. All participants reported normal or corrected vision and had no history of any neurological or psychiatric disorder. Mean age was 20 years (range 18–28 years), and 16 of the participants were female. The study was approved by the local ethics committee, and all participants provided written informed consent and were debriefed following participation.

Design and Procedure

Participants read 282 sentences that varied in contextual constraint (half being "strong constraint," with expected word cloze probability >0.67 and the other half being "weak constraint," with cloze probability of the most expected word <0.42) and that ended with either the most expected or an unexpected (cloze ≈ 0.03), but plausible word. Stimuli were counter-balanced into two lists, such that half of the sentences completed by an expected ending in one list were completed by an unexpected ending in the second list, and vice versa. Thus, there were four basic conditions of sentence final words: strong constraint expected (SCE), strong constraint unexpected (SCU), weak constraint expected (WCE), and weak constraint unexpected (WCU), with ~70 sentences in each condition. Sentence frames were matched in length, and lexical properties (word length, word frequency) of sentence endings were matched.

Words prior to the final word of the sentence, or pre-final words, were primarily made up of determiners or prepositions ("the," "a," "his," etc.; 65% of pre-final stimuli). Pre-Final words did not reliably differ in length across sentence constraint ($P = 0.06$) but did differ in log frequency ($P < 0.01$). Additional linear mixed effect model analyses were run to include lexical confounds as predictors for experimental effects of interest. Mixed effect models were conducted in R, using the lme4 package (Bates et al. 2015), and statistical significance of fixed effects were estimated with *t*-tests using the Satterthwaite method in the lmerTest package (Kuznetsova et al. 2017).

Association strength between words within sentences and sentence-ending words was measured using the Edinburgh Associative Thesaurus. For each sentence in each condition, the forward and backward association was calculated between each word in the sentence and the sentence final word, and the number of instances of associations greater than 0.2 were counted. Overall, there were few associations between sentence words and sentence-ending words: 4% of SCE sentences contained at least one association greater than 0.2, and for all other conditions (SCU, WCE, WCU), only 1% of sentences contained at least one association. The mean association strength between sentence final words and all words in the sentence was less than 0.005 for each of the four conditions. Thus, observed effects are unlikely to arise simply due to word level associations (but, nevertheless, a control analysis was also performed to further rule out this possibility as a basis for the critical effects).

Participants viewed the sentences on a 21″ CRT monitor in an electrically shielded booth. Each sentence was presented word-by-word in the center of the screen, with each word appearing for 200 ms with an interstimulus interval of 300 ms. Sentences were separated by a 3-s pause. Participants were instructed to attend and read the sentences for comprehension and told that they would be asked questions about what they had read at the end of the experiment.

EEG Recording and Processing

EEG data were recorded from 26 tin electrodes embedded into a flexible elastic cap distributed over the scalp in an equidistant arrangement (see Supplementary Fig. 1). Additional electrodes included one on each mastoid, one on each outer canthus of the eye (for monitoring eye movements), and one below the lower eyelid of the left eye (for monitoring blinks). Electrode impedances were kept below 5 kΩ. Signals were amplified by a Grass amplifier with a bandpass filter of 0.01–100 Hz and a sampling rate of 250 Hz. During recording, the left mastoid electrode was used as a reference; offline, the data were rereferenced to the average of the left and right mastoid electrodes.

Raw EEG time series were filtered with a 0.2–60 Hz digital Butterworth bandpass filter with a 12 dB/oct roll-off. Filter parameters were chosen a priori to remove high-frequency noise and large drifts but retain some higher frequencies that could potentially contribute meaningful variance to the RSA; however, a second analysis with a 0.2–30 Hz filter produced nearly identical results. Note that the high-pass filter was implemented to reduce noise from low-frequency drifts but was not so high as to induce confounds in the similarity analysis. Previous work has demonstrated that filter settings in this range do not produce distortions in electrophysiological measurements over these time scales (Tanner et al. 2015; Wang, Kuperberg, et al. 2018b). To correct ocular artifacts, the data were decomposed into independent components with the AMICA algorithm (Palmer et al. 2012). Component time-courses that correlated with a bipolar vertical electrooculogram channel at Pearson $r > 0.6$ were removed, and the data were reconstructed with the remaining components. The corrected data were then submitted to a sliding window artifact scan to identify extreme amplitude excursions (>90 μV). Any trial in which either the pre-final word or final word was marked as an artifact was excluded from analysis. RSA results could potentially be influenced by differences in trial numbers (Dimsdale-Zucker and Ranganath 2018), and so for, each participant the number of trials in each condition was equated by randomly dropping trials from conditions with more trials than the condition with the minimum number. This resulted in an average of 64 trials in each condition for each participant.

The remaining trials were corrected with a z-scoring baseline correction method (Ciuparu and Mureşan 2016), in which pre-trial baseline periods are fused together and used to z-score the trial data. This method reduces potential biases of single-trial normalization techniques (Grandchamp and Delorme 2011) and allows for scaling of neural measurements, which is important for multivariate analyses. Separate baselines were created for strong constraint and weak constraint sentences to reduce any contamination from sentential constraint (as strong constraint

baselines could differ somewhat from weak constraint baselines). Additionally, both the pre-final word and final word data were corrected with pre-final word baseline data, so as not to introduce any bias by using pre-final word data both as a baseline and in the similarity analysis.

Spatial RSA

Spatial RSA is focused on the similarity of neural activity across the scalp at each timepoint of two time series. It is thus able to reveal if some aspect(s) of the timecourse of neural activity elicited by a pre-final word anticipates similar processing when the predicted word is actually encountered. Of course, it is possible that the preactivation of information would not align in time with when that information would normally be evoked by a word input (that possibility is further assessed using a time-generalization analysis, described next). However, given that word processing follows a characteristic functional and neural time-course as revealed by ERP componentry, it is not unlikely that, for example, semantic preactivation of the final word could arise during semantic processing of the pre-final word (which, as revealed by studies of the N400 component, occurs in a very stable time window around 400 ms; for a review, see Federmeier et al. 2016), and thus show the kind of temporal relationship assessed by spatial RSA.

For each trial of data from each participant, the vector of amplitudes from each of the 26 scalp channels of the pre-final word data was correlated (Pearson's *r*) with the vector of channel data of the final word data at each and every timepoint from 1 to 500 ms postword onset. This resulted in a time-series of correlations between pre-final and final word activity for each trial. For bin-based visualizations and analyses, these time-series were averaged across trials within each of the four conditions (SCE, SCU, WCE, and WCU) for each participant, and grand averages were created by averaging across participants. Additionally, a grand average across all items was created, which was used to identify timepoints of interest for analysis so as not to bias our decision by viewing the condition data (see Supplementary Fig. 2; Luck and Gaspelin 2017). While this method is subjective, it is unbiased in terms of condition and allows for greater statistical power than the more conservative mass univariate approaches. Peak similarity in the grand average was observed at 185 ms; statistical analyses were then conducted in a 50-ms window around that peak, using repeated-measures ANOVAs for testing factors of expectancy and constraint.

Time Generalization Analysis

Time generalization follows the same steps as the spatial RSA, but the channel activity at each timepoint of the pre-final word is correlated with the channel activity of every timepoint of the final word, producing a matrix of correlations with the matching timepoints on the diagonal. This analysis thus tests for additional correlations between pre-final and final word activity that are not precisely aligned in time. Here, only the first 300 ms of the pre-final word activity was correlated with 1–500 ms of the final word activity. The later timepoints of the pre-final word are close in proximity to the early timepoints of the final word, and thus, the overall similarity is greatly increased; however, this is likely not due to prediction, but simply due to autocorrelation of the time-series, and the large correlation values produced could influence the results of statistical analyses. Thus, the time-range of the pre-final word was limited to 1–300 ms to avoid this issue. See Supplementary Fig. 3 for more details.

