# Dissociating the Pre-activation of Word Meaning and Form During Sentence Comprehension: Evidence From EEG Representational Similarity Analysis

Lin Wang[1,2], Trevor Brothers[1,2], Ole Jensen[3], Gina R. Kuperberg[1,2]

*[1] Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, 02129, USA*

*[2] Department of Psychology, Tufts University, Medford, MA, 02155, USA*

*[3] Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham B15 2TT, U.K.*

Corresponding authors' email addresses:

lwang48@mgh.harvard.edu

# Abstract

During language comprehension, the processing of each incoming word is facilitated in proportion to its predictability. Here, we asked whether anticipated upcoming linguistic information is actually *pre-activated* before new bottom-up input becomes available, and if so, whether this pre-activation is limited to the level of semantic features, or whether extends to representations of individual word-forms (orthography/phonology). We carried out Representational Similarity Analysis on EEG data while participants read highly constraining sentences. Prior to the onset of the expected target words, sentence pairs predicting semantically-related words (*financial* "bank" – "loan") and form-related words (*financial* "bank" – *river* "bank") produced more similar neural patterns than pairs predicting unrelated words ("bank" – "lesson"). This provides direct neural evidence for item-specific semantic and form predictive pre-activation. Moreover, the semantic pre-activation effect preceded the form pre-activation effect, suggesting that top-down pre-activation is propagated from higher to lower levels of the linguistic hierarchy over time.

# Introduction

Prediction plays a vital role in language comprehension, as it allows us to process language more quickly and accurately (Kuperberg & Jaeger, 2016). However, there is ongoing debate regarding the nature of these predictions, and the levels of representation at which they are generated (Kuperberg & Jaeger, 2016). There is a growing consensus that we use prior contexts to pre-activate the *meanings* of upcoming words. For example, given the context *"I went to deposit the check at the...."*, we are likely to predict the semantic features of the expected word, "bank" (e.g. <financial> and <service>), which will facilitate the processing of "bank" if it is subsequently encountered (Federmeier & Kutas, 1999; Kutas & Federmeier, 2011). A far more contentious issue is whether we additionally pre-activate the linguistic *form* of anticipated words. For example, do we additionally pre-activate the orthographic representation, /B-A-N-K/, before this information actually becomes available from the bottom-up input? Addressing this question is crucial, as it would offer valuable insights into the computational principles and mechanisms of predictive language processing, and the nature of predictive processing more generally.

One possibility is that we only predict at the semantic level. For example, during incremental language comprehension, the evolving representation of the prior context may interact or "resonate" with words stored in long-term semantic memory (e.g. Gerrig & McKoon, 1998; Myers & O'Brien, 1998; Van Berkum, 2009), or with higher-level representations of common events or states, which may passively spread activation to associated semantic features of upcoming words (Paczynski & Kuperberg, 2012; Sanford et al., 2011; and see Kuperberg et al., 2011; Lau et al., 2013 for discussion).

An alternative theory, grounded in hierarchical generative frameworks (Kuperberg & Jaeger, 2016, Section 5), such as predictive coding (Mumford, 1992; Rao & Ballard, 1999), proposes that predictions are propagated down from higher to lower levels of the representational hierarchy over time. On this account, the highest-level ongoing interpretation of the prior context actively propagates predictions downward to pre-activate upcoming lexico-semantic information (the semantic features associated with upcoming words), which, in turn, propagate predictions further down to pre-activate specific upcoming word-form representations at the level of orthography/phonology.

Thus far, existing studies have not been able to distinguish between these accounts. Some studies using event-related potentials (ERPs) have reported effects of predictability on early ERP components that are linked to the processing of word-forms (Brothers et al., 2015; Connolly & Phillips, 1994; Federmeier et al., 2005; Groppe et al., 2010; Lau et al., 2013), while others have reported an attenuation of the N400, a measure of lexico-semantic predictability, in response to unexpected words that are orthographically related to anticipated words (DeLong et al., 2019; DeLong et al., 2021; Ito et al., 2016; Laszlo & Federmeier, 2009). ERP effects have also been reported on pre-target words that are incongruous with the form-level features of upcoming expected words, such as "an" *versus* "a" when the context strongly constrains for the expected word, "kite" (DeLong et al., 2005). However, several of these results have proved difficult to replicate (e.g. Nieuwland et al., 2018), and others may not offer unequivocal evidence of word-form pre-activation (see Nieuwland, 2019 for a review). Most importantly, because ERPs index neural activity that is time-locked to the onset of new input, they assess the *effects* of prediction on bottom-up processing, rather than directly isolating predictive activity itself.

To circumvent this issue, some researchers have examined oscillatory activity *before* the onset of predictable *versus* unpredictable words. There have been some reports of effects that localize to lower-level cortical regions, which have been interpreted as reflecting anticipatory activity at the level of orthography (Wang, Hagoort, et al., 2018) or even low-level visual features (Dikker & Pylkkänen, 2013). Others, however, have argued that these effects reflect pre-stimulus differences in attention or arousal, rather than prediction *per se* (Terporten et al., 2019).

In the present study, we took a different approach. We employed Representational Similarity Analysis (RSA) to ask whether the brain generates distinct neural patterns that correspond to the pre-activation of the meaning (e.g. <financial> and <service>) and the form (i.e. orthography and phonology; e.g. /B-A-N-K/ and /bæŋk/) of specific upcoming words before this information became available from the bottom-up input. The basic assumption of RSA is that representationally similar items produce patterns of neural activity that are more similar to each other than representationally distinct items (Kriegeskorte et al., 2008). When used in conjunction with time-sensitive neurophysiological methods, like EEG or MEG, RSA can also tell us *when* this representationally-specific information becomes active (Cichy et al., 2014; Stokes, 2015).

In a recent MEG study (Wang, Kuperberg, et al., 2018), we used RSA to show that, during sentence comprehension, the patterns of neural activity produced before the appearance of upcoming words were more similar when pairs of contexts predicted the same words (e.g. "baby" – "baby") than when they predicted different words (e.g. "baby" – "rose"). This provided evidence that item-specific information associated with upcoming words was pre-activated before the bottom-up input actually appeared. However, it was unclear whether comprehenders only pre-activated the semantic features of upcoming words (e.g. <little>, <cute>, <crying>), or whether predictions were additionally generated at the level of word-form (e.g. /B-A-B-Y/, /ˈbeɪbi/).

To address this question directly, we implemented a novel experimental design that aimed to dissociate the pre-activation of meaning and form. We developed triplets of highly constraining sentence contexts that gave rise to two types of *related* sentence pairs. First, *semantically-related* pairs constrained for upcoming words that shared meaning but not form (e.g. "I went to deposit the check at the… [bank]" and "His college was very expensive, so he had to take out a student… (Vuong & Martin)"). Second, *form-related* pairs constrained for homographs that shared form but not meaning (e.g. "I went to deposit the check at the… [bank]" and "The muddy sides of the river are called the river… [bank]") [1]. Across triplets, *unrelated* sentence pairs constrained for words that shared neither meaning nor form, and therefore served as a control condition (e.g. "I went to deposit the check at the… [bank]" and "After waking up, the student went to his first… [class]").

