When the Subtitle Doesn't Match: Semantic Tracking of Naturalistic Speech Predicts Cross-Modal N400 Priming for Visual Word Probes

Ian Martindale (imartindale@ucsd.edu) Haoyin Xu (hyx002@ucsd.edu)
Seana Coulson (scoulson@ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093, USA

Abstract

Researchers studying semantic processing during continuous speech have proposed temporal response function (TRF)–based measures of semantic tracking as a complement to the N400 event-related potential (ERP). However, the interpretation of these measures and their relation to the well-established N400 remain unclear. Here, we asked whether semantic tracking during naturalistic podcast listening relates to cross-modal semantic priming, as reflected in the N400 elicited by intermittently presented visual word probes. EEG was recorded while participants listened to podcast excerpts containing time-locked visual probes that varied in probe–context relatedness. In parallel, TRF decoding models were used to estimate individual semantic tracking performance from the speech stream. We found that less related probes elicited larger N400 amplitudes, and N400 sensitivity to probe–context semantic distance scaled with semantic tracking performance. These results suggest that TRF-based semantic tracking and probe-evoked N400 priming reflect a shared discourse-level semantic representation during naturalistic listening.

Keywords: EEG, Temporal Response Function, Semantic Processing

Introduction

A central question in psycholinguistics concerns how comprehenders construct and update semantic representations in real time as language unfolds. One major source of evidence comes from the event-related potential (ERP) literature, particularly the studies of the N400, a negative-going ERP that peaks approximately 400 ms after the onset of a potentially meaningful stimulus (Kutas & Hillyard, 1980; Kutas & Federmeier, 2011). Across decades of carefully manipulated word-pair and sentence-level paradigms, researchers have shown that the N400 component is larger (more negative) when a target word is less expected or less compatible with the preceding context (Van Petten & Luka, 2012; Frank, Otten, Galli, & Vigliocco, 2015a). For this reason, the N400 is widely interpreted as an index of context-sensitive semantic processing, with larger negativities reflecting greater difficulty.

However, most such research has relied on artificial paradigms of limited ecological validity. In real-world settings, meaning is often built up gradually over extended narratives, and listeners interpret speech while concurrently processing information in other modalities (e.g., scrolling on a phone while listening to others). Several ERP studies have therefore moved beyond isolated sentences, demonstrating that discourse-level context can modulate N400 amplitude (van Berkum, Hagoort, & Brown, 1999; Nieuwland & Van Berkum, 2006; Otten & Van Berkum, 2007). Nonetheless, much of this evidence still depends on carefully manipulated written materials, and comparatively fewer studies test discourse-level representations during naturalistic, continuous speech comprehension.

Most recently, EEG researchers have begun to address this gap by modeling word-by-word neural responses during story listening, revealing sensitivity to contextual semantic information in narratives (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018; Levari & Snedeker, 2024). A prominent approach in this line of work is temporal response function (TRF) modeling, which estimates how stimulus features (e.g., acoustic, lexical, or semantic properties of the speech) predict the ongoing EEG signal. Related decoding models use similar functions to predict features of the speech from the EEG. TRF-based semantic measures are often assumed to reflect comprehension because they are sensitive to speech intelligibility and show timing and scalp distributions reminiscent of the classic N400 (Broderick et al., 2018; Crosse, Di Liberto, Bednar, & Lalor, 2016). However, it remains unclear what these measures uniquely reflect, as they may index semantic processing but also covary with factors like attention and comprehension. Linking TRF measures to better established neural signatures of context-dependent meaning can help clarify what aspects of semantic processing they reflect (Kutas & Federmeier, 2011; Van Petten & Luka, 2012).

Current Study

To address these issues, we recorded EEG while participants listened to excerpts from a podcast and time-locked ERPs to visually presented probe words that appeared intermittently. The probe words were designed so that their relationship to the ongoing discourse context varied, with each probe presented once in a semantically related and once in an unrelated context. This allowed us to test whether discourse-level semantic representations constructed during naturalistic speech processing are available to modulate the probe-evoked N400 as a function of probe–context relatedness.