Time-generalization matrices were created for each trial and averaged across trials for each of the four conditions for each participant. The resultant average matrices were submitted to cluster-based permutation analyses (Maris and Oostenveld 2007) to test for significant differences in similarity between two conditions. Here, *t*-tests were performed at each timepoint testing for differences between conditions. Clusters were identified in the time × time matrix by grouping adjacent timepoints where the *t*-test was significant ($P < 0.05$), and the magnitude of each observed cluster was determined by summing the *t*-values within the cluster. A surrogate distribution was then created by shuffling the condition labels, performing *t*-tests at each timepoint, identifying significant clusters, and recording the largest cluster statistic. The largest statistic was recorded as both a positive and negative value in order to perform a two-sided test (the null hypothesis distribution was assumed to be symmetric). This shuffling procedure was repeated 1000 times, and the observed clusters of the actual data were then compared with the surrogate distribution of cluster statistics to test for significance. The observed clusters were considered significant if 97.5% of the surrogate cluster values were smaller than the observed cluster value, or if 97.5% of the surrogate cluster values were larger than the observed cluster value. Multiple permutation tests were performed to examine differences between the four conditions.

Temporal RSA

Temporal RSA is focused on the similarity of two neural time series at each channel across the scalp, allowing for the visualization of the topography of the similarity. For each trial of data from each participant, the vector of amplitudes from the 75-ms time window (150–225 ms) following the pre-final word was correlated with the 150–225 ms time series following the final word at each of the 26 scalp channels. Note that the window used for the temporal RSA was slightly larger than that used for the spatial RSA. The window was widened to include more points in the temporal RSA correlation for a more stable estimate. This analysis resulted in a scalp map of correlations between pre-final and final word activity for each trial. As with the spatial RSA, these scalp maps were then averaged across trials within each of the four conditions for each participant and averaged across participants to create a grand average scalp map.

This method was additionally extended to a sliding window approach to explore the results from the time generalization analysis. For the correlation approach to be possible, the correlated time series must be the same length; thus, we used 60-ms windows of time from both the pre-final word and the final word activity. A 60-ms time window (slightly larger than the original window used for spatial RSA) was used to capture the entire extent of the significant clusters observed. The time series of pre-final word activity correlated with 60-ms windows of postfinal word activity in successive 4-ms steps for each trial. The scalp maps at each time step were then averaged across trials within each of the four conditions.

Results

EEG was recorded while participants read sentences that varied in contextual constraint and ended with either an expected or

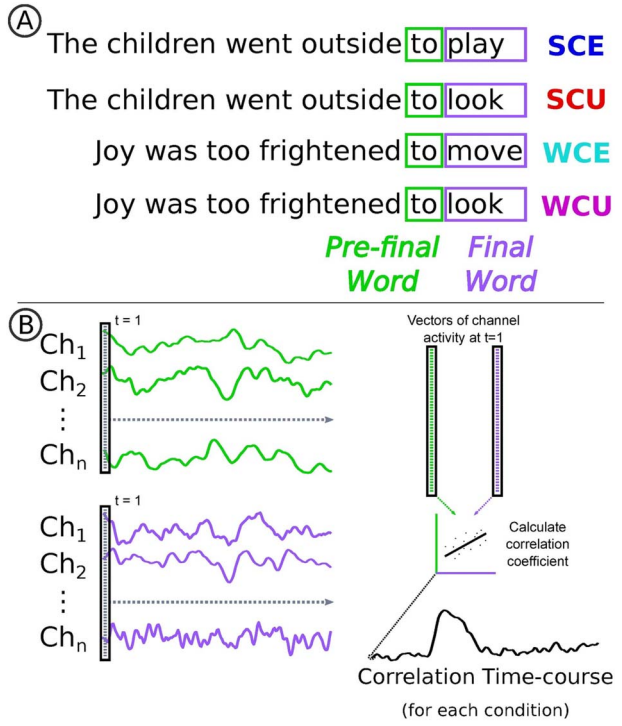

Figure 1. Example of experimental materials and RSA diagram. (*A*) Examples of sentences from each of the four conditions: SCE, SCU, WCE, and WCU. The sentence final words are highlighted in purple, and the pre-final words are highlighted in green. (*B*) In spatial RSA, the vector of EEG channel activity at the first time-point of the pre-final word (shown in green) is correlated with the vector of activity at the first time-point of the final word (shown in purple). EEG activity correlations are calculated at each successive time-point, resulting in a time-course of correlations.

unexpected word (Fig. 1A). We used spatial RSA to compare neural response patterns to pre-final words and final words by correlating the amplitude values across sensors at each timepoint of the two time-series and averaging the resulting similarity time-series within each condition (Fig. 1B).

Spatial RSA

The first analysis correlated neural activity following the pre-final word with neural activity following the sentence final word at matching time-points. Spatial RSA revealed a peak in neural similarity beginning around 100 ms and continuing to about 350 ms following word onset (Fig. 2A). This peak in similarity appeared to vary with both expectancy and sentential constraint. Indeed, expected endings (combining SCE and WCE) showed greater similarity overall compared with unexpected endings (combining SCU and WCU; $t = 2.77$, $P < 0.01$). To assess the impact of expectancy in a more fine-grained manner, an item-level analysis was performed. The similarity values were averaged across subjects for each expected ending with the same cloze probability, and a linear regression was run predicting neural similarity from cloze probability. Cloze probability significantly predicted neural similarity for the expected endings ($t = 2.91$, $r^2 = 0.18$, $P < 0.01$); see Figure 2B. Note that although the majority of the range of cloze probabilities was sampled with these stimuli, there were no items in the middle range of cloze probability (~40–60%). However, relationships between neural activity (e.g., N400s) and cloze probability are usually linear (Wlotko and Federmeier 2012), and thus, it is unlikely that the inclusion of middle-range cloze items would produce a nonlinear relationship.

To assess whether constraint—and hence the likelihood of being able to make a specific prediction, even if it is never realized—also affected responses to the unexpected items, an additional regression was run predicting neural similarity for the unexpected endings (combining SCU and WCU) from graded sentential constraint (i.e., the cloze probability of the most expected ending for that sentence). As seen in Figure 2B, there was a significant linear relationship ($t = 2.24$, $r^2 = 0.12$, $P = 0.03$), suggesting that the prediction signal was graded with sentential constraint.

Controlling for Lexical Characteristics

It is possible that neural similarity might have varied across expectancy or constraint due to lexical characteristics of the pre-final and final words. To test for this possibility, we used linear mixed-effects models to predict pattern similarity with cloze probability and lexical characteristics on a single-trial level. For each trial, the average similarity derived from spatial RSA from 160 to 210 ms was extracted. The first analysis predicted trial-level similarity values from cloze probability for expected sentence endings. Given the hypothesis that lexical similarity could explain the observed effect, we created differences measures by taking the absolute value of the difference between the pre-final word frequency and the final word frequency, as well as word lengths (Orthographic neighborhood size is also a lexical variable known to affect electrophysiological responses to words. However, word length is highly correlated with orthographic neighborhood. To simplify the model, we only included word length and frequency.), and included these measures in the model. The model predicting trial-level neural similarity included random intercepts for participants and items (the final word), and random slopes for cloze probability and lexical properties (word length and frequency) were added.

Significance of fixed effects of the model was assessed with *t*-tests using the Satterthwaite method of approximation for degrees of freedom. The results of this analysis are reported in Table 1. Cloze probability remained a significant predictor of neural similarity, even with lexical characteristics included. Neither word length nor frequency was significant.

The second analysis focused on unexpected sentence endings. Here, the model was similar to the first model described previously, but the cloze probability of the expected ending of the sentence was included instead of the cloze probability of the unexpected ending. Thus, the model was constructed with random intercepts for participants and items (the final word), as well as random slopes for cloze probability of the expected endings and lexical properties (word length and frequency).

The fixed effects results are reported in Table 2. As before, cloze probability remained a significant predictor of similarity even after including lexical variables. However, word frequency was also a significant predictor of similarity. Thus, while it is probable that differences in word frequency contributed to the effect for unexpected endings, there was also a continued influence from sentence-level constraint and, hence, the extent to which a prediction was likely to have been formed from that context. Additional model output (e.g., random effects) is reported in the Supplementary Appendix.