We measured EEG as participants silently read these sentences for comprehension. At each time point from the onset of the pre-target word until the onset of the predicted target word, we computed the similarity between the spatiotemporal patterns of neural activity produced by each type of sentence pair. If readers pre-activate the *semantics* of upcoming words, then the neural patterns produced by the *semantically-related* pairs should be significantly more similar than the patterns produced by the *unrelated* pairs. If readers additionally pre-activate the *form* of upcoming words, then the patterns produced by the *form-related* pairs should be significantly more similar than those produced by the *unrelated* pairs. Finally, if top-down predictions are propagated from higher to lower levels of the representation over time, then the semantic pre-activation effect should precede the form pre-activation effect.

## Methods

[1] In line with all accounts of predictive language comprehension, and all theories of post-onset homograph processing, we assume that in highly constraining contexts, the appropriate meaning of the homograph is pre-activated. For example, the dominant financial-related meaning of the word "bank" should be pre-activated following a dominant-constraining context, while the subordinate river-related meaning of the word "bank" should be pre-activated following the subordinate-constraining context. In the psycholinguistic literature, there has been some debate about whether, *after* a homograph is encountered from bottom-up input, we additionally access its contextually-inappropriate meaning. Addressing this question is outside the scope of the present study.

## Design and development of stimuli

### Experimental design

We constructed 84 triplets of highly constraining sentences that each predicted a specific upcoming word (Figure 1A). The contexts within each triplet were distinct from one another, and they had no content words in common. Each triplet was comprised of three types of sentence contexts: (1) dominant-constraining contexts that predicted the dominant meaning of a homograph (e.g. the *financial* meaning of "bank"), (2) non-homograph-constraining contexts that predicted a word that was semantically related to the dominant meaning of the homograph (e.g. "loan"), and (3) subordinate-constraining contexts that predicted the same homograph's subordinate meaning (e.g. the *river* meaning of "bank"). This yielded two types of sentence pairs within each triplet: *semantically-related* pairs (n = 84), in which the predicted words were associated in meaning but not in form (e.g. *financial* "bank" – "loan"), and *form-related* pairs (n = 84), in which the predicted homographs overlapped in form but not in meaning (e.g. *river* "bank" – *financial* "bank"). Across different triplets, there were 31,374 *unrelated* sentence pairs, in which the predicted words were always distinct in both form and meaning ("loan" – "lesson"; "bank" – "class", etc.). This large number of *unrelated* pairs served as a control condition.

To construct these stimuli, we carried out a series of cloze norming tests (Taylor, 1953). Native English speakers, recruited through Amazon Mechanical Turk, read the sentence contexts without the target words (e.g. "I went to deposit the check at the…") and completed each sentence by writing down the first word that came to mind. Items were counterbalanced across lists to ensure that each participant observed only one item from each triplet. After three rounds of cloze norming, responses for each item were obtained from an average of 39 participants (range: 33 – 48). Based on these norms, we selected the final set of 84 sentence triplets in which at least 67% of participants provided the expected target word across all three sentences (mean lexical constraint = 88%, SD = 8%).

In the EEG experiment, participants saw all three types of sentence contexts within each triplet (dominant-constraining, non-homograph-constraining, and subordinate-constraining), pseudorandomized and separated by at least 30 trials (Mean separation = 69 trials). To avoid repeating the same homograph twice within an experimental session, half of the dominant-constraining contexts and half of the subordinate-constraining contexts ended with the expected homographs, and the other half ended with another plausible but unexpected completion that was never provided in the cloze norming study (0% cloze). This excluded the possibility that, instead of directly reflecting representational activity related to form pre-activation itself, any form similarity effect reflected participants' recognition of a match between the form of the word that they had just predicted and a homograph that they had seen earlier in the experiment. The pairing of contexts with the expected and unexpected endings was counterbalanced across two experimental lists, ensuring that each participant saw an equal number of dominant-constraining and subordinate-constraining contexts that ended with expected and unexpected words. Because the non-homograph-constraining contexts always ended with the expected word (88% cloze), two thirds of all sentences in the experimental session ended with expected words.

## Characterization of stimuli

### *Semantic similarity between pairs of predicted target words*

Our design rested on the assumption that the predicted words in the *semantically-related* pairs were more semantically related to each other than in the *unrelated* pairs. To confirm that this was the case, we quantified the semantic similarity between the pairs of predicted word using

WordNet from the Natural Language Toolkit (NLTK) (Loper & Bird, 2002). WordNet is an English lexical database that organizes words based on their semantic relations in a hierarchical network, such as their super-subordinate relations (e.g. "meal" – "breakfast") and part-whole relations (e.g. "sand"– "beach"). Unlike other tools for measuring semantic relatedness, WordNet assigns different "senses" associated with a particular word to different Synsets. For example, "bank" has ten synsets, such as (a) *Synset('bank.n.01')*: sloping land (especially the slope beside a body of water) and (b) *Synset('depository_financial_institution.n.01')*: a financial institution that accepts deposits and channels the money into lending activities.

We manually identified the sense of each predicted target word in each sentence context within each triplet. Then, based on the identified senses, we calculated pairwise semantic similarity values using a path-based approach described by Wu & Palmer (ranging from 0 to 1, indicating low to high semantic similarity) (Wu & Palmer, 1994). To check that there were statistically significant differences in the semantic similarity between the predicted words in the *semantically-related* and the *unrelated* sentence pairs, we used a permutation approach. Specifically, we computed the mean similarity values for the *semantically-related* and *unrelated* pairs separately and took their difference value as our test statistic. We then randomly re-assigned the similarity values across these two types of sentence pairs, re-calculated the mean difference of the permuted values, and took the mean difference value for each randomization (1000 times) to build a null distribution. We considered the observed test statistic significant if it fell within the highest or lowest 2.5% of the null distribution. This analysis confirmed that the predicted words in the *semantically-related* pairs were indeed more semantically similar (Mean WordNet semantic similarity = 0.58, SD = 0.29) than the predicted words in the *unrelated* pairs (Mean WordNet semantic similarity = 0.25, SD = 0.16, $p < 0.001$.

We used the same methods to determine whether the semantic similarity of the predicted words differed between the *form-related* and *unrelated* sentence pairs. We found a small but significant difference (*form-related* vs. *unrelated*: Mean = 0.30 *vs.* 0.25, SD = 0.18 *vs.* 0.16, *p* = 0.011). As discussed in the Results, to ensure that this difference did not confound the form neural similarity effect, we carried out an RSA using a subset of *unrelated* pairs (56 in total) in which the semantic similarity of the predicted target words as well as other properties of the pre-target words were matched with the *form-related* pairs.