In parallel, we used TRF decoding models (Crosse et al., 2016) to quantify each participant's semantic tracking performance, indexing how reliably their neural activity entrained to semantic-related information in the speech stream. Critically, this approach enables us to examine the relationship be-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

tween (i) semantic tracking measures derived from naturalistic TRF modeling and (ii) a well established neural signature of context-sensitive meaning processing, namely the cross-modal semantic priming effect indexed by the N400 elicited by visual probes (Holcomb & Anderson, 1993). If semantic tracking reflects discourse-level semantic processing, then individuals with better semantic decoding in the podcast should show larger cross-modal priming effects on the visually presented word probes.

Materials and Methods

Participants

16 students (mean age = 20.4) from a large public university in the United States participated and received course credit as compensation. All participants were native English speakers who reported normal hearing, normal or corrected-to-normal vision, and no known neurological conditions or impairments.

Stimuli Design

Auditory Stimulus Participants listened to ten one-minute podcast clips, all sourced from the episode "House on Loon Lake" of NPR's *This American Life* (Glass, 2001). The podcast focuses on a group of adults recalling time they spent exploring an abandoned house during their childhood. The clips were extracted to end at logical narrative breaks. They were also selected to give a summarized overview of the narrative presented in the full podcast episode to support comprehension of the general themes and plot points embedded in the discourse.

Visual Probes Visual probe words were chosen to manipulate semantic relatedness to the local discourse context of the podcast. A total of 172 visual probes were used in the ten clips of the podcast. Probe words were counterbalanced such that each lexical item served as a related probe at one point in the podcast and as an unrelated probe at another point (86 per condition). In terms of lexical features, the mean word length of the visual probes was 5.27 ± 1.48 letters, the mean log word frequency was 3.50 ± 0.65, and mean concreteness was 3.75 ± 1.05 based on established norms for English (Brysbaert & New, 2009; Brysbaert, Warriner, & Kuperman, 2014). Visual probe words were presented simultaneously with the auditory podcast, appearing 300 ms after the onset of the corresponding spoken auditory prime utterance.

Probe-Context Distance To quantify the semantic distance between each visual probe and the preceding spoken discourse context, we used GloVe word embeddings (Pennington, Socher, & Manning, 2014). GloVe represents words as vectors derived from word co-occurrence statistics in large corpora. Although these static embeddings assign each word a single context-invariant representation, semantically similar words tend to cluster together in vector space. Accordingly, GloVe-based distance measures have been shown to predict semantic priming in lexical decision and naming tasks (Auguste, Rey, & Favre, 2017), as well as N400 amplitude (Michaelov, Bardolph, Van Petten, Bergen, & Coulson, 2024).

For each probe, the relevant discourse context was defined as the content words from the clause immediately preceding probe onset. For example, when the visual probe word *dinner* appeared during the clause "a salt and pepper shaker sat on a kitchen table," the words *salt*, *pepper*, *shaker*, *kitchen*, and *table* were used to quantify its related context. In the unrelated context, where *dinner* appeared after the clause "some eye glasses were folded on the nightstand," the words *eye*, *glasses*, *folded*, and *nightstand* were used instead. Context-word embeddings were averaged to create a semantic representation of the podcast prime (Michaelov et al., 2024), and cosine distance was computed between this context vector and the probe-word embedding (Mikolov, Chen, Corrado, & Dean, 2013). Smaller distances indicated greater probe-context relatedness, whereas larger distances indicated lower relatedness.

Experimental Procedure

Each participant listened to all ten one-minute podcast clips in a single EEG session. During listening, participants were instructed to fixate on a cross in the center of a computer screen. Periodically, a visually-presented word replaced the fixation cross for 300 ms.