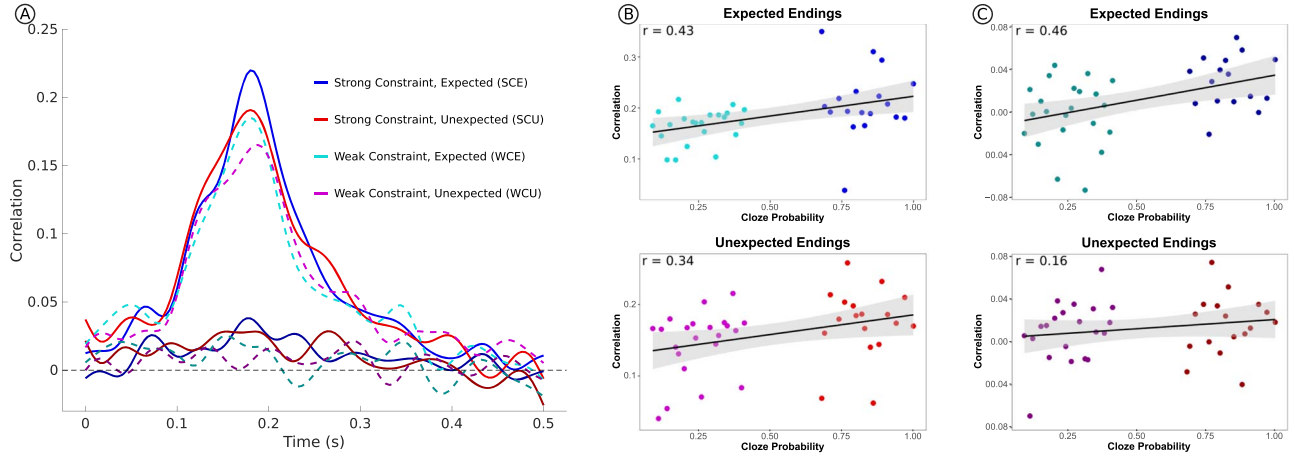

Figure 2. Results of the spatial RSA. (*A*) The similarity time-course is shown for each of the four conditions. The darker lines show the similarity time-course after subtracting the average word response, whereas the lighter colors show the similarity time-course without subtraction. Without subtraction, a peak in overall similarity is observed, which varies with constraint and expectancy. With subtraction, overall similarity does not vary over time, but condition-related differences remain. (*B*) Correlations for data without word subtraction. The correlation between neural similarity and cloze probability for sentences with expected endings is presented on top, and the correlation between neural similarity and sentential constraint (i.e., cloze probability of the expected sentence endings) for sentences with unexpected endings is presented on bottom. (*C*) Correlations between cloze probability/constraint and neural similarity after word subtraction.

Table 1 Fixed effect estimates and tests of significance for mixed effects model predicting trial level similarity values for expected endings derived from spatial RSA

Estimate	Std. Error	Dg. freedom	*t*-value	*P*-value	
Intercept	0.161	0.027	57.83	6.09	<0.01*
Cloze	0.049	0.018	43.09	2.81	<0.01*
Length	0.005	0.003	266.3	1.32	0.19
Freq	−0.009	0.005	55.54	−1.55	0.13

Table 2 Fixed effect estimates and tests of significance for mixed effects model predicting trial level similarity values for unexpected endings derived from spatial RSA

	Estimate	Std. Error	Dg. freedom	*t*-value	*P*-value
Intercept	0.166	0.026	57.78	6.28	<0.01*
E_Cloze	0.031	0.016	3525	1.97	0.04*
Length	0.004	0.003	189.7	1.55	0.12
Freq	−0.013	0.004	242.4	−2.92	<0.01*

Controlling for General Word Processing Activity

The global peak in neural similarity observed in the time window of the effect is likely due to overall similarity in the neural activity elicited when processing visual words. To assess how much such general activity associated with visual word processing might have affected the observed effect pattern, a control analysis was performed to try to subtract out "baseline word activity" prior to running spatial RSA. For each participant, the time-course of neural activity following each word in every sentence, except for first words, pre-final, and final words, was averaged to create an average "word" ERP. This ERP was then subtracted from every pre-final and final word time-course, and spatial RSA was conducted again on this "word-corrected" data.

The resultant RSA time-courses are shown in Figure 2A (the darker lines). Although the global peak in similarity was indeed greatly reduced, the same pattern of condition-related differences is apparent. To test for sensitivity to cloze probability and constraint, trial-level neural similarity was extracted from the word-corrected data, and the linear mixed-effects models previously described were run with the word-corrected trial values. The correlations are presented in Figure 2C. For expected sentence endings, the relationship between cloze probability and neural similarity remained significant ($t = 2.32$, $P = 0.02$), and lexical variables were not significant predictors ($P > 0.05$). Thus, for expected endings, baseline word activity differences did not explain the pattern of observed similarity results. For unexpected sentence endings, neither cloze probability nor word frequency remained as significant predictors after correcting for baseline word activity ($P > 0.05$). The full model outputs are reported in the Supplementary Appendix. This result suggests that general word characteristics may have contributed to the effect of constraint on the similarity pattern for unexpected endings. Since, by design, specific features of unexpected endings would not have been preactivated, it makes sense that the effect of constraint on similarity patterns for these words would be driven by more general word features that are also present in the baseline word activity. However, this baseline word subtraction is likely a coarse correction method, as some signals of prediction could be subtracted out as well.

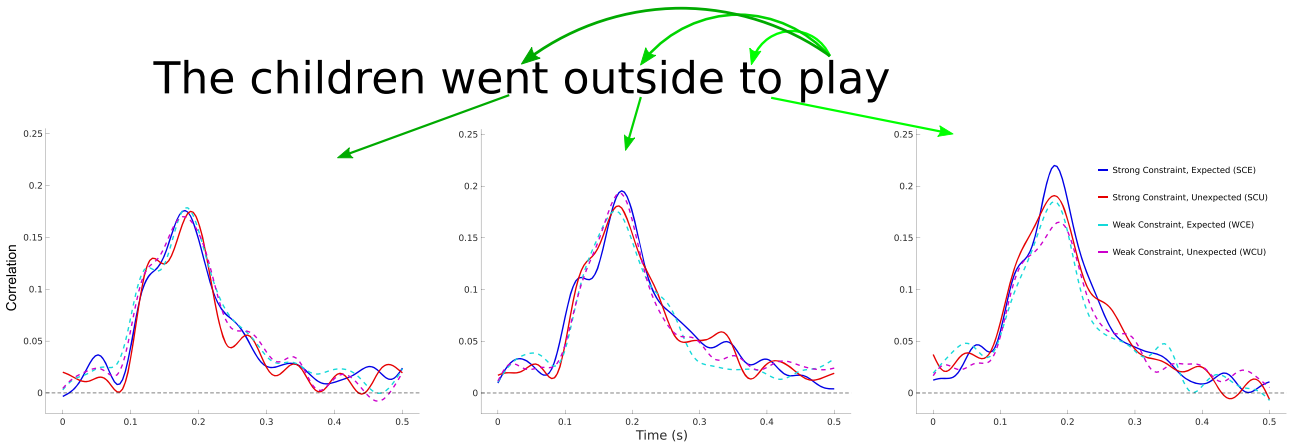

Figure 3. Spatial RSA for sentence final words and pre-final words at three different positions: immediately preceding (right plot), two words back (center plot), and three words back (left plot). Significant differences are observed only for pre-final words immediately preceding sentence final words.

Assessing Which Sentence Position(s) Show Evidence of Preactivation

Predictive preactivations may have been generated or may be detectable prior to the onset of the pre-final word. To assess, we performed the same spatial RSA method comparing neural activity patterns of the final word to the word prior to the pre-final word (2 word positions back). An overall peak in pattern similarity was found in the same time window as the pre-final word analysis, but effects of condition were less apparent (Fig. 3). Indeed, there was no significant difference between expected and unexpected endings at this word position ($t = 0.28$; $P = 0.78$). Additionally, a linear regression predicting pattern similarity from cloze probability was not significant for expected endings ($t = 1.65$, $r^2 = 0.07$, $P = 0.11$) and constraint did not affect pattern similarity for unexpected endings ($t = 0.21$, $r^2 < 0.01$, $P = 0.84$). A spatial RSA comparing final word activity to words 3 word positions back also showed no significant effects of predictability (all P-values > 0.05; Fig. 3). Thus, prediction-related pattern similarity differences were only reliable immediately prior to the onset of the sentence final word.