*Semantic similarity between pairs of pre-target words*

We also examined the semantic similarity between the pre-target words within each pair. Because these pre-target words varied in their syntactic category, instead of using WordNet, which only provides semantic similarity values for words within the same syntactic category (Loper & Bird, 2002), we used Word2Vec (Mikolov et al., 2013). We computed the cosine distance between the 300-dimensional vectors that corresponded to each pair of pre-target words using the Gensim natural language processing library in Python. A permutation-based test showed that the pre-target words in the *semantically-related* pairs were slightly more semantically similar than those in the *unrelated* pairs (*semantically-related* vs. *unrelated*: Mean Word2Vec semantic similarity = 0.16 *vs.* 0.11, SD = 0.26 *vs.* 0.17, *p* = 0.006). As discussed in the Results, to ensure that this difference did not confound our neural similarity effect, we compared the neural similarity values produced by the *semantically-related* and a subset of *unrelated* pairs (again in a total of 56 pairs) in which the semantic similarity of the pre-target words as well as other properties of the pre-target words were matched.

We found no significant differences in the semantic similarity of the pre-target words between the *form-related* and *unrelated* pairs (Mean = 0.13 *vs.* 0.11, SD = 0.22 *vs.* 0.17), $p = 0.17$.

*Similarity structure of other properties of the pre-target word*

To ensure that any observed neural similarity effects were not driven by other differences in the similarity structure of the pre-target words, we also examined several of their other properties: (1) their lexical probabilities, based on their preceding contexts, which we operationalized using the large language model, GPT-3 (Brown et al., 2020) [2]; (2) length, i.e., number of letters; (3) orthographic Levenshtein distance (OLD20, Balota et al., 2007); (4) concreteness (Brysbaert et al., 2014); (5) log frequency, based on the SUBTLEX database (Brysbaert & New, 2009); and (6) syntactic class, i.e., whether it was a content or a function word. For each of these variables, we calculated the pairwise absolute difference values and compared these difference values between the *semantically-related* and *unrelated* pairs, as well as between the *form-related* and *unrelated* pairs. Our permutation tests revealed no statistically significant differences (all *p*s > 0.09).

*Similarity structure of more general properties of the prior sentence contexts*

Finally, we extracted two additional measures for all sentences: (1) context length, i.e. the number of words prior to the target word (this ranged from 4-21 words); (2) contextual constraint, i.e. the probability of the most expected target word, as estimated using our cloze norms (this ranged from 67% to 100%). For each of these measures, we calculated the absolute difference values for every pair of contexts. Permutation-based statistical tests once again showed no

[2] Previous studies have shown that GPT-derived probability values correlate with human-based cloze estimates (e.g. Szewczyk & Federmeier, 2022). Indeed, in the present study, GPT-derived estimates of the lexical probabilities of the predicted target words correlated strongly with the cloze estimates of lexical probability that we had obtained from human participants ($r = 0.92$, $p < 0.001$).

difference between either the *semantically-related* and *unrelated* pairs, or between the *form-related* and *unrelated* pairs (all *p*s > 0.07).

## Participants

A power analysis showed that a sample size of 34 is required to achieve 80% power for detecting a medium-sized similarity effect (*d* = 0.5: (Wang, Kuperberg, et al., 2018; Wang et al., 2020); α = 0.05 for two-tailed paired t-tests). In total, 36 participants took part in the EEG experiment. The data of three participants were subsequently excluded due to termination of the experiment by one participant and excessive artifacts in two participants, leaving a final dataset of 33 participants (mean age: 20 years, range: 20 – 28; 20 males). All participants were native speakers of English and had no exposure to other languages before the age of 5. They were all right-handed with normal or corrected-to-normal vision. They were not taking psychoactive medication, and had never been diagnosed with a psychiatric or neurological disorder. Each participant gave informed consent following procedures approved by the Tufts University Social, Behavioral, and Educational Research Institutional Review Board.

## Procedure

The sentences were presented on a computer monitor, positioned approximately one meter away from the participant, using PsychoPy 2.0 software. Stimuli were presented in white Arial font on a black background, with three characters covering approximately one degree of visual angle. The trial started with a fixation ("++++") for 1200ms, followed by a 400ms blank. Then, each sentence was presented word by word, with each word being presented for 300ms, followed by a 400ms blank screen. The final word was presented together with a period, followed by a

1200ms blank. After one-sixth of the trials, at random, participants read a True/False statement based on the sentence that they had just read and judged whether these statements were correct/incorrect by pressing one of two buttons with their left hand. This helped ensure that they read the sentences for comprehension. In all other trials, the word "NEXT" appeared, and participants pressed another button to proceed to the next trial.

All 252 sentences were presented over nine blocks, with each block containing 28 sentences and lasting about six minutes. The experiment lasted about 1.5 hours, including EEG preparation, instructions and a short practice session consisting of five sentences.

## EEG data acquisition and preprocessing

The EEG dataset was acquired using a Biosemi ActiveTwo system from 64 active electrodes arranged according to the standard 10-20 system montage. Signals were digitized at 512Hz, with a passband of DC – 104Hz. The EEG dataset was preprocessed using Fieldtrip, an open-source Matlab toolbox (Oostenveld et al., 2011). The raw EEG signals were referenced offline to the average of the left and right mastoid electrodes. Visually identified channels with excessive noise were excluded (removed) from further analysis (on average, 5 out of the 64 channels across participants) [3]. A 4th-order Butterworth bandpass filter of 0.1-30Hz was applied to the data to remove slow drift and high-frequency noise (Luck, 2014) [4]. Next, we segmented the data into epochs, which spanned a time window from -1500ms to 1500ms, relative to the onset of the target word. Trials that exhibited high variance across channels were excluded from further

[3] We chose to exclude rather than interpolate bad channels to avoid any potential distortion of spatiotemporal patterns over the entire set of channels, which would occur if the true electrical activity at a bad channel was not highly correlated with activity at its neighboring channels.

[4] High-pass filters can sometimes produce spurious multivariate effects (van Driel et al., 2021) and/or distort the timing of such effects (Acunzo et al., 2012). However, when we carried out the RSA without applying any high-pass filter, we found qualitatively similar effects.

analysis. We then carried out an Independent Component Analysis (ICA; Bell & Sejnowski, 1997), and identified and removed any components that showed temporal and spatial characteristics of eye movements (on average, one component per participant was removed). After these steps, the data was visually inspected and any remaining artifacts were removed. As noted earlier, we excluded the data of two participants who had a trial rejection rate of over 30%. For the remaining 33 participants, 13% of trials (on average) were rejected, with no significant differences in rejection rates across the three types of contexts within the sentence triplets (i.e. dominating-constraining, subordinate-constraining, non-homograph-constraining: $F_{(2,96)} < 1$).