EEG Acquisition and Preprocessing EEG was recorded in an electromagnetically shielded and sound attenuated chamber using a Biosemi ActiveTwo acquisition system. 32 active scalp electrodes were placed at standard international 10-20 sites. Two additional electrodes were placed at the left and right mastoid processes to serve as reference electrodes during recording; two were placed next to the corner of each eye to detect horizontal eye movements; and one was placed under the left eye to detect blinks. The EEG signal was digitized at 2048 Hz. EEG data were preprocessed using EEGLAB, with additional steps completed in ERPLAB. Prior to any analysis, all raw EEG data were bandpass filtered between 0.1 Hz and 40 Hz and then decomposed using independent component analysis (ICA) via the "runica" algorithm in EEGLAB (Jung et al., 2000). Components were visually inspected and removed if they clearly reflected ocular artifacts (e.g., blinks or horizontal eye movements). Remaining artifacts were further rejected on a trial-by-trial basis using the sliding window peak-to-peak function in ERPLAB, followed by visual inspection of individual trials.

Cross-modal Priming of Visual Probe Words

To examine whether semantic information from one modality (spoken words in a podcast) impacted how participants understood words presented in another (visual probe words), we tested whether the N400 elicited by probes was larger when it was preceded by semantically unrelated terms in the spoken podcast discourse relative to related ones. Specifically, we used the semantic distance between each probe and

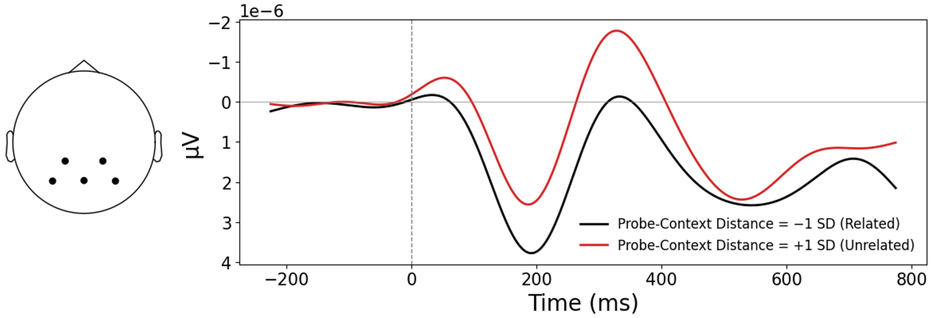

Figure 1: rERPs. Predicted voltages averaged from centro-parietal sites (CP1/CP2/Pz/P3/P4) for probes whose semantic distance from the preceding podcast context (viz., probe-context distance) is -1 standard deviation (SD), corresponding to related probes (black), and +1 SD, corresponding to unrelated probes (red). Negative voltage is plotted upwards per convention in the ERP language literature.

its podcast context to predict single-trial N400 amplitude via regression-based (rERP) methods (Troyer, Urbach, & Kutas, 2020).

Single-trial N400 analysis

Analysis utilized a Linear Mixed Effects (LME) regression model. Our outcome variable was single trial N400 amplitude operationalized by mean amplitude measurements between 300-500 ms post-visual probe onset taken from a centro-parietal electrode cluster (CP1,CP2, Pz, P3, P4). The fixed effect was probe-context distance and the model included random intercept terms for subject, clip, and electrode. A significant negative effect of probe-context distance was observed (Estimate: -0.68, confidence interval (CI): [-0.87, -0.50], $p < 0.001$), such that probe words that were less semantically related to the preceding podcast context (i.e., farther away in semantic space) tended to elicit a larger and more negative N400. This finding suggests the spoken words in the podcast provide semantic information that subsequently modulates the neural response to upcoming visual words.

rERP waveforms

To visualize the cross-modal priming effect, we used the EEG data described above to compute the regression ERP (rERP) for visual probe words. Targeting the same centro-parietal ROI (CP1, CP2, Pz, P3, P4) from the LME analysis, probe distance was used to predict voltage at each time point, beginning 200 ms before probe onset until 800 ms after. Model-predicted rERPs were derived by adding the estimated average of the probe-locked intercept term to the semantic distance estimate at each time point. Figure 1 plots rERPs for probe words that were close to their podcast primes in z-scored measures of semantic space (viz., probe distance = -1 standard deviation from the mean) in black versus those that were more distant (probe distance = 1 standard deviation from the mean) in red. Consistent with the LME, words with greater probe-context distances were more negative during the N400 time window.