Generalization Analysis

To probe for similarity that is not temporally aligned and also to assess the fate of predicted representations across time, we employed a time generalization analysis. This differed from the first Spatial RSA analysis, in that, instead of correlating activity only at matching time-points following the pre-final and final words, the activity at each timepoint following the pre-final word was correlated with the activity at each timepoint following the final word. To reduce influences from auto-correlation, the pattern of activity across channels at each timepoint from 0 to 300 ms following the pre-final word was correlated with every timepoint from 0 to 500 ms following the final word. The resulting time × time matrices were analyzed by submitting pair-wise contrasts to cluster-based permutation tests.

Permutation tests revealed four clusters of interest that reached a significant cluster-wise threshold of $P < 0.05$ (Fig. 4A). First, in the contrast of expected words (SCE—WCE), a positive cluster was found (pre-final word time: 150–275 ms; post-final word time: 160–265 ms; $P < 0.01$), with SCE similarity greater than WCE. This cluster likely reflects the same effect

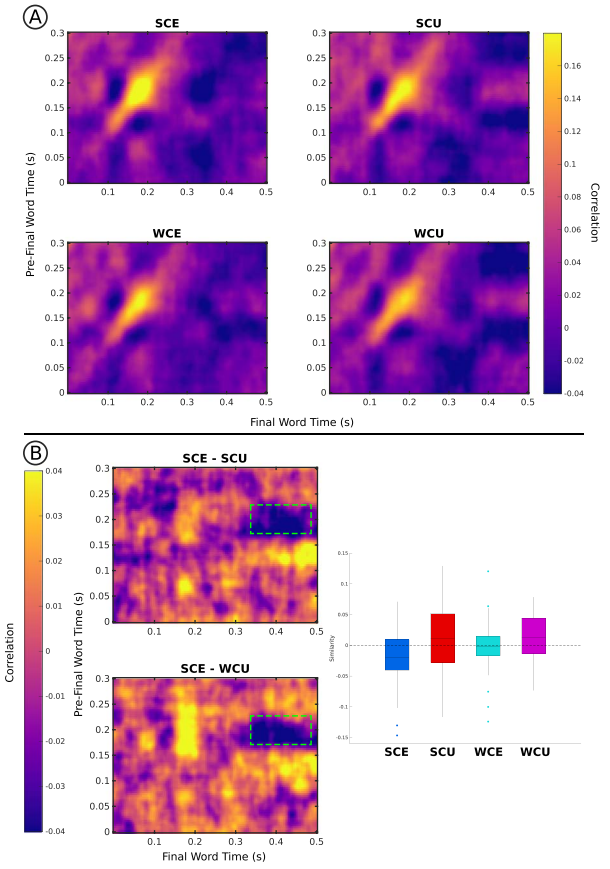

Figure 4. Results of the spatial generalization RSA. (A) Generalization matrices for each condition are plotted. The color intensity represents neural similarity. The strong increase observed in each condition reflects the previously observed spatial RSA peak (Fig. 2A). (B) Differences in generalization matrices for SCE-SCU and SCE-WCU. The green box highlights the significant cluster found for both differences. The bar-plot displays similarity values extracted from this time window, with SCE similarity significantly below zero, demonstrating anticorrelation.

found with the initial spatial RSA analysis, and a similar positive cluster was found in the contrast of SCE and WCU words (pre-final word time: 150–250 ms; postfinal word time:

155–210 ms; $P = 0.01$). These positive clusters demonstrate that early pre-final word and postfinal word similarity is greater for more predictable sentence endings. Comparisons of other pairs did not yield significant positive clusters. It is important to note that the cluster-based permutation tests are more statistically conservative than the analysis focused on the specific peak. Critically, even under a conservative analytical approach, similarity is found to be greater for words with high cloze probability compared with not only unexpected words but also expected words with lower cloze probability.

The permutation tests also identified two similar negative clusters, with SCE words showing reduced similarity compared with both SCU words (pre-final word time: 170–230 ms; post-final word time: 315–475 ms; $P < 0.01$) and WCU words (pre-final word time: 175–230 ms; postfinal word time: 340–500 ms; $P < 0.01$). Note that this effect reflected similarity of pre-final word activity in the time window of the previously reported spatial RSA effect and postfinal word activity in a later time window, roughly the timewindow of the N400 component of the ERP. To examine this effect further, we performed exploratory post-hoc analyses. Follow-up analyses on extracted similarity values (pre-final word time: 175–230 ms; postfinal word time: 340–475) showed that similarity for SCE words was reduced compared with all other conditions (WCE: $t_{(30)} = -2.60$, $P = 0.01$; SCU: $t_{(30)} = -3.62$, $P < 0.01$; WCU: $t_{(30)} = -4.33$, $P < 0.01$), and, in fact, was significantly less than 0 ($t_{(30)} = -2.62$, $P = 0.01$). This pattern is depicted in Figure 4B.

Controlling for Univariate Effects

An alternative explanation of these results is that the similarity difference results reflect a confound of univariate activation magnitude, which has been shown to affect pattern similarity results in fMRI studies (Coutanche 2013). The observed early clusters could reflect differences in early visual ERP component amplitudes (e.g., N1/P2), and the later clusters were in the time window of the N400 following the final word, which does show amplitude differences in a similar pattern to the reported similarity pattern. We performed an additional analysis to test for this possibility. RSA was used to compare activity elicited by sentence final words with activity elicited by pre-final sentence words from different sentences that were the same as the pre-final word of the same sentence. For instance, the sentence "Father carved the turkey with *a* knife" has the same pre-final word as "His touch was light as *a* feather." Here, we measured the similarity of the activity from the word knife in the first sentence to activity from the word *a* in the second sentence (a between-sentence comparison). This was done for all pre-final words that matched the pre-final word in the sentence, and the resultant correlation time-courses across trials were averaged. This allowed us to compare similarity when the pre-final word and final word were exactly the same, but only the sentence differed. This means only the activity at the preword differed, as the univariate effect of the N400 at the final word was the same as in the original analysis.

The results of this analysis are shown in Figure 5. Similarity was greatly reduced for the between-sentence comparison compared with the within-sentence comparison, and similarity between conditions did not significantly differ (all $P > 0.05$). Additionally, similarity was greater in the within-sentence comparison than in the between-sentence comparison for all four conditions of interest (all $P < 0.05$). Finally, the time generalization analysis for the between-sentence comparison revealed no

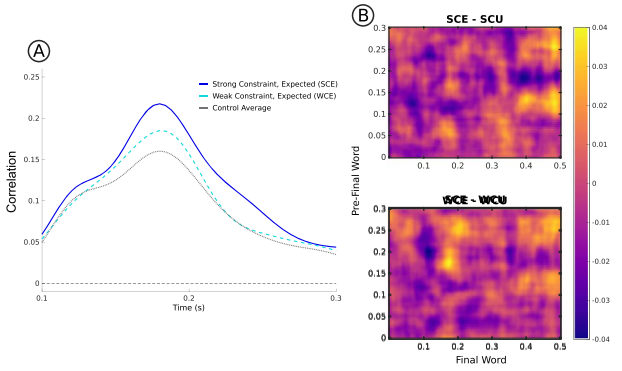

Figure 5. Results of the between-sentence similarity analysis. (*A*) The spatial similarity time-courses, zoomed in on 100–300 ms. The SCE and WCE time-courses from the initial analysis are plotted for comparison. The average across conditions for the control between-sentence analysis is plotted in gray. All conditions had lower similarity than the WCE similarity, and conditions did not differ. (*B*) Time generalization results for the between-sentence control analysis. The plots show differences between conditions (SCE-SCU for the top plot, SCE-WCU for the bottom plot). The later negative cluster found in the within-sentence generalization analysis is not observed.

significant clusters; i.e., the negative cluster was not observed. Thus, the results are unlikely to be driven by univariate confounds, as the amplitude of the ERPs to the final word did not differ in the between-sentence analysis and likely did not largely differ for the pre-final words, but similarity between the two was nonetheless greatly reduced. This control analysis also attests that the observed condition effects cannot be explained by any form of low-level orthographic or lexical similarity between the pre-final and final words, as the same lexical pairing does not yield those results if the pre-final words are simply taken from a different sentence context.