After artifact rejection, there remained, on average, 56 full triplets of trials across all participants. Because we were interested in brain activity reflecting the *pre-activation* of upcoming target words, we did not apply any baseline correction. Instead, we mean-centered activity within each trial by subtracting the average activity of the full epoch from the activity at each sampling point. This approach effectively eliminated any potential shift caused by slow drift.

## Representational Similarity Analysis (RSA)

We carried out an RSA on the single-trial EEG data (Figure 1B), using MATLAB 2020a (MathWorks) together with custom-written scripts. When combining EEG/MEG with RSA, one can potentially examine similarities amongst spatial patterns, temporal patterns, or a combination of both, depending on the research question. For example, in a previous MEG study (Wang, Kuperberg, et al., 2018), we examined spatial and temporal patterns separately because we were interested in detecting item-specific *spatial patterns* at the head surface (stemming from distributed underlying sources), and in source localizing the specific neuroanatomical regions that produced item-specific *temporal patterns*. In another MEG and EEG study (Wang et al., 2020), we only examined spatial patterns because we were interested in detecting the pre-activation of animacy-related semantic features at the head surface (again assumed to stem from distributed underlying cortical sources). In the present EEG study, our primary goal was to determine whether there was *any* evidence for a dissociation in the pre-activation of meaning *versus* form. Therefore, following previous RSA studies (e.g., Choi et al., 2021; He et al., 2021; Lyu et al., 2019), we *combined* spatial and temporal information to maximize our chances of detecting pre-activation effects.

In each participant, on each trial, we extracted vectors of EEG data that represented the spatiotemporal patterns of activity produced across all EEG channels at consecutive sampling intervals. We chose a sampling interval of 31ms because this corresponded to a full cycle of oscillatory activity at approximately 30Hz — the low-pass filter that we applied during pre-processing [5]. Given the 512Hz sampling rate, each 31ms sampling interval contained 16 data points. Thus, each sampling interval captured sufficient temporal information while maintaining reasonable temporal precision. Within our time window of interest, i.e. -700 to 0ms relative to the onset of the target words, at each successive sampling interval, using a step size of one sampling point, we computed a Pearson's *r* value for every pair of trials and averaged these values across all *semantically-related* pairs (*financial* "bank" – "loan") and across all unrelated pairs. We replaced the values of any sampling interval that lay outside the time window of interest with zeros (see Michelmann et al., 2016).

As discussed in our previous work (Wang & Kuperberg, 2023; Wang et al., 2020), Pearson's *r* neural similarity measure can capture *both* item-specific as well as non-item-specific activity (i.e. evoked responses). The large number of *unrelated* pairs across triplets (31,374 pairs)

[5] Following a reviewer's suggestion, we repeated the RSA using different sampling interval sizes (i.e. 0ms, 20ms, and 40ms). We found qualitatively similar effects for all interval sizes.

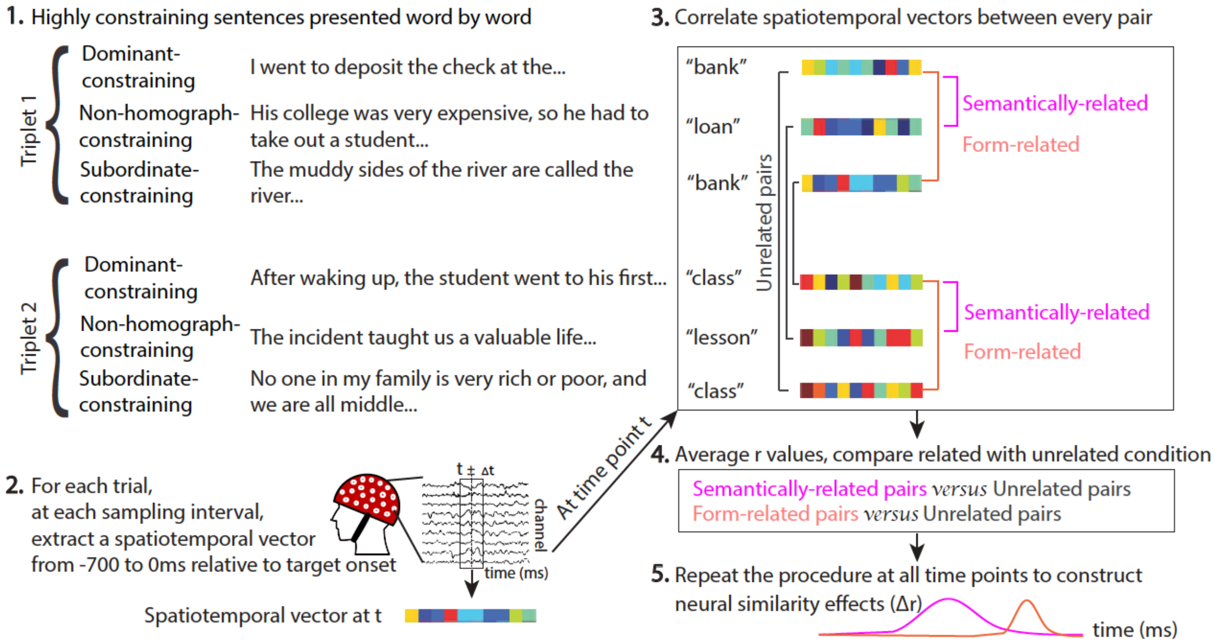
provided a highly reliable measure of non-item-specific neural similarity (Wang, Kuperberg, et al., 2018). Therefore, to isolate the effect of item-specific semantic similarity, at each sampling point from the onset of the pre-target word (t = -700ms) until target word onset (t = 0), we carried out paired t-tests comparing the neural similarity values produced by the *semantically-related* and the *unrelated* pairs. We took the same approach to isolate the effect of item-specific form similarity.

To account for multiple comparisons, we used cluster-based permutation tests (Maris & Oostenveld, 2007). Adjacent data points that showed effects at $p \leq 0.05$ were considered temporal clusters, and within each cluster, we summed the individual t-statistics and took the summed t-value as the cluster-level test statistic. We then built a permutation distribution by randomly shuffling the conditional labels of the neural similarity values at all time points in all participants 10,000 times. We calculated cluster-level statistics on each permutation and selected the largest cluster-level statistic to construct a null distribution. We compared the observed cluster-level test statistic against this null distribution, and took any temporal clusters falling within the highest or lowest 2.5% of the distribution to be significant.

To visualize the item-specific semantic and form pre-activation effects, we subtracted the average neural similarity values produced by *unrelated* pairs from those produced by the *semantically-related* pairs and *form-related* pairs. The time series of neural similarity values produced by all three types of sentence pairs are shown in Supplementary Materials.