Semantic Tracking and Priming Effect

We then proceeded to temporal response function (TRF) analysis to examine the relationship between individual discourse-level semantic tracking and the cross-modal semantic priming effect. Two computational metrics were used to index semantic tracking performance: lexical surprisal and contextual semantic dissimilarity. Lexical surprisal quantifies the conditional probability of a word given its preceding context, whereas contextual semantic dissimilarity captures how much a word's meaning diverges from the evolving discourse representation. Collectively, these two measures allow us to probe semantic information processing at both lexical and discourse levels during continuous speech comprehension.

Feature Extraction

Surprisal of Words in the Podcast Surprisal denotes the log-transformed conditional probability of a word given its preceding context. When operationalized this way, surprisal estimates based on word co-occurrence statistics have been shown to predict the amplitude of the N400 elicited by words, with less probable words eliciting larger (more negative) N400 responses (Frank, Otten, Galli, & Vigliocco, 2015b). More recently, surprisal estimates derived from large language models have proven especially effective for modeling single-trial N400 amplitude for words embedded in extended text and discourse contexts (Michaelov et al., 2024). Accordingly, we used the Qwen3-8B language model (Team, 2025) via the Hugging Face platform (Wolf et al., 2020) to compute the lexical surprisal of each spoken word in the podcast, based on a 10-word rolling context window. Here, Qwen3-8B was selected because the Qwen3 family provides both open-source autoregressive and embedding models, supporting a consistent feature-extraction pipeline across surprisal and semantic dissimilarity measures. The average surprisal across clips was 7.05 (SD = 6.15).

Table 1: Linear Mixed Effect Regression Models and AICs

LME Models	Normalized AIC
N400 amp $\sim$ Surprisal Decoding * Probe Distance + RV	-7
N400 amp $\sim$ Semantic Decoding * Probe Distance + RV	-12
N400 amp $\sim$ Probe-Context Distance $\times$ (Surprisal Decoding + Semantic Decoding) + RV	**-20**

RV: Random variables

Table 2: Linear Mixed Effects Regression Model Results

N400 Amplitude $\sim$ Probe-Context Distance $\times$ (Surprisal Decoding + Semantic Decoding) + RV			
Predictors	Estimates	CI	p-value
(Intercept)	0.11	[-1.15, 1.36]	0.87
Probe Distance	0.03	[-0.16, 0.22]	0.76
Surprisal Decoding	0.14	[-0.06, 0.35]	0.18
Semantic Decoding	-0.28	[-0.50, -0.07]	**0.011**
Probe-Context Distance $\times$ Surprisal Decoding	-0.30	[-0.49, -0.12]	**0.002**
Probe-Context Distance $\times$ Semantic Decoding	-0.32	[-0.51, -0.13]	**0.001**

RV: random variables

Contextual Semantic Dissimilarity of Words in the Podcast In a seminal study of semantic processing in discourse contexts, Broderick and colleagues found that TRF models of each word's semantic dissimilarity to its context yielded a neural response that resembled the N400 and was sensitive to participants' ability to interpret the eliciting speech (Broderick et al., 2018). While they employed static embeddings to compute semantic dissimilarity, we surmised that contextualized embeddings derived from a large language model (LLM) might perform even better as an index of the semantic processing load experienced by humans during podcast listening (Goldstein, Ham, Schain, Honey, & Hasson, 2025). Similar to our surprisal estimates, we used the Qwen3-8B embedding model (Zhang et al., 2025) via the Hugging Face platform (Wolf et al., 2020) to compute the semantic dissimilarity of each word in the podcast relative to its contextual meaning. Contextual semantic dissimilarity was defined as the cosine distance between the embedding vector representing the preceding 10-word context and the embedding vector for the target word. Both vectors were obtained by querying the Qwen3 embedding model, and cosine distance was computed using the SciPy package in Python (Virtanen et al., 2020). Larger distances indicate greater semantic dissimilarity between the meaning of the preceding context and the current word. The average semantic dissimilarity across clips was 0.54 (SD = 0.09)

Stimulus Stream After feature extraction, each semantic measure was organized into a 1000Hz stream, with values annotated using the precise timing of each corresponding spoken word provided by Whisper (Radford et al., 2022). Each measure was annotated from the onset to the offset of the corresponding spoken word in the clip. Prior to the decoding analysis, both stimulus streams were downsampled to 128Hz to align with the EEG data.