Effect Topographies

To characterize the topography of the early spatial RSA effect, we used temporal RSA. Recall that in the spatial RSA, signals across space were correlated at different timepoints. Here, instead, signals across time were correlated at different spatial locations: This analysis correlated the entire time series of pre-final word and final word activity in the selected window of analysis at each sensor on the scalp. Note that because spatial RSA relates spatial patterns across time, whereas temporal RSA relates temporal patterns across space, the statistical pattern of these results may differ. Temporal RSA showed that similarity was greatest over occipital channels across all conditions (Fig. 6).

The topographies for each of the four conditions did not significantly differ from one another. However, this analysis is focused on the similarity in time at each channel, not the similarity across channels at each time-point. Thus, the resultant topography plots highlight the channels where the pre-final and final word time-series were the most similar. It is thus not surprising that the four conditions would not differ in this analysis, as they are likely to all reflect the same process, which varies in degree with prediction strength and level of match between the prediction and the input.

We performed a similar analysis to characterize the late negative cluster found in the time generalization analysis. We implemented a sliding window approach, in which the time series of pre-final word activity from 170 to 230 ms were correlated with final word activity from 340 to 480 ms across

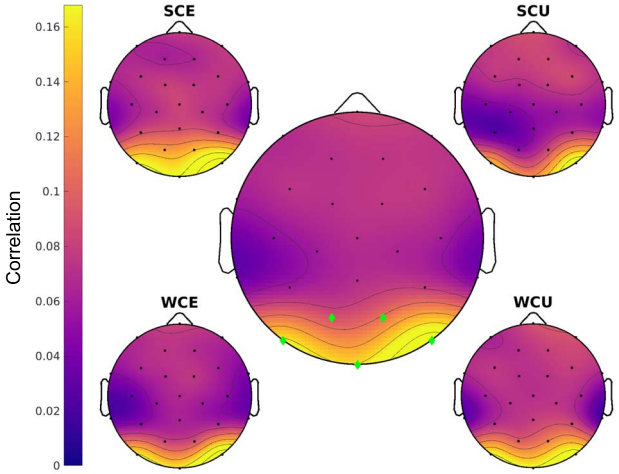

Figure 6. Temporal RSA results for the spatial RSA peak. The central topography plot shows the average across conditions, with the channels with the largest similarity values as green diamonds. The topography for each condition is also plotted. A strong occipital topography is observed.

successive 60-ms windows (Fig. 7). Note that these time windows were designated by identifying the minimum and maximum time values of the significant clusters reported previously. An initial dissimilarity was observed over occipital channels, which was more pronounced and sustained for SCE words. This was followed by an increase in similarity over central channels for unexpected words, but not for expected words. This time-course corroborates the results from the cluster analyses; namely, strongly expected words show less similarity later in time compared with less expected words.

Discussion

Numerous studies have investigated the neural consequences of predictability during language comprehension, but the specific mechanisms, timing, and extent of anticipatory preactivation have remained elusive. Here, we used RSA to compare the patterns of neural activity prior to a sentence-final word to the activity following a sentence-final word, allowing us to examine the timing of generating predictions, as well as their specificity. Our results provide persuasive evidence that people predict upcoming information probabilistically and that those preactivations affect word processing quite rapidly. Additionally, predictions appear to be generated or allocated at specific times—that is, close in time to the upcoming final word. Finally, we provide novel evidence that activity patterns representing preactivations are anticorrelated with later activity following confirmation of predictions. Altogether, these results elucidate the neural mechanisms of prediction during the comprehension of language and potentially provide insight into general mechanisms of prediction in the brain.

Spatial RSA revealed an increase in neural similarity between pre-final word activity and final word activity that extended from ~100–300 ms following pre-final word onset. This similarity was significantly reduced for unexpected but semantically plausible final words compared with expected final words, suggesting that the preactivated features were at some level specific to the expected word. Moreover, this similarity was graded with the cloze probability of the sentence final word, such that similarity decreased as the word became less predictable, showing that predictions are graded. The graded relationship manifested over and above the general peak in similarity and was still observed in a control analysis that subtracted out baseline "word" neural activity. The relationship between neural similarity and cloze probability may reflect a graded degree of effort, in which neural resources are allocated toward anticipatory processing and the level of resource allocation is dependent on predictability and, thus, a function of constraint. Alternatively, this pattern may reflect a graded degree of success, in which the probability of a match between the predicted and actual outcome is greater with higher levels of sentential constraint. In either case, we provide the first results that, when reading language, features of specific upcoming words are rapidly preactivated prior to their onset, and the magnitude of this anticipatory processing is graded with predictability.

Examining pattern similarity of the sentence final word and words prior to the pre-final word revealed that a significant relationship between similarity and predictability was present only for the pre-final word. In other words, evidence of anticipatory preactivation was found only immediately prior to the word being predicted. This finding could reflect that predictions were generated rapidly following the onset of the pre-final word, and not before. Alternatively, the observed signal may not reflect the time at which information became available to the system, but, instead, the time at which the production system allocated

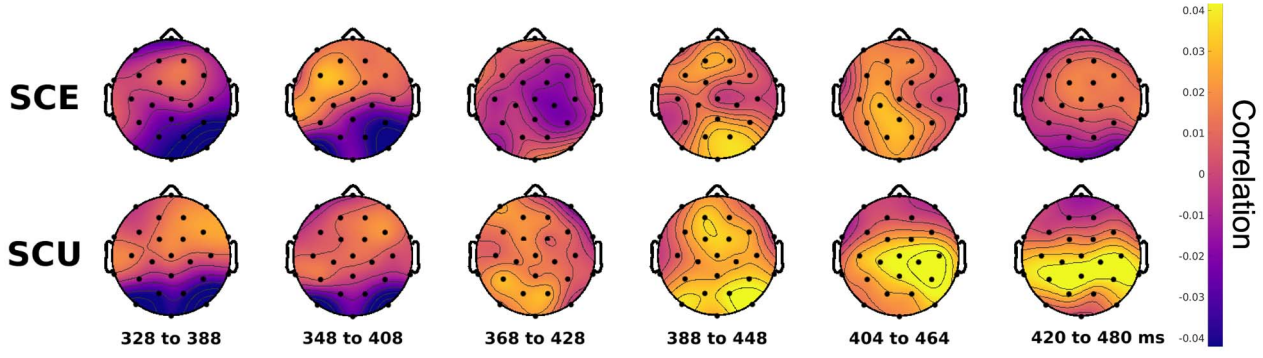

Figure 7. Sliding window temporal RSA results for the late cluster. Similarity topographies are shown at different time windows, where the activity from the final word in the displayed time window is correlated with pre-final word activity from 170 to 230 ms. SCE shows broad dissimilarity that stays near zero over time. SCU shows early dissimilarity, followed by positive similarity over posterior channels.

resources toward explicitly forming a prediction of a particular type (Dell and Chang 2014). In a recent study, neural preactivation of a series of expected simple visual stimuli occurred in visual cortex only after the first stimulus in the train was presented (Ekman et al. 2017); thus, there is precedent that the pre-final word could serve as a cue for preactivation of visual features of the upcoming final word. In the current study, a sentence like "The bad boy was sent to his room" may have allowed some level of anticipation of the final word "room" even at the time of the word "sent" based on the semantics of the sentence. However, as previously described, the contents of prediction are multifaceted in nature, and these different features may be generated at different times. Sentence level semantic predictions may manifest in different neural signals than those observed here and may influence the predictions generated at other levels (e.g., orthographic or syntactic). We do not claim that the preactivation signal reported here reflects the preactivation of all linguistic information; there are likely other signals left to be identified relating to other levels of prediction. What our data reveal is that some aspects of prediction—possibly, as discussed next, prediction of specific word forms—seem to be specifically timed, perhaps cued by, for example, the preceding (usually function) word suggesting the imminent arrival of the target noun. This process could be epiphenomenal—that is, stimuli automatically lead to the preactivation of features of associated stimuli that may follow. Alternatively, such a mechanism could be tailored by the nervous system to be beneficial for efficient behavior during language processing; for example, cued preactivation could guide eye movement behavior during reading to reduce reading times and/or skip over easily predicted information (Ehrlich and Rayner 1981).

Temporal RSA revealed an occipital topography in the time window of the preactivation. Given EEG's limited spatial resolution, it is difficult to pinpoint the exact networks that were involved in the observed preactivation. However, recent related work may provide insight into the neural systems involved in anticipatory processing. One proposed mechanism is that memory systems in the brain, such as the hippocampus, coordinate sensory preactivation of upcoming information through a pattern completion process (Hindy et al. 2016; Kok and Turk-Browne 2018). This proposal is in line with MEG results from a picture-word matching paradigm that demonstrated the prediction of visual word form features (Dikker and Pylkkänen 2013). In that study, MEG source localization of activity prior to predicted target words revealed left medial temporal and occipital sources, with temporal activity slightly preceding occipital activity. Thus, temporal structures could preactivate lexical information, leading to preactivation of form features in sensory cortex. Our combined results—the timing of the similarity effect, as well as the strongly occipital topography—are consistent with this account, and suggest that the increased similarity may have reflected overlap of preactivated and observed lower level orthographic lexical features.

Consistent with the idea that the observed RSA signal might reflect the prediction of word orthography, we observed that, although the pre-final and final word similarity was reduced for unexpected sentence completions, it was not abolished, and it varied with constraint. This differs from the pattern of sem antic-based facilitation seen on the N400, wherein unexpected sentence completions elicit large N400s that do not differ by constraint (Federmeier et al. 2007). Although, by design, there was likely little to no semantic overlap between expected and unexpected completions, orthographic space is more constrained, such that even unexpected words are likely to sometimes carry expected low-level features (shared length, a shared letter, etc.). The observed RSA pattern suggests that, although the unexpected endings were not predicted, there was, in some cases, some level of—likely orthographic—featural overlap. Note that when average "visual word processing activity" was subtracted prior to spatial RSA, the relationship between sentential constraint and neural similarity for the unexpected endings was no longer significant, suggesting that overall word-related or visual processing activity could contribute to this effect. In particular, the average word subtraction method may have removed some of the signal related to general orthographic features that could be preactivated, especially in more constraining contexts. This hypothesis could be tested further by examining neural similarity to unexpected items that are specifically designed to be orthographically similar but semantically dissimilar to expected words, or similarity to unexpected items that remove orthographic features (e.g., pictures or Gabor patches).

The constraint-sensitivity of the similarity response for unexpected words suggests that the prediction signal itself is variable in strength and/or fidelity, based on the predictability of the upcoming final word, an idea consistent with a probabilistic prediction account (Levy 2008; Kuperberg and Jaeger 2016). More weakly constraining sentences, by their nature, permit a wider range of completions, both at the semantic and orthographic level. Thus, if prediction were ubiquitous, or if equal strength of prediction could be allocated to every potential completion, then the possibility of a match, at any level of analysis, would tend to be higher under weak constraint; yet, similarity was reduced for expected as well as unexpected items in weakly compared with more strongly constraining contexts. One possibility is that predictions are generated to the same degree on every trial but the preactivation signal is not distributed probabilistically, such that the brain selects one (or a very small set) of likely items from the set of possibilities and then preactivates features only of those selected items (also referred to as "preupdating"; Kuperberg and Jaeger 2016). Across trials, then, the selection is more likely to match the observed stimulus when constraint is higher, leading to greater neural similarity. However, this account (alone) cannot explain why the similarity for unexpected items also tended to vary with constraint, since these endings were essentially unpredictable (near 0% cloze probability) and thus should always result in a mismatch.

We argue that rather than predicting to the same degree and/or level every time, the brain may essentially utilize a generative model to rationally and optimally allocate resources toward anticipatory processing based on the available contextual evidence and projected costs of preactivation. This resource allocation would lead to a more defined or "word-like" neural representation in strongly constraining contexts, whereas if the context does not strongly constrain toward a particular outcome, the representation would be less precise. A similar mechanism has been proposed to explain generalization and adaptation during speech perception (Kleinschmidt and Jaeger 2015). If we recognize a familiar speaker, we may generate more specific predictions or activate more features of upcoming utterances from this speaker compared with an unfamiliar one; essentially, the distribution of possibilities narrows. Here, individuals may have generated more specific predictions—perhaps including orthographic features—when the sentential context was highly biased toward a particular outcome. Since the preactivation

occurred prior to the final word, the representational similarity was graded by cloze probability even for unexpected sentence endings, as the representation of the final word was weaker when predictability was lower, leading to less of a possibility of even incidental featural overlap. To our knowledge, this result is one of the first to demonstrate rapid probabilistic preactivation. Further work using this method may shed light on the debate between serial and parallel predictions; namely, whether similarity for other possible sentence endings varies with their completion probability, or a graded response is only found for a single word representation. Additionally, future experiments examining context-specific (Nieuwland and Berkum 2006) and speaker-specific (Ryskin et al. 2019) predictions may benefit from application of RSA to better understand how these factors affect neural preactivation.

To further probe the fate of predicted representations, we conducted a time generalization analysis, in which the spatial pattern of neural activity at each time point following the pre-final word was correlated with the pattern at each timepoint following the final word. This revealed a surprising finding: Early pre-final word activity was less similar to later final word activity for strongly constrained expected endings compared with unexpected endings, and in fact had negative similarity or anticorrelation. Anticorrelation has only rarely been observed in other studies utilizing RSA but has been found in episodic memory studies; namely, hippocampal firing patterns representing events occurring in different contexts are anticorrelated (McKenzie et al. 2014). Similarly, hippocampal representations of overlapping spatial routes become anticorrelated or demonstrate "repulsion" or "differentiation," over time (Chanales et al. 2017). Here, final word activity became differentiated from pre-final word activity after the early increase in similarity. The fact that this occurred most in cases wherein strong predictions were formed and confirmed suggests that the representation of the pre-final word was not anticorrelated, but in fact, the features of the final word that were preactivated were anticorrelated. Such a result is in line with other findings that, downstream, individuals have impoverished representations and impaired memory for predicted words (Rommers and Federmeier 2018b; Hubbard et al. 2019). This provides a view of anticipatory processing during language comprehension in which prediction allows for rapid verification of incoming words, leading to more efficient processing in the moment, but, once the predicted information is verified, the brain shifts processing away, essentially "leaving behind" the predicted information.

Another insight from the time generalization analysis is that the similarity increase during the pre-final period was not sustained over the delay prior to the onset of the word, as might have been predicted by analogy to some accounts of the maintenance of information in working memory (Fuster and Alexander 1971). Indeed, maintained neural firing to sustain predictions would seem a highly inefficient and metabolically costly strategy for the brain and thus not likely to be the mechanism of preactivation. Even sustained working memory signals are observed after averaging many trials; in actuality, spiking during delay periods on single trials is sparse and varies in time (Shafi et al. 2007). More recent evidence suggests that activity at specific frequencies coordinates in bursts to produce rapid synaptic weight changes, which efficiently code information (Miller et al. 2018). A similar mechanism seems likely to be utilized for rapid predictive coding and would explain how a lack of maintained delay-related activity could still produce preactivation of upcoming information.

Our results contribute to a growing literature on prediction during language comprehension, not only in the domain of cognitive and neuroscientific experiments but also in the field of natural language processing and neural network models designed to predict upcoming words given a particular input (Bengio et al. 2003; Radford et al. 2019). The outputs and contextual representations of these models have been used to predict or compare the fMRI data (Jain and Huth 2018), as well as EEG data (Hashemzadeh et al. 2020), allowing for novel tools to probe predictive processing in the brain. Conversely, recent work has used insights from neural data to fine-tune these models to better predict human language output (Schwartz et al. 2019), as these models are trained on language corpora, which do not show the same properties as human-generated predictions (Smith and Levy 2011; Eisape et al. 2020). Future work in this area could use network models to further probe the preactivation observed in the current study, or this preactivation signal could be used to better fine-tune network models for more accurate and human-like predictions.

As a set, our findings demonstrate that the brain rapidly preactivates specific features of upcoming words during language comprehension and does so in a timely manner to allow for efficient processing. These results not only conclusively demonstrate that individuals utilize prediction to preactivate information during comprehension but also shed light on one of the fundamental functions of the brain. Environmental cues may lead to the generation of predictions of associated information or stimuli, and once the preactivated stimulus is encountered and confirmed, the brain may rapidly shift processing away to focus resources on other objectives. While continued research is necessary to better understand the precise mechanisms of generating predictions in the brain, our work moves toward providing answers to the "what" and "when" of prediction during language comprehension (Huettig 2015).

Supplementary Material

Supplementary material is available at *Cerebral Cortex* online.

Funding

National Institute on Aging (R01AG026308 to K.D.F.); Beckman Institute Postdoctoral Fellowship to R.J.H.

Authors' Contributions

R.J.H. designed and performed the analyses. R.J.H. and K.D.F. wrote the paper together.

Data Availability

For protection of participants' privacy, the data are available upon request to the authors. Please email the corresponding author for more information.

Notes

Conflict of Interest: The authors declare no competing financial interests.

References

Altmann GTM, Kamide Y. 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*. 73(3):247–264. doi: 10.1016/S0010-0277(99)00059-1.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme 4. *J Stat Softw*. 67(1):1–48. doi: 10.18637/jss.v067.i01.

Bengio Y, Ducharme R, Vincent P, Jauvin C. 2003. A neural probabilistic language model. *J Mach Learn Res*. 3(Feb):1137–1155. doi: 10.1162/153244303322533223.

Chanales AJH, Oza A, Favila SE, Kuhl BA. 2017. Overlap among spatial memories triggers repulsion of hippocampal representations. *Curr Biol*. 27(15):2307, e5–2317. doi: 10.1016/j.cub.2017.06.057.

Cichy RM, Pantazis D. 2017. Multivariate pattern analysis of MEG and EEG: a comparison of representational structure in time and space. *Neuro Image*. 158:441–454. doi: 10.1016/j.neuroimage.2017.07.023.

Cichy RM, Ramirez FM, Pantazis D. 2015. Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *Neuro Image*. 121:193–204. doi: 10.1016/j.neuroimage.2015.07.011.

Ciuparu A, Mureşan RC. 2016. Sources of bias in single-trial normalization procedures. *Eur J Neurosci*. 43(7):861–869. doi: 10.1111/ejn.13179.

Coutanche MN. 2013. Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? *Cogn Affect Behav Neurosci*. 13(3):667–673. doi: 10.3758/s13415-013-0186-2.

Dell GS, Chang F. 2014. The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philos Trans R Soc Lond B Biol Sci*. 369(1634):20120394. doi: 10.1098/rstb.2012.0394.

DeLong KA, Quante L, Kutas M. 2014. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*. 61:150–162. doi: 10.1016/j.neuropsychologia.2014.06.016.

DeLong KA, Urbach TP, Kutas M. 2005. Probabilistic word preactivation during language comprehension inferred from electrical brain activity. *Nat Neurosci*. 8(8):1117–1121. doi: 10.1038/nn1504.

Dikker S, Pylkkänen L. 2011. Before the N400: effects of lexical–semantic violations in visual cortex. *Brain Lang*. 118(1–2):23–28. doi: 10.1016/j.bandl.2011.02.006.

Dikker S, Pylkkänen L. 2013. Predicting language: MEG evidence for lexical preactivation. *Brain Lang*. 127(1):55–64. doi: 10.1016/j.bandl.2012.08.004.

Dikker S, Rabagliati H, Farmer TA, Pylkkänen L. 2010. Early occipital sensitivity to syntactic category is based on form typicality. *Psychol Sci*. 21(5):629–634. doi: 10.1177/0956797610367751.

Dimsdale-Zucker, H. R., & Ranganath, C. 2018. Representational similarity analyses: a practical guide for functional MRI applications. In: *Handbook of Behavioral Neuroscience*, Vol. 28. London, UK: Elsevier, pp. 509–525.

Ehrlich SF, Rayner K. 1981. Contextual effects on word perception and eye movements during reading. *J Verbal Learning Verbal Behav*. 20(6):641–655. doi: 10.1016/S0022-5371(81)90220-6.

Eisape T, Zaslavsky N, Levy R. 2020. Cloze distillation improves psychometric predictive power. In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Amsterdam, the Netherlands, pp. 609–619.

Ekman M, Kok P, de Lange FP. 2017. Time-compressed preplay of anticipated events in human primary visual cortex. *Nat Commun*. 8(1):1–9. doi: 10.1038/ncomms15276.

Federmeier KD. 2007. Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology*. 44(4):491–505. doi: 10.1111/j.1469-8986.2007.00531.x.

Federmeier KD, Kutas M. 1999. A rose by any other name: long-term memory structure and sentence processing. *J Mem Lang*. 41(4):469–495. doi: 10.1006/jmla.1999.2660.

Federmeier KD, Kutas M, Dickson DS. 2016. A common neural progression to meaning in about a third of a second. In Hickok GS and Small SL, editors. *Neurobiology of language*. Holland: Elsevier, pp. 557–568. doi:10.1016/B978-0-12-407794-2.00045-6

Federmeier KD, Wlotko EW, De Ochoa-Dewald E, Kutas M. 2007. Multiple effects of sentential constraint on word processing. *Brain Res*. 1146:75–84. doi:10.1016/j.brainres.2006.06.101

Freunberger D, Roehm D. 2017. The costs of being certain: brain potential evidence for linguistic preactivation in sentence processing: brain-potential evidence for linguistic preactivation. *Psychophysiology*. 54(6):824–832. doi: 10.1111/psyp.12848.

Friston K, Kiebel S. 2009. Predictive coding under the free-energy principle. *Philos Trans R Soc Lond, B, Biol Sci*. 364(1521):1211–1221. doi: 10.1098/rstb.2008.0300.

Fuster JM, Alexander GE. 1971. Neuron activity related to short-term memory. *Science*. 173(3997):652–654. doi: 10.1126/science.173.3997.652.

Grandchamp R, Delorme A. 2011. Single-trial normalization for event-related spectral decomposition reduces sensitivity to noisy trials. *Front Psychol*. 2:236–236. doi: 10.3389/fpsyg.2011.00236.

Hashemzadeh M, Kaufeld G, White M, Martin AE, Fyshe A. 2020. *From Language to Language-ish: How Brain-like is an LSTM's Representation of Nonsensical Language Stimuli?* Stroudsburg, PA: Association for Computational Linguistics.

Heikel E, Sassenhagen J, Fiebach CJ. 2018. Time-generalized multivariate analysis of EEG responses reveals a cascading architecture of semantic mismatch processing. *Brain Lang*. 184:43–53. doi: 10.1016/j.bandl.2018.06.007.

Hess DJ, Foss DJ, Carroll P. 1995. Effects of global and local context on lexical processing during language comprehension. *J Exp Psychol Gen*. 124(1):62–82. doi: 10.1037/0096-3445.124.1.62.

Hindy NC, Ng FY, Turk-Browne NB. 2016. Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat Neurosci*. 19(5):665–667. doi: 10.1038/nn.4284.

Hubbard RJ, Rommers J, Jacobs CL, Federmeier KD. 2019. Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Front Hum Neurosci*. 13. doi: 10.3389/fnhum.2019.00291.

Huettig F. 2015. Four central questions about prediction in language processing. *Brain Res*. 1626:118–135. doi: 10.1016/j.brainres.2015.02.014.

Huettig F, Guerra E. 2019. Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Res*. 1706:196–208. doi: 10.1016/j.brainres.2018.11.013.