**Figure 1. Schematic illustration of our Representational Similarity Analysis stream**. **1.** Participants read triplets of highly constraining sentences (pseudorandomly presented). **2.** For each trial, and at the center of each sampling interval ($t \pm \Delta t$) within our time window of interest (from 700 to 0ms relative to the target onset), we extracted a vector of EEG data that represented the spatiotemporal pattern of activity produced across all EEG channels. **3.** We calculated Pearson's *r* between each pair of spatiotemporal vectors to quantify the neural similarity between the patterns of EEG activity produced by predicted upcoming words in (a) *semantically-related* pairs where

the predicted words shared semantic but not form features, (b) *form-related* homograph pairs where the predicted words shared form but not semantic features, and (c) *unrelated* pairs where the predicted words shared neither form nor semantic features. **4.** To isolate item-specific semantic and form similarity effects, we averaged the *r* values separately for the *semantically-related* and *form-related* pairs, and statistically compared them with the patterns produced by the *unrelated* pairs. **5.** We subtracted the average neural similarity values of the *unrelated* pairs from those of *semantically-related* and *form-related* pairs (Δr) at each sampling point to construct time series of the *semantic* and *word-form* pre-activation effects.


1. Highly constraining sentences presented word by word
Triplet 1
Dominant-constraining
I went to deposit the check at the...
Non-homograph-constraining
His college was very expensive, so he had to take out a student...
Subordinate-constraining
The muddy sides of the river are called the river...
Triplet 2
Dominant-constraining
After waking up, the student went to his first...
Non-homograph-constraining
The incident taught us a valuable life...
Subordinate-constraining
No one in my family is very rich or poor, and we are all middle...
2. For each trial, at each sampling interval, extract a spatiotemporal vector from -700 to 0ms relative to target onset
t ± Δt
channel
time (ms)
Spatiotemporal vector at t
At time point t
3. Correlate spatiotemporal vectors between every pair
"bank"
"loan"
"bank"
"class"
"lesson"
"class"
Unrelated pairs
Semantically-related
Form-related
Semantically-related
Form-related
4. Average r values, compare related with unrelated condition
Semantically-related pairs versus Unrelated pairs
Form-related pairs versus Unrelated pairs
5. Repeat the procedure at all time points to construct neural similarity effects (Δr)
time (ms)

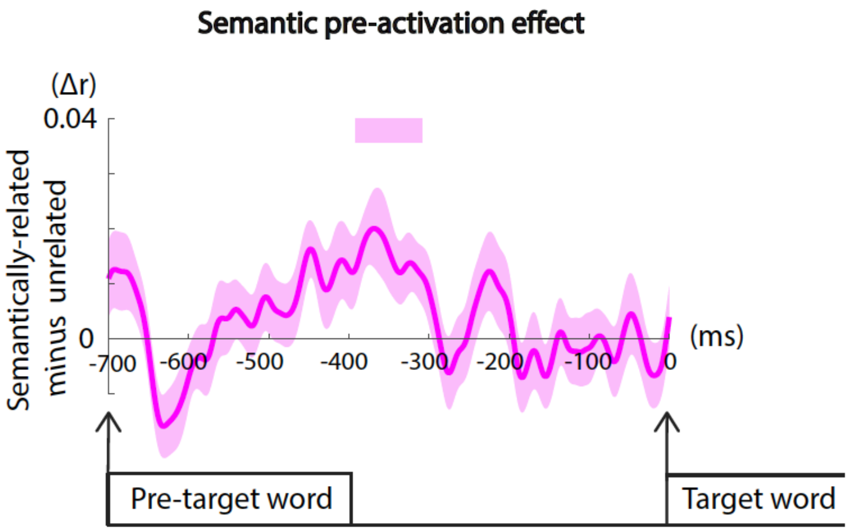
Finally, we compared the *timing* of the *form* and *semantic* pre-activation effects. For this analysis, we took a bootstrapping approach because of its flexibility and robustness, even when the distribution of the underlying data is unknown (e.g. Manly, 1997). First, we identified the peak latency for each effect, based on its smoothed grand-average time course (for smoothing, we used a moving average of 20 sampling points). We then calculated the difference between the two identified peak latencies and compared this value against a sampling distribution, which we built by resampling the neural similarity effect of all participants with replacement 5000 times, and taking the difference in the identified peak latencies between -700 and 0ms for each iteration. We considered the observed peak latency difference to be significant if it fell within the highest or lowest 2.5% of the sampling distribution.

## Results

### Semantic pre-activation effect

When we compared the neural similarity values produced by pairs that predicted semantically-related words with those produced by pairs that predicted unrelated words, we found a significant effect between -391 and -309ms before the onset of the predicted target word, i.e. between 309 and 391ms following the onset of the pre-target word (cluster-based permutation test: *p* = 0.003), see Figure 2.

**Figure 2.** The time course of the semantic pre-activation effect. Shown is the mean difference in neural similarity values between the *semantically-related* and *unrelated* pairs (Δr). Standard errors are indicated with shading. Between -391 and -309ms before the onset of the predicted target word, i.e. between 309 and 391ms after the onset of the pre-target word, indicated with the pink horizontal bar, the *semantically-related* pairs produced significantly greater neural similarity values than the *unrelated* pairs. The predicted target word was presented at 0ms, and the pre-target word was presented at -700ms, with a duration of 300ms.



As noted in the Methods, pairs of *pre-target* words were also slightly more semantically similar in the *semantically-related* sentence pairs than in the *unrelated* sentence pairs. Therefore, to ensure that the observed effect was driven by the semantic similarity between the *predicted* upcoming words, rather than the pre-target words, we selected a subset of *unrelated* pairs in which the pre-target words were matched on semantic similarity as well as other properties of the pre-target words with the *semantically-related* pairs. This analysis again revealed a significant effect (between -412 and -320ms prior to the onset of the target word, i.e. between 288 and 380ms following the onset of the pre-target word, $p = 0.001$). This suggests that the neural similarity effect indexed the pre-activation of semantic features of the upcoming target word rather than the semantic features of the pre-target word.