Decoder Model Training

The TRF approach was used to model the relationship between EEG data and each of the semantic stimulus features described above (Crosse et al., 2016). Specifically, a backward TRF model (hereafter referred to as the decoder model) was implemented to predict semantic features from the EEG signal. For each time point 200-700 ms post-word onset (an interval chosen to capture the N400), EEG data from all 28 channels were used to construct a linear regression model predicting either the lexical surprisal or contextual semantic dissimilarity values. Thus, each decoder model characterizes the relationship between a stimulus feature and the temporally associated EEG signal.

Prior to model training, continuous EEG and stimulus signals were segmented into 10-second trials. Decoder models were trained separately for each trial. To prevent data leakage, the target trial was held out as a test set and was not included during training. Model parameters were optimized using a leave-one-out cross-validation procedure across the remaining training trials, such that each training trial served as a test trial once. Ridge regression parameters were selected by averaging across models from the training trials, and the resulting decoder was applied to the held-out test trial. Decoder performance was quantified as the Pearson correlation between the predicted and true stimulus feature time series for the test trial (this correlation coefficient is hereafter referred to as the surprisal decoding score and semantic dissimilarity decoding score). Separate decoding models were constructed for each subject and each clip. All TRF analyses were done using the multivariate temporal response function (mTRF) MATLAB toolbox (Crosse et al., 2016).

Statistical Significance of the Decoder Models To verify the statistical significance of our decoding models, we established a chance-level baseline as a null model. The null model

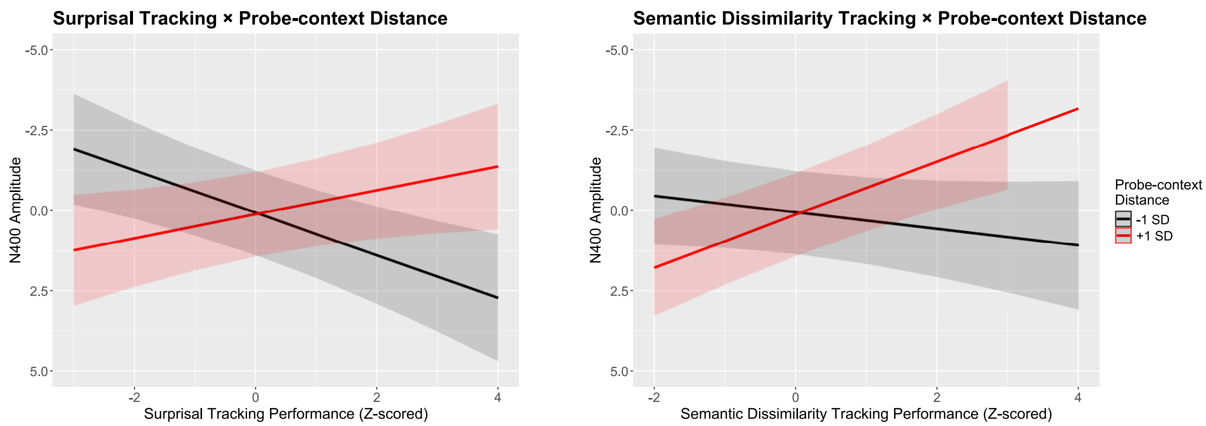

Figure 2: Interactions between semantic tracking performance and probe-context distance on predicting N400 amplitude. Left: Interaction between surprisal tracking performance and probe-context distance measure. Right: Interaction between semantic dissimilarity tracking performance and probe-context distance measure. Stronger semantic tracking is associated with a larger modulation of N400 amplitude by relatedness, particularly for contextual semantic distance.

was computed by constructing unrelated stimulus features and training the decoder using the identical procedure as in the true model. Specifically, to disrupt the stimulus–EEG correspondence while preserving the within-clip structure of the data, we randomly shuffled the stimulus feature values within each clip prior to model training. Decoding performance obtained from this null model was then compared against decoding performance from the true (unshuffled) model using paired t-tests, confirming that the true-model decoding scores were significantly higher than chance level. Under this procedure, the averaged surprisal decoding scores and semantic dissimilarity decoding scores were 0.14 ± 0.005 and 0.18 ± 0.006, respectively. The paired t-test confirmed that they were both higher than chance level ($p < 0.001$).

Statistical Model Comparison The decoding scores from each testing trial for each audio clip and participant were entered into a series of increasingly complex Linear Mixed Effects (LME) models to examine whether semantic tracking performance modulates the cross-modal priming effect at the individual level. The outcome variable in all models was the mean N400 amplitude elicited by the visual probe, measured between 300-500 ms post-probe onset from a centro-parietal electrode cluster (CP1, CP2, Pz, P3, P4). All models shared the same random-effects structure, including random intercepts for clips and subjects. Fixed effects included surprisal decoding score, semantic dissimilarity decoding score, and probe-context distance. Probe-context distance was included to model the cross-modal semantic relatedness effect directly and to test whether its magnitude is modulated by individual differences in semantic tracking performance. The regression model with the lowest Akaike information criterion (AIC) was selected as the best-fitting model (Akaike, 1973).

Results

The best-fitting model for predicting the cross-modal priming effect (i.e., the magnitude of the N400 component) was the full model including scores for both surprisal decoding and semantic dissimilarity decoding (see Table 1 for AIC comparison and Table 2 for model output). In this model, semantic dissimilarity decoding score showed a significant main effect on the cross-modal priming effect, whereas neither probe-context distance nor surprisal decoding score showed a significant main effect (see Table 2). Rather, significant interactions were observed between probe-context distance and both surprisal decoding score and semantic dissimilarity decoding score (see Table 2), indicating that the strength of the N400 priming effect varies as a function of individual differences in semantic tracking performance (see Figure 2). Specifically, stronger neural tracking of semantic dissimilarity, and to a lesser extent surprisal, was associated with a larger modulation of N400 amplitude by probe-context distance. These results suggest that discourse-level semantic representations, as indexed by neural semantic tracking measures, shape how strongly probe-context distance modulates the retrieval process during discourse comprehension.

To rule out the possibility that the absence of the cross-modal priming effect reflects general data quality (i.e., noisier EEG producing both weaker N400s and poorer "semantic tracking" scores), we ran a control decoding analysis using an acoustic feature, pitch, that should be sensitive to signal quality but is theoretically orthogonal to semantic integration. Accordingly, we extracted continuous pitch (F0) from the audio using Parselmouth (Jadoul, de Boer, & Ravignani, 2024) and computed a pitch TRF decoding score using the same preprocessing and decoding pipeline as for the semantic models.

However, the pitch decoding was not a significant predictor of N400 to visual probes (Estimate: -0.12, CI: [-0.39, 0.14], p = 0.36). Moreover, when the pitch decoding score was added as a covariate to the model show in table 2, the semantic decoding scores remained predictive of the cross-modal priming effect. Taken together, these results suggest that the association between semantic tracking and N400 amplitude is unlikely to be a byproduct of noisy recordings.

Discussion

In the present study, we asked whether the fidelity of discourse-level semantic representations formed during naturalistic podcast listening influences the amplitude of the N400 elicited by visual word probes that varied in their relatedness to the spoken discourse. To address this question, we employed a cross-modal semantic priming paradigm in which visual words were intermittently presented while participants listened to a podcast. Our initial analyses examined whether the podcasts elicited cross-modal priming effects, indexed by the N400 evoked by the probes. When contextual relatedness was quantified using a continuous measure of semantic distance, we found that greater probe–context distance was associated with larger N400 amplitudes.

Next, we constructed regression models to decode metrics related to the meaning of words in the podcast from the concurrently recorded EEG data. We were able to use participants' EEG to successfully decode both the surprisal and the semantic dissimilarity of each word in the podcast. We then asked whether the magnitude of cross-modal priming effects on the probe words were related to our decoding scores for the podcasts. Results revealed that participants with better decoding scores for the podcast also exhibited greater sensitivity to probe-context distance on the visual probes. Taken together, these findings suggest our measures of semantic tracking for speech in the podcast and the N400 to visual probes are both sensitive indices of participants' understanding of the ongoing discourse.

Cross-modal Semantic Priming

Our first key finding is that cross-modal visual probes that were less related to the ongoing podcast context produced larger centro-parietal N400 than related probes. This pattern suggests that semantic information extracted from speech in the podcast can rapidly influence the processing of visual words, consistent with the view that semantic representations are not strictly modality-bound but can interact across different modalities during comprehension (Holcomb & Anderson, 1993; Holcomb, Anderson, & Grainger, 2005). In our paradigm, as listeners follow the podcast narrative, they incrementally build a discourse-level situation model that constrains the interpretation of new information (van Berkum, Brown, Hagoort, & Zwitserlood, 2003; Hald, Steenbeek-Planting, & Hagoort, 2007). When a visual probe appears, it activates its own semantic representation that will interact with the currently active discourse context. Probes that are more compatible with the discourse model benefit from contextual facilitation (reduced N400), whereas probes that are weakly supported require greater semantic retrieval effort, yielding an enhanced N400. Overall, results suggest discourse-derived semantic context is available online during naturalistic speech processing and can rapidly affect semantic processing of input beyond the speech stream.

Semantic Tracking and Cross-modal Priming Effect

The second key finding was that the cross-modal relatedness effect on the N400 scaled with listeners' semantic tracking performance: participants with better semantic tracking of the podcast showed a stronger N400 sensitivity to the probe–context semantic distance. One straightforward interpretation is that this interaction reflects online engagement with the discourse. TRF-based semantic tracking has been shown to depend strongly on whether listeners are successfully comprehending the speech (Broderick et al., 2018; Broderick, Di Liberto, Anderson, Rofes, & Lalor, 2021), consistent with the suggestion that these measures index higher-level semantic processing during naturalistic listening. Further, classic ERP work shows that reduced attention and competing task demands attenuate N400 context effects, implying these effects rely on the availability of attentional resources to build and maintain the relevant semantic representation (Hohlfeld, Martín-Loeches, & Sommer, 2015).

A related, more representational-level interpretation is that semantic tracking performance in any given audio clip indexes the quality of the situation model at the moment the probe appears. In our paradigm, listeners who better track semantic information in the podcast may effectively maintain a more coherent discourse model, which in turn yields greater facilitation for related probes and a larger mismatch cost for unrelated ones. From this perspective, the presence of a cross-modal N400 priming effect primarily among better semantic "trackers" suggests that discourse-level context only exerts a strong influence on the probe when listeners are actively representing the semantic content of the podcast.

One limitation of the present study is that our computational measures captured context within relatively local linguistic windows. Probe-context distance was computed from the clause immediately preceding the visual probe, and both surprisal and contextual semantic dissimilarity were estimated using a 10-word rolling context. Thus, although the stimuli were embedded in a naturalistic podcast narrative, these predictors may reflect local sentential or near-discourse context more than broader discourse-level representations. Future work should compare alternative operationalizations of context, including longer windows, sentence- or paragraph-level embeddings, and models sensitive to narrative structure, to determine whether semantic tracking and N400 priming reflect local contextual fit, broader discourse-level meaning, or both.

In sum, our findings suggest that TRF-based semantic tracking reflects discourse-level semantic representations that also govern context-sensitive N400 priming during naturalistic speech comprehension.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the 2nd international symposium on information theory* (pp. 267–281). Akadémiai Kiadó. doi: 10.1007/978-1-4612-1694-0$_1$5

Auguste, J., Rey, A., & Favre, B. (2017). Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd workshop on evaluating vector space representations for nlp* (pp. 21–26).

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3. doi: 10.1016/j.cub.2018.01.080

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., & Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Scientific Reports*, *11*, 4963. doi: 10.1038/s41598-021-84597-9

Brysbaert, M., & New, B. (2009). Moving beyond kucera 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*(4), 977–990. doi: 10.3758/BRM.41.4.977

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. doi: 10.3758/s13428-013-0403-5

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604. doi: 10.3389/fnhum.2016.00604

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015a). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. doi: 10.1016/j.bandl.2014.10.006

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015b). The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1-11. doi: https://doi.org/10.1016/j.bandl.2014.10.006

Glass, I. (2001). *House on loon lake.* Audio podcast episode, No. 199, *This American Life*. Retrieved from `https://www.thisamericanlife.org/199/house-on-loon-lake` (NPR)

Goldstein, A., Ham, E., Schain, M., Honey, C. J., & Hasson, U. (2025). Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, *16*, 10529. doi: 10.1038/s41467-025-65518-0

Hald, L. A., Steenbeek-Planting, E. G., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. evidence from the n400. *Brain Research*, *1146*, 210-218. Retrieved from `https://www.sciencedirect.com/science/article/pii/S000689` (Mysteries of Meaning) doi: https://doi.org/10.1016/j.brainres.2007.02.054

Hohlfeld, A., Martín-Loeches, M., & Sommer, W. (2015). Is semantic processing during sentence reading autonomous or controlled? evidence from the N400 component in a dual task paradigm. *Advances in Cognitive Psychology*, *11*(2), 42–55. doi: 10.5709/acp-0170-2

Holcomb, P. J., Anderson, J., & Grainger, J. (2005). An electrophysiological study of cross-modal repetition priming. *Psychophysiology*, *42*(5), 493–507. doi: 10.1111/j.1469-8986.2005.00348.x

Holcomb, P. J., & Anderson, J. E. (1993). Cross-modal semantic priming: A time-course analysis using event-related brain potentials. *Language and Cognitive Processes*, *8*(4), 379–411. doi: 10.1080/01690969308407583

Jadoul, Y., de Boer, B., & Ravignani, A. (2024). Parselmouth for bioacoustics: automated acoustic analysis in Python. *Bioacoustics*, *33*(1), 1–19. doi: 10.1080/09524622.2023.2259327

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, *62*, 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*(2), 99–116. doi: 10.1016/0301-0511(80)90046-0

Levari, T., & Snedeker, J. (2024). Understanding words in context: A naturalistic EEG study of children's lexical processing. *Journal of Memory and Language*, *137*, 104512. doi: 10.1016/j.jml.2024.104512

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*. doi: 10.1162/nol_a_00105

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. doi: 10.1162/jocn.2006.18.7.1098

Otten, M., & Van Berkum, J. J. (2007). What makes a discourse constraining? comparing the effects of discourse message and scenario fit on the discourse-dependent n400 effect. *Brain Research*, *1153*, 166-177. Retrieved from `https://www.sciencedirect.com/science/article/pii/S000689` doi: https://doi.org/10.1016/j.brainres.2007.03.058

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1162/ doi: 10.3115/v1/D14-1162

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.* Retrieved from https://arxiv.org/abs/2212.04356

Team, Q. (2025). *Qwen3 technical report.* Retrieved from https://arxiv.org/abs/2505.09388

Troyer, M., Urbach, T. P., & Kutas, M. (2020). Lumos!: Electrophysiological tracking of (wizarding) world knowledge use during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(3), 476.

van Berkum, J. J. A., Brown, C. M., Hagoort, P., & Zwitserlood, P. (2003). Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension. *Psychophysiology*, *40*(2), 235–248. doi: 10.1111/1469-8986.00025

van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the N400. *Journal of Cognitive Neuroscience*, *11*(6), 657–671. doi: 10.1162/089892999563724

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. doi: 10.1016/j.ijpsycho.2011.09.015

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Brew, J. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., … Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.