Jain S, Huth A. 2018. Incorporating context into language encoding models for fmri. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*. pp. 6629–6638.

Kim A, Lai V. 2012. Rapid interactions between lexical semantic and word form analysis during word recognition in context: evidence from ERPs. *J Cogn Neurosci*. 24(5):1104–1112. doi: 10.1162/jocn_a_00148.

King J-R, Dehaene S. 2014. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn Sci*. 18(4):203–210. doi: 10.1016/j.tics.2014.01.002.

Kleinschmidt DF, Jaeger TF. 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol Rev*. 122(2):148–203. doi: 10.1037/a0038695.

Kok P, Turk-Browne NB. 2018. Associative prediction of visual shape in the hippocampus. *J Neurosci*. 38(31):6888–6899. doi: 10.1523/JNEUROSCI.0163-18.2018.

Kriegeskorte N, Mur M, Bandettini PA. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Front Syst Neurosci*. 2. doi: 10.3389/neuro.06.004.2008.

Kuperberg GR, Jaeger TF. 2016. What do we mean by prediction in language comprehension? *Lang Cogn Neurosci*. 31(1):32–59. doi: 10.1080/23273798.2015.1102299.

Kutas M, DeLong KA, Smith NJ. 2011. A look around at what lies ahead: prediction and predictability in language processing. In: Bar M, editor. *Predictions in the brain: using our past to generate a future*. New York, NY: Oxford University Press, pp. 190–207.

Kuznetsova A, Brockhoff PB, Christensen RH. 2017. Lmer test package: tests in linear mixed effects models. *J Stat Softw*. 82(1):1–26. doi: 10.18637/jss.v082.i13.

Laszlo S, Federmeier KD. 2009. A beautiful day in the neighborhood: an event-related potential study of lexical relationships and prediction in context. *J Mem Lang*. 61(3):326–338. doi: 10.1016/j.jml.2009.06.004.

Lau EF, Holcomb PJ, Kuperberg GR. 2013. Dissociating N400 effects of prediction from association in single-word contexts. *J Cogn Neurosci*. 25(3):484–502. doi: 10.1162/jocn_a_00328.

León-Cabrera P, Rodríguez-Fornells A, Morís J. 2017. Electrophysiological correlates of semantic anticipation during speech comprehension. *Neuropsychologia*. 99:326–334. doi: 10.1016/j.neuropsychologia.2017.02.026.

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

Luck SJ, Gaspelin N. 2017. How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*. 54(1):146–157. doi: 10.1111/psyp.12639.

Maess B, Mamashli F, Obleser J, Helle L, Friederici AD. 2016. Prediction signatures in the brain: semantic pre-activation during language comprehension. *Front Hum Neurosci*. 10. doi: 10.3389/fnhum.2016.00591.

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

McKenzie S, Frank AJ, Kinsky NR, Porter B, Rivière PD, Eichenbaum H. 2014. Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron*. 83(1):202–215. doi: 10.1016/j.neuron.2014.05.019.

Miller EK, Lundqvist M, Bastos AM. 2018. Working memory 2.0. *Neuron*. 100(2):463–475. doi: 10.1016/j.neuron.2018.09.023.

Nieuwland MS, Van Berkum JJ. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *J Cogn Neurosci*. 18(7):1098–1111. doi:10.1162/jocn.2006.18.7.1098

Palmer JA, Kreutz-Delgado K, Makeig S. 2012. *AMICA: an adaptive mixture of independent component analyzers with shared components*. Swartz Center for Computatonal Neursoscience, University of California San Diego, Tech. Rep.

Pickering MJ, Gambi C. 2018. Predicting while comprehending language: a theory and review. *Psychol Bull*. 144(10):1002–1044. doi: 10.1037/bul0000158.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. 2019. Language models are unsupervised multitask learners. *Open AI Blog*. 1(8):9.

Roll M, Söderström P, Frid J, Mannfolk P, Horne M. 2017. Forehearing words: pre-activation of word endings at word onset. *Neurosci Lett*. 658:57–61. doi: 10.1016/j.neulet.2017.08.030.

Rommers J, Federmeier KD. 2018a. Lingering expectations: a pseudo-repetition effect for words previously expected but not presented. *Neuro Image*. 183:263–272. doi: 10.1016/j.neuroimage.2018.08.023.

Rommers J, Federmeier KD. 2018b. Predictability's aftermath: downstream consequences of word predictability as revealed by repetition effects. *Cortex*. 101:16–30. doi: 10.1016/j.cortex.2017.12.018.

Ryskin R, Ng S, Mimnaugh K, Brown-Schmidt S, Federmeier KD. 2019. Talker-specific predictions during language processing. *Lang Cogn Neurosci*. 35(6):797–812. doi: 10.1080/23273798.2019.1630654.

Schwanenflugel PJ, LaCount KL. 1988. Semantic relatedness and the scope of facilitation for upcoming words in sentences. *J Exp Psychol Learn Mem Cogn*. 14(2):344–354. doi: 10.1037/0278-7393.14.2.344.

Schwartz D, Toneva M, Wehbe L. 2019. Inducing brain-relevant bias in natural language processing models. *Adv Neural Inf Process Syst*. 32:14123–14133.

Shafi M, Zhou Y, Quintana J, Chow C, Fuster J, Bodner M. 2007. Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience*. 146(3):1082–1108. doi: 10.1016/j.neuroscience.2006.12.072.

Smith N, Levy R. 2011. Cloze but no cigar: the complex relationship between cloze, corpus, and subjective probabilities in language processing. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33, No. 33. Austin, TX. https://escholarship.org/uc/item/69s3541f

Söderström P, Horne M, Frid J, Roll M. 2016. Pre-activation negativity (PrAN) in brain potentials to unfolding words. *Front Hum Neurosci*. 10. doi: 10.3389/fnhum.2016.00512.

Staub A, Clifton C. 2006. Syntactic prediction in language comprehension: evidence from either ... or. *J Exp Psychol Learn Mem Cogn*. 32(2):425–436. doi: 10.1037/0278-7393.32.2.425.

Szewczyk JM, Schriefers H. 2013. Prediction in language comprehension beyond specific words: an ERP study on sentence comprehension in polish. *J Mem Lang*. 68(4):297–314. doi: 10.1016/j.jml.2012.12.002.

Szewczyk JM, Schriefers H. 2018. The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Lan Cogn Neurosci*. 33(6):665–686. doi: 10.1080/23273798.2017.1401101.

Tanner D, Morgan-Short K, Luck SJ. 2015. How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*. 52(8):997–1009. doi: 10.1111/psyp.12437.

Thornhill DE, Van Petten C. 2012. Lexical versus conceptual anticipation during sentence processing: frontal positivity and N400 ERP components. *Int J Psychophysiol*. 83(3):382–392. doi: 10.1016/j.ijpsycho.2011.12.007.

Van Berkum JJA, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. 2005. Anticipating upcoming words in discourse: evidence from erps and reading times. *J Exp Psychol Learn Mem Cogn*. 31(3):443–467. doi: 10.1037/0278-7393.31.3.443.

Vissers CTWM, Chwilla DJ, Kolk HHJ. 2006. Monitoring in language perception: the effect of misspellings of words in highly constrained sentences. *Brain Res*. 1106(1):150–163. doi: 10.1016/j.brainres.2006.05.012.

Wang L, Hagoort P, Jensen O. 2018a. Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. *J Cogn Neurosci*. 30(3):432–447. doi: 10.1162/jocn_a_01190.

Wang L, Kuperberg G, Jensen O. 2018b. Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *Elife*. 7:e39061. doi: 10.7554/eLife.39061.

Wicha NYY, Moreno EM, Kutas M. 2003. Expecting gender: an event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in spanish. *Cortex*. 39(3):483–508. doi: 10.1016/S0010-9452(08)70260-0.

Wlotko EW, Federmeier KD. 2012. So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *Neuro Image*. 62(1):356–366. doi: 10.1016/j.neuroimage.2012.04.054.

Wlotko EW, Federmeier KD. 2015. Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*. 68:20–32. doi: 10.1016/j.cortex.2015.03.014.