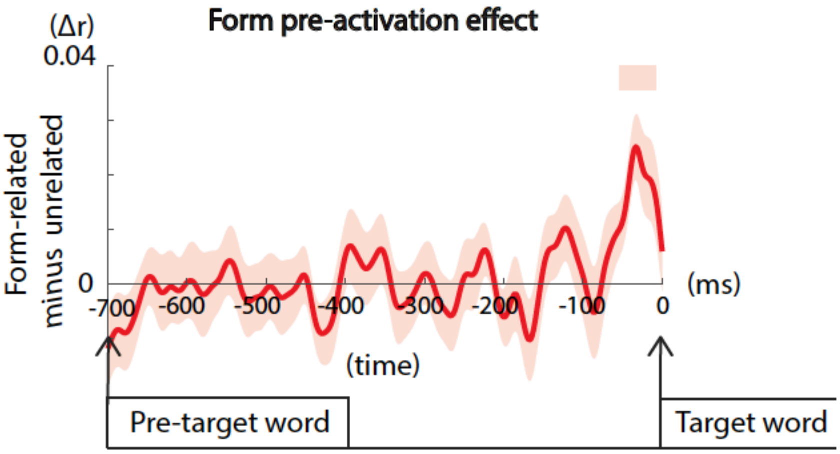
Finally, to ensure that the observed effect was not driven by the general differences in the semantic similarity of the prior context, we carried out a cluster-based permutation test from -1400 to -700ms *before* the onset of the pre-target word. We found no significant difference between the *semantically-related* and *unrelated* pairs ($p = 0.169$).

### Form pre-activation effect

When we compared the neural similarity values produced by pairs that predicted form-related words with those produced by pairs that predicted unrelated words, we found a significant effect between -53ms and -8ms before the onset of the target word, i.e. between 647 and 692ms following the onset of the pre-target word (cluster-based permutation test: $p$ = 0.025), see Figure 3.

As noted in the Methods, pairs of the predicted target words were very slightly more semantically similar to each other in the *form-related* sentence pairs than in the *unrelated* sentence pairs. However, when we repeated the analysis using a subset of *unrelated* pairs in which the predicted target words were matched in semantic similarity with the *form-related* pairs, we again found a significant effect (between -47 and -8ms prior to the onset of the target word, i.e. between 643 and 694ms following the onset of the pre-target word), $p$ = 0.025.

**Figure 3.** The time course of the form pre-activation effect. Shown is the mean difference in neural similarity values between the *form-related* and *unrelated* pairs (Δr). Standard errors are indicated with shading. Between -53ms and -8ms before the onset of the predicted target word, i.e. between 647 to 692ms after the onset of the pre-target word (indicated with the red horizontal bar), the *form-related* pairs produced significantly greater neural similarity values than the *unrelated* pairs. The predicted target word was presented at 0ms, and the pre-target word was presented at -700ms, with a duration of 300ms.



### Timing of the semantic pre-activation effect *versus* the form pre-activation effect

The analysis above suggested that the semantic pre-activation effect preceded the form pre-activation effect, with the semantic pre-activation effect peaking at -367ms, and the form pre-activation effect peaking at -25ms, each relative to the target word onset. To determine whether this difference in peak latency was statistically significant, we used a bootstrapping sampling approach. This confirmed that the peak latency was significantly earlier for the semantic pre-activation than the form pre-activation effect ($t_{(32)} = 7.29$, $p < 0.001$) [6].

## Discussion

A fundamental question in language comprehension is whether the brain generates predictions of both the meaning and the form of upcoming words during language comprehension.

[6] Following a reviewer's request, we carried out two additional post-hoc tests. First, we statistically compared the peak latency difference between the semantic and form similarity effects using a Jackknife resampling approach. This also showed a significant effect. Second, we directly compared the neural similarity values produced by the *semantically-related* and *form-related* pairs within the two time windows that revealed significant differences in the analysis above, i.e. in comparison with the *unrelated* control condition. Between -391 and -309ms, the *semantically-related* pairs produced greater neural similarity values than the *form-related* pairs although this effect was marginal ($t_{(32)} = 1.99$, $p = 0.055$). Between -53 and -8ms, the *form-related* pairs produced significantly greater neural similarity values than the *semantically-related* pairs ($t_{(32)} = 2.53$, $p = 0.017$).

In the present study, we addressed this question by combining EEG with RSA. We found that, relative to *unrelated* pairs, pairs of contexts that predicted words with related meanings but distinct linguistic forms (e.g. *financial* "bank" – "loan") produced more similar neural patterns from 300-400ms before the onset of the predicted target. Conversely, pairs of contexts that predicted words with the same form but different meanings (i.e., homographs, e.g. *financial* "bank" – *river* "bank") produced more similar neural patterns immediately before target word onset.

Although contexts that predict *semantically-related* words are naturally more similar to one another than those that predict *unrelated* words, we detected a semantic pre-activation effect even when the semantic similarity of the pre-target words was matched between the *related* and *unrelated* pairs, and the effect only became significant *after* pre-target onset. We also ensured that the similarity of both types of *related* pairs was matched to that of the *unrelated* pairs with respect to context length and contextual constraint, as well as multiple other properties of the pre-target words (word length, orthographic neighborhood, concreteness, word frequency, syntactic category, and lexical predictability). We therefore take these neural similarity effects as evidence that the meaning and form of the expected words were predictively pre-activated before new bottom-up input became available.

## Pre-activation of meaning

Our finding of a semantic pre-activation effect provides direct support for the theory that comprehenders use prior contexts to pre-activate the semantic features associated with expected upcoming words. Previous support for this theory has been somewhat indirect, coming from studies reporting facilitated processing of unexpected words that share semantic features with expected words (e.g. Federmeier & Kutas, 1999). In the present study, our use of RSA allowed us to directly capture neural representations of these pre-activated semantic features before the arrival of new bottom-up input. For example, the semantic features associated with the pre-activated word "bank" and "loan" (e.g., < financial >, <service>) overlapped with each other. In the brain, semantic features are thought to be represented in a distributed fashion across multiple cortical regions (Huth et al., 2016), which project to the scalp surface as unique spatiotemporal patterns. Therefore, the spatiotemporal patterns produced by predicted words that shared semantic features were more similar than those produced by predicted words in *unrelated* pairs that shared no semantic features. This finding also extends our previous RSA work showing that comprehenders also pre-activate distributed semantic features in contexts that, instead of constraining for single words, constrain for broader semantic categories such as animacy (e.g., <animate> following "*The lifeguard cautioned the ...*") (Wang et al., 2020).

**Pre-activation of word-form**

Of even greater theoretical significance is our finding of an increase in similarity between the neural patterns produced by predicted words that overlapped only in their linguistic form. As outlined in the Introduction, several studies have claimed to find evidence of form-level prediction. Again, however, this previous evidence primarily comes from ERP studies that examined the *effects* of prior predictions on processing new bottom-up input. Moreover, the interpretation of these ERP effects is challenging, particularly as some findings have not been replicated (Nieuwland, 2019; Nieuwland et al., 2018).

In the present study, our use of RSA methods, in combination with our novel homograph design, allowed us to circumvent many of these interpretational issues, and to directly dissociate the pre-activation of linguistic form from meaning. For example, in the contexts, "*I went to deposit* *the check at the ...[bank]"* and *"The muddy sides of the river are called the river ...[bank]"*, the predicted concepts (<financial-bank> and <river-bank>) do not share common semantic features. However, both concepts are associated with the same word-form, "BANK". In the brain, form features are thought to be encoded in a distributed fashion along the posterior ventral visual pathway (Dehaene et al., 2005). Therefore, when projected to the scalp surface, the pre-activation of these same word-forms produced spatiotemporal patterns that were more similar to one another than the patterns produced by predicted words in the *unrelated* pairs that did not share word-form features. As such, our findings provide direct neural evidence that, at least in highly constraining sentences, the brain is able to generate predictions not only at the level of semantic features, but also at level of specific word-forms (see Kuperberg & Jaeger, 2016, Section 5) [7].

## The pre-activation of meaning preceded the pre-activation of form

Our findings also shed important light on *when* the meaning and form of pre-activated words became available. The pre-activation of semantic features was detected ~300ms after the onset of the pre-target word. This is consistent with previous RSA studies (Hubbard & Federmeier, 2021; Wang, Kuperberg, et al., 2018; Wang et al., 2020), suggesting that the brain uses prior

[7] A reviewer raised the possibility that the form pre-activation effect actually reflected an indirect semantic pre-activation effect. On this account, the pre-activation of a homograph's contextually appropriate meaning (e.g. <financial-bank>) *indirectly* led to the additional pre-activation of its contextually inappropriate meaning (e.g. <river-bank>), and so the form pre-activation effect that we detected actually reflected overlap of these indirectly pre-activated semantic features. Note that account this would still be consistent with the claim that comprehenders pre-activate upcoming word-form forms. This is because the only way that the alternative meaning of the homograph could have been pre-activated if its *form* was also pre-activated. Nonetheless, we explored this possibility by carrying out a post-hoc analysis that contrasted the neural similarity values produced by *indirectly related* pairs (e.g. *river* "bank" – "loan") with those produced by the *unrelated* pairs, each averaged across the -53 – -8ms time window time window where we detected the form pre-activation effect. This analysis did not reveal a significant effect. Of course, this does not exclude the possibility that the alternative meaning of the predicted homographs was pre-activated in another time window. However, a comprehensive investigation of how different meanings associated with homographs are pre-activated goes beyond the scope of the present study. To investigate this, one would need to construct an additional set of highly constraining sentence contexts that predict words that are semantically related to the subordinate meanings of the predicted homographs.

contexts to generate semantic predictions as soon as it is able to do so. Also consistent with our previous RSA studies using different presentation rates (1000ms per word: Wang, Kuperberg, et al., 2018; 550ms per word: Wang et al., 2020), the effect was transient — lasting for only 100ms and disappearing 400ms before the onset of the target word.

In contrast, the form pre-activation effect was detected later — ~600ms following the onset of the pre-target word, i.e. ~100ms prior to the target word. This later detection of word-form pre-activation is consistent with hierarchical predictive accounts of language comprehension, including predictive coding, which posit that pre-activated information is propagated down the linguistic hierarchy over time — from higher-levels representing semantic features to lower levels that represent form features (e.g. see Kuperberg & Jaeger, 2016, Section 5).

Notably, the time lag between the semantic and word-form pre-activation effects was relatively long — 300ms. This mirrors the temporal separation between conceptual retrieval and word-form encoding during language production (Indefrey & Levelt, 2004). As such, these latency results are consistent with the proposal that strong top-down predictions generated during language comprehension may be implemented by the same circuitry that supports language production (Federmeier, 2007; Martin et al., 2018; Pickering & Garrod, 2013).

On the other hand, it is important to consider an alternative account of why the form pre-activation effect was detected so late: It is possible that participants actually pre-activated word-form information earlier, but that this pre-activation was difficult to detect using our signal-averaging methods. Word-form features are thought to be encoded over a shorter time scale than semantic features (Kiebel et al., 2008). Therefore, any jitter in the timing of a highly-transient word-form pre-activation effect across trials would have made it difficult to detect the initial generation of form-level predictions. However, immediately before the presentation of the upcoming word, when this pre-activated information became relevant for bottom-up form-level analysis, its latent representation might have been "reignited", and brought into an active state across all trials (Sprague et al., 2016; Stokes, 2015), making the word-form pre-activation effect easier to detect.

To investigate this possibility, future studies could carry out the same experiment using a faster presentation rate (e.g. 500ms per word). If the form-level pre-activation effect disappears at this faster rate, then this would imply that there was not enough time for pre-activation to reach form features before the arrival of new bottom-up input, and that the interval between the pre-activation of meaning and form is fixed (i.e. 300ms). This would provide support for a prediction-by-production account in which the generation of form-level predictions is limited to slow presentation rates (see Ito et al., 2016). If, on the other hand, a form pre-activation effect can still be detected just before the presentation of the target word, then this would suggest that such predictions are generated earlier and brought online just before they are needed, regardless of presentation rate. This would, in turn, imply that, in predictive contexts, form-level predictions are generated rapidly and routinely during real-time natural language comprehension (DeLong et al., 2021).

## Conclusion

To summarize, by combining EEG with RSA, we provide direct neural evidence for the pre-activation of both the meaning and form of expected words in predictive language contexts. Moreover, our finding that the pre-activation of meaning preceded the pre-activation of form is consistent with hierarchical predictive accounts of language comprehension, which posit that information flows from higher to lower levels of linguistic representations before new input becomes available.

## Open Practices Statement

The data and materials for the experiment are available at https://osf.io/u6tpx/. The experiment was not preregistered.

## Acknowledgments

This research was supported by the National Institutes of Health under Award Number R01HD082527. We thank Sophie Greene, Feng Cheng, and Edward Alexander for their assistance with stimuli preparation and data collection.

## References

Acunzo, D. J., Mackenzie, G., & van Rossum, M. C. (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. *J Neurosci Methods*, *209*(1), 212-218. https://doi.org/10.1016/j.jneumeth.2012.06.011

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327-3338. https://doi.org/10.1016/s0042-6989(97)00121-1

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: prediction takes precedence. *Cognition*, *136*, 135-149. https://doi.org/10.1016/j.cognition.2014.10.017

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 1877-1901. https://doi.org/10.5555/3495724.3495883

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911. https://doi.org/10.3758/s13428-013-0403-5

Choi, H. S., Marslen-Wilson, W. D., Lyu, B., Randall, B., & Tyler, L. K. (2021). Decoding the Real-Time Neurobiological Properties of Incremental Semantic Interpretation. *Cereb Cortex*, *31*(1), 233-247. https://doi.org/10.1093/cercor/bhaa222

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455-462.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256-266. https://doi.org/10.1162/Jocn.1994.6.3.256

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7), 335-341.

DeLong, K. A., Chan, W. H., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, *56*(4), e13312. https://doi.org/10.1111/psyp.13312

DeLong, K. A., Chan, W. h., & Kutas, M. (2021). Testing limits: ERP evidence for word form preactivation during speeded sentence reading. *Psychophysiology*, *58*(2), e13720. https://doi.org/10.1111/psyp.13720

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, P.H.S.]. *Nature Neuroscience*, *8*(8), 1117-1121. https://doi.org/10.1038/nn1504

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.]. *Brain and Language*, *127*(1), 55-64. https://doi.org/10.1016/j.bandl.2012.08.004

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension [Research Support, N.I.H., Extramural

Review]. *Psychophysiology*, *44*(4), 491-505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469-495. https://doi.org/10.1006/Jmla.1999.2660

Federmeier, K. D., Mai, H., & Kutas, M. (2005). Both sides get the point: hemispheric sensitivities to sentential constraint. *Memory and Cognition*, *33*(5), 871-886. https://doi.org/10.3758/BF03193082

Gerrig, R. J., & McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes*, *26*(2-3), 67-86.

Groppe, D. M., Choi, M., Huang, T., Schilz, J., Topkins, B., Urbach, T. P., & Kutas, M. (2010). The phonemic restoration effect reveals pre-N400 effect of supportive sentence context in speech perception. *Brain Research*, *1361*, 54-66. https://doi.org/10.1016/J.Brainres.2010.09.003

He, T., Boudewyn, M. A., Kiat, J. E., Sagae, K., & Luck, S. J. (2021). Neural correlates of word representation vectors in natural language processing models: Evidence from representational similarity analysis of event-related brain potentials. *Psychophysiology*, e13976. https://doi.org/10.1111/psyp.13976

Hubbard, R. J., & Federmeier, K. D. (2021). Representational Pattern Similarity of Electrical Brain Activity Reveals Rapid and Specific Prediction during Language Comprehension. *Cereb Cortex*, *31*(9), 4300-4313. https://doi.org/10.1093/cercor/bhab087

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453-458. https://doi.org/10.1038/nature17637

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101-144. https://doi.org/10.1016/j.cognition.2002.06.001

Ito, A., Corley, M., Pickering, M., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157-171. https://doi.org/10.1016/j.jml.2015.10.007

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A Hierarchy of Time-Scales and the Brain. *PLoS Computational Biology*, *4*(11), e1000209. https://doi.org/Artn E1000209

10.1371/Journal.Pcbi.1000209

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32-59. https://doi.org/10.1080/23273798.2015.1102299

Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: an ERP study [Comparative Study

Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *Journal of Cognitive Neuroscience*, *23*(5), 1230-1246. https://doi.org/10.1162/jocn.2010.21452

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP) [Research Support, N.I.H., Extramural

Review]. *Annual Review of Psychology*, *62*, 621-647. https://doi.org/10.1146/annurev.psych.093008.131123

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326-338. https://doi.org/10.1016/j.jml.2009.06.004

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts [Comparative Study

Randomized Controlled Trial

Research Support, N.I.H., Extramural]. *Journal of Cognitive Neuroscience*, *25*(3), 484-502. https://doi.org/10.1162/jocn_a_00328

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. *arXiv preprint* cs/0205028. https://arxiv.org/abs/cs/0205028

Luck, S. J. (2014). Chapter 7: Basics of Fourier Analysis and Filtering.

Lyu, B., Choi, H. S., Marslen-Wilson, W. D., Clarke, A., Randall, B., & Tyler, L. K. (2019). Neural dynamics of semantic composition. *Proceedings of the National Academy of Sciences of* *the United States of America*, *116*(42), 21318-21327. https://doi.org/10.1073/pnas.1903402116

Manly, B. F. J. (1997). *Randomization, bootstrap, and Monte Carlo methods in biology* (2nd ed.). Chapman & Hall.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports*, *8*(1), 1079. https://doi.org/10.1038/s41598-018-19499-4

Michelmann, S., Bowman, H., & Hanslmayr, S. (2016). The temporal signature of memories: Identification of a general mechanism for dynamic memory replay in humans. *PLoS Biol*, *14*(8), e1002528. https://doi.org/10.1371/journal.pbio.1002528

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations (ICLR), Workshop Track Proceedings, Scottsdale, Arizona.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops [Comparative Study]. *Biological Cybernetics*, *66*(3), 241-251. https://doi.org/10.1007/BF00198477

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*(2&3), 131-157. https://doi.org/10.1080/01638539809545042

Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neurosci Biobehav Rev*, *96*, 367-400. https://doi.org/10.1016/j.neubiorev.2018.11.019

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., . . . Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *Elife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

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*, 1. https://doi.org/10.1155/2011/156869

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language*, *67*(4), 426-448. https://doi.org/10.1016/j.jml.2012.07.003

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension [Research Support, Non-U.S. Gov't]. *Behavioral and Brain Sciences*, *36*(04), 329-347. https://doi.org/10.1017/S0140525X12001495

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79-87. https://doi.org/10.1038/4580

Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. S. (2011). Anomalies at the borderline of awareness: an ERP study. *Journal of Cognitive Neuroscience*, *23*, 514-523.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron*, *91*(3), 694-707. https://doi.org/10.1016/j.neuron.2016.07.006

Stokes, M. G. (2015). ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394-405.

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*. https://doi.org/10.1016/j.jml.2021.104311

Taylor, W. (1953). 'Cloze' procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415-433.

Terporten, R., Schoffelen, J. M., Dai, B., Hagoort, P., & Kosem, A. (2019). The relation between alpha/beta oscillations and the encoding of sentence induced contextual information. *Scientific Reports*, *9*(1), 20255. https://doi.org/10.1038/s41598-019-56600-x

Van Berkum, J. J. A. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and Pragmatics: From Experiment to Theory* (pp. 276-316). Palgrave Macmillan.

van Driel, J., Olivers, C. N. L., & Fahrenfort, J. J. (2021). High-pass filtering artifacts in multivariate classification of neural time series data. *J Neurosci Methods*, *352*, 109080. https://doi.org/10.1016/j.jneumeth.2021.109080

Vuong, L. C., & Martin, R. C. (2013). Domain-specific executive control and the revision of misinterpretations in sentence comprehension. *Language, Cognition and Neuroscience*, *29*(3), 312-325. https://doi.org/10.1080/01690965.2013.836231

Wang, L., Hagoort, P., & Jensen, O. (2018). Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. *Journal of Cognitive Neuroscience*, *30*(3), 432-447. https://doi.org/10.1162/jocn_a_01190

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

Wang, L., & Kuperberg, G. R. (2023). Better together: integrating multivariate with univariate methods, and MEG with EEG to study language comprehension. *Language, Cognition and Neuroscience*. https://doi.org/10.1080/23273798.2023.2223783

Wang, L., Wlotko, E., Alexander, E. J., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural evidence for the prediction of animacy features during language comprehension: Evidence from MEG and EEG Representational Similarity Analysis. *Journal of Neuroscience*, *40*(16), 3278-3291. https://doi.org/10.1101/709394

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM.