Exploring Neural Correlates of Predictability in Natural Face-to-Face Conversation

Oussama Silem[1,2] (jo_silem@esi.dz), Maïwenn Fleig[3,4,5] (maiwenn-annie.fleig@univ-amu.fr), Philippe Blache[4,5] (blache@ilcb.fr), Auriane Boudin[4,5] (auriane.boudin@univ-amu.fr), Houda Oufaida[1] (h_oufaida@esi.dz), Leonor Becerra[3,5] (leonor.becerra@univ-amu.fr)

Inria, Paris, France[1], Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria[2],
Aix Marseille Univ, CNRS, LIS, Marseille, France[3], Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France[4],
Institute of Language,Communication and the Brain, France[5]

Abstract

Prediction is central in human language processing, as the brain continuously predicts upcoming words using prior knowledge and context. Surprisal theory quantifies predictability using word surprisal. While previous studies link neural activity to surprisal during passive listening or reading, we investigate how surprisal is tracked in dynamic face-to-face conversations. Two key challenges arise: estimating surprisal as well as identifying predictions in EEG data in natural conversation. We address the first challenge by adapting a pre-trained large language model to a dataset of spontaneous conversation capturing features like hesitations and repetitions. We then relate the surprisal estimated by the adapted model to EEG data using temporal response functions. Our experimental results show neural tracking of surprisal at different time lags after word onset, supporting the surprisal theory in face-to-face conversation. To the best of our knowledge, we are the first to address the application of surprisal theory in such interactive settings.

Keywords: Surprisal, EEG, Temporal Response Function, Natural Conversations

Introduction

During a spoken conversation, the human brain processes what other persons are saying while simultaneously preparing a response, effortlessly enabling a reply in less than half a second once the speaker finishes. The predictive coding theory (Rao & Ballard, 1999; Friston, 2005) offers an explanation for this ability, which requires real-time language understanding to keep up with the rapid pace of speech -two to three words per second- (Ryskin & Nieuwland, 2023). This framework views human thought and behavior as a result of constant predictions carried out by the brain. From this perspective, during a conversation, a person constantly makes predictions about what their interlocutor will say next, drawing from both the context and their prior knowledge encoded in an internal generative model (Wang et al., 2023).

The surprisal theory (Hale, 2001; Levy, 2008) is a computational theory of language processing that provides a strong formalization of predictability (Pickering & Gambi, 2018). This theory posits that the cognitive effort required to process a word is proportional to its surprisal (Shannon, 1948), a measure of a word's predictability given its context (calculated as the negative logarithm of the conditional probability). The surprisal theory is supported by different studies that demonstrated the existence of a correlation between a word's surprisal and language processing difficulty (Oh, Clark, & Schuler, 2022) estimated by behavioral measures such as reading time (Smith & Levy, 2013; Kliegl, Nuthmann, & Engbert, 2006; McDonald & Shillcock, 2003; Monsalve, Frank, & Vigliocco, 2012) or neural measures, such as neural responses to individual words (Frank & Willems, 2017; Kutas & Hillyard, 1984; Frank, Otten, Galli, & Vigliocco, 2015).

However, most studies on surprisal theory have focused on passive tasks, where participants were engaged in activities that primarily involved language understanding. This includes reading sessions (Haller, Bolliger, & Jäger, 2024; Dambacher, Kliegl, Hofmann, & Jacobs, 2006) or listening to narratives (Willems, Frank, Nijhof, Hagoort, & Van den Bosch, 2016; Heilbron, Ehinger, Hagoort, & De Lange, 2019), with very few studies exploring the plausibility of this theory in more interactive situations, such as during a spontaneous conversation where both language understanding and production are involved. Additionally, as the surprisal and predictive coding theories are grounded on predictability, a major question arises: How can we efficiently estimate the probability of different linguistic features in a way that aligns with human language processing? While earlier works relied on probabilities estimated through experimental procedures (e.g. Cloze Task (Taylor, 1953)) or by leveraging advancements in NLP like Recurrent Neural Networks (Goodkind & Bicknell, 2018; Bhattasali & Resnik, 2021), recent studies have entirely shifted toward the use of Large Language Models (LLMs) (Michaelov, Bardolph, Van Petten, Bergen, & Coulson, 2024; Oh & Schuler, 2023). These models, with their billions of parameters trained on extensive corpora, demonstrate remarkable language comprehension and generation capabilities (Minaee et al., 2024), which has motivated their application in studying phenomena related to human language and its neural basis (Michaelov et al., 2024; Oh & Schuler, 2023).

Yet, LLMs are predominantly trained on written data (e.g. text scraped from Wikipedia and social media) which limits their ability to fully capture the unique

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

characteristics of natural speech. Spoken conversations are marked by distinctive phenomena such as turn taking, hesitations, filled pauses, repetitions, and other disfluencies, which are either poorly represented or entirely absent in the training corpora of most LLMs. Therefore, to effectively study language prediction in interactive contexts, such as natural spoken conversations, LLMs must be adapted to these conversational settings. This adaptation would allow for a more accurate exploration of prediction phenomena within natural dialogue.

In contrast to other studies that rely on passive tasks such as listening and reading, we present a study focused on the brain correlates of prediction mechanisms in a more interactive context: **face-to-face spoken conversation**. We developed a new conversational model by adapting a pre-trained large language model to spoken interactions, which we then leveraged to investigate the neural correlates of two linguistic features: *Surprisal* and *Word Onset*. Our results suggest evidence in support of the surprisal theory in interactive settings, reflected as an early and late neural tracking of word surprisal.

Related Work

Brain correlates have been studied using different approaches. In Brennan and Hale (2019), the authors employed multiple regression to determine which model of linguistic expectations most accurately represents EEG brain activity. Further, brain correlates of linguistic phenomena can be studied using a technique known as Temporal Response Functions (TRFs) (Ding & Simon, 2012). The core idea is to identify the relationship between linguistic predictors and EEG (or MEG) signals (Crosse, Di Liberto, Bednar, & Lalor, 2016; Brodbeck & Simon, 2020). In this approach, a linguistic predictor is represented as a continuous signal, and a corresponding response function is learned. This response function, when combined with the predictor, generates a prediction of the brain signal. The function is learned by training a model that maps the input (predictor) to the output (brain signal).

Early studies using TRFs revealed a connection between the speech spectral envelope and brain oscillatory dynamics across different frequency bands. For instance, the rhythm of the acoustic signal has been found to influence the rhythm of brain activity (Lalor & Foxe, 2010; Wong et al., 2018; Brodbeck et al., 2023). More recent research has focused on the influence of various linguistic predictors, ranging from low-level (e.g. word onset, position, and frequency) to higher-level features (e.g. surprisal, precision, and similarity). Among high-level linguistic features, the surprisal has long been a key tool in psycholinguistics (Hale, 2001) for assessing the difficulty of integrating a word into its context, as reflected in neural activity. Several studies have shown that cortical activity tracks word surprisal (Weissbart, Kandylaki, & Reichenbach, 2020; Chalehchaleh, Winchester, & Di Liberto, 2025; Broderick, Di Liberto, Anderson, Rofes, & Lalor, 2021). These studies have revealed several key effects. Notably, when surprisal is high, a response is observed around 450 ms, primarily in EEG channels over the temporal and occipital regions of the left hemisphere (Weissbart et al., 2020). This effect is present across multiple frequency bands (delta, beta, gamma) and is thought to reflect the brain's predictive mechanisms. Statistical significance is determined by comparing the observed results to a baseline function that shuffles surprisal values (Weissbart et al., 2020; Chalehchaleh et al., 2025). Several studies have also explored estimating word similarity within context to gauge the amount of new linguistic information introduced by a word (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018; Heilbron, Armeni, Schoffelen, Hagoort, & De Lange, 2022; Gillis, Vanthornhout, Simon, Francart, & Brodbeck, 2021). Gillis et al. (2021) found that different linguistic representations (including surprisal) explain neural responses beyond acoustic responses to speech, while Broderick et al. (2021) showed different neural correlates of word surprisal between younger and older persons.

Data

We used the SMYLE corpus[1], an audio-video and neurophysiological corpus in French (Boudin et al., 2023). It contains 16 hours of recordings, with 30 pairs of participants engaged in dyadic 1) face-to-face storytelling followed by 2) a free conversation task. The storytelling consists of three successive narratives: #1 retelling a video clip (the pear story (Watson-Gegeo, 1981)), #2 narrate the pitch of a movie/book/video game and #3 narrate favorite vacation. The storytelling task comprises two listening conditions: "*attentive*" or a "*distracted*" listener (between-subject design). In normal condition, listeners were instructed to carefully listen to the storytelling and freely react. In the distracted condition, listeners received the supplementary instruction to count all words produced by the storyteller, which starts with the sound /*t*/, without the storyteller discovering this hidden task. This study uses a subset of 25 dyads (50 participants) of SMYLE (12.92 hours), and focus exclusively on the storytelling task (6.6 hours), where one participant assumes the role of a listener, resulting in minimal noise in the EEG data. The corpus includes enriched orthographic transcription (Blache et al., 2017), segmented into Inter-Pausal Units[2] (IPUs). These transcriptions include information such as laughter, repetitions, disfluencies, broken words, elisions. The transcriptions were subsequently standardized and segmented into tokens, aligned with the audio signal using

[1] `https://www.ortolang.fr/market/item/smyle`

[2] a segment of speech delimited by pauses of 200 ms

the SPPAS software[3] (Bigi, 2012, 2015).

Quantifying Linguistic Features

In our study, we are interested in the neural correlates of two linguistic features: **Surprisal**, a high-level feature which we estimated using an LLM fine-tuned on spoken conversational data, and **Word Onset**, a low-level feature marker of word boundaries.

Data Preparation for Fine-tuning

For fine-tuning the language model, we used the transcriptions from both the storytelling and free conversation tasks. Standardized IPUs were used, preserving all present disfluencies, such as repeated or truncated words. We aimed to maintain the data as natural as possible to enable the model to capture a wide range of characteristics inherent in human spontaneous conversation. To form a conversational turn, we grouped one participant's IPUs until a production from the other participant occurred.

The model was fine-tuned using samples consisting of 10 consecutive turns separated by the '<p>' marker. The training dataset was constructed using a sliding window approach (stride = 10) and further augmented with a modified version of Weighted Random Sampling where additional training examples were constructed by taking 10 turns starting from a randomly sampled turn in the dataset. The samples were then shuffled to ensure random ordering. Twenty-one dyads were used for training (555,458 token), 2 for cross-validation (8055 tokens), and 2 for testing (7388 token).

Fine-tuning Conversational LLM

To estimate word probabilities in our experiment, we used a French version of GPT-2 adapted for natural conversation. Specifically, we used the base version of GPT-fr (Simoulin & Crabbé, 2021), which contains 1.3 billion parameters, as the backbone model and fine-tuned it using LoRA (Hu et al., 2022) on the conversational dataset constructed from the SMYLE corpus. The LoRA modules were applied to all the layers and across every module of the backbone model. The fine-tuning lasted for 5 epochs, using the AdamW optimizer with the following hyperparameter settings: LoRA rank = 32, $\alpha = 32$, learning rate = 2×10^{-4} with a linear warmup phase of 500 steps, batch size = 8, dropout ratio = 0.05, and gradient clipping with a norm of 1.

Construction of Linguistic Features

To calculate the conditional probability for each word in the transcription of each conversation, we used a sliding window approach with a window size of 5 and a stride of 1. Each window was constructed by concatenating 5 turns from the same speaker. The window was then passed through the fine-tuned language model to calculate the probabilities for each word in the window. We then retained the probabilities of the words in the last turn of the window, resulting in an estimation of the probabilities for the words of a turn using a context consisting of the previous 4 turns from the same speaker.

Since TRF models work with continuous signals, we followed the approach from Weissbart et al. (2020) to construct a continuous surprisal signal for each dyad from the word probabilities. First, the surprisals for different words were calculated using the equation below:

$$\text{surprisal}(w_i) = -\log\left(P(w_i|w_1...w_{i-1})\right)$$

We then created a continuous signal of the same duration as the conversation, initialized with zeros. Spikes were added at word onset timestamps (provided with the SMYLE corpus), with amplitudes representing the word's surprisal estimated previously, resulting in a continuous surprisal signal. Additional cropping was then performed to align the signal with the EEG data, with an 11-second offset at the beginning of the surprisal signal and a 1-second offset at the end for both the EEG data and the surprisal signal.

Since the word surprisal spikes were placed at word boundaries, and to ensure that the observed neural responses reflect the information conveyed by word surprisal rather than the timing of word boundaries, we considered another linguistic feature -Word Onset- for a more controlled interpretation of the results. We followed the same approach to construct a continuous word onset signal, but set the amplitude of the spikes at the word onset timestamps to 1.

Neural Encoding Model

In this section, we detail the pre-processing of EEG data recorded during spontaneous conversation and the development of models to predict EEG responses based on a linguistic feature signal.

EEG Data Pre-processing

The electrophysiological signal in SMYLE was acquired with two 64-channel Biosemi active electrodes, arranged according to the 10/20 positioning system and configured for signal acquisition at a frequency of 2048 Hz. For the listening participants, we identified the segment from the first task, segmented it with a 1-second offset and compared the length of the resulting segment to the duration reported in the SMYLE corpus. Dyads with mismatched durations were excluded from the analysis. To denoise the EEG signal, a band-pass filter with a low-cut threshold value of 0.3 Hz and a high-cut threshold value of 62 Hz was applied to the EEG data. The raw data were processed using a band-pass FIR filter (Hamming window, cutoff -6 dB at 0.15 Hz for the lower threshold and -6 dB at 69.75 Hz for the upper threshold),

[3] https://sppas.org/ (V.4.10)

preserving frequencies within the range [0.3, 62] Hz to retain brain activity across all frequency bands. This filter effectively removed low-frequency noise (e.g. electrode wire movement) and high-frequency noise (e.g., muscle contractions). A notch filter (band-stop FIR filter, Hamming window, cutoff -6 dB at 49.12 Hz and 50.88 Hz) was then applied to eliminate power line noise from electrical interference around 50 Hz. To reduce computation time in the following steps of the experiment, the EEG data was down-sampled to 256 Hz. All channels were subsequently re-referenced to the channel average. Artifacts were removed from the EEG data using Independent Component Analysis (ICA) implemented in the MNE library (Gramfort et al., 2013). Principal Component Analysis (PCA) was first applied to the filtered data to retain components explaining 98% of variance, followed by decomposition into ICs using the FastICA algorithm (Hyvarinen, 1999). Artifact ICs were identified through visual inspection of scalp topographies for eye-related artifacts (e.g. eye blinks and movements), followed by the MNE function `find_bads_muscle` for muscle-related artifacts, with a threshold of 0.7 . Identified artifact ICs were zeroed out and the EEG data were reconstructed through inverse transformation, resulting in cleaned signals. Due to issues with signal synchronization, EEG data quality, and ICA decomposition, we excluded 8 dyads from the analysis, leaving a total of 17 dyads (average length = 896.47 s (14.94 min), standard deviation = 449.55 s (7.49 min)).

TRF Model Training

We used a linear encoding model to assess the neural response to the surprisal. Specifically, we trained a model to predict the EEG recording of the brain activity from the listener, using as input the surprisal signal (or word onset signal) constructed from the speech of the speaker. Our model consisted of Temporal Response Functions (TRFs) (Ding & Simon, 2012), which are linear kernels that allow for a precise temporal description of the neural response to a linguistic feature at different time lags.

The TRF describes the EEG signal $r(t,n)$ from channel $n \in [1,...,64]$ at timestamps t as a weighted sum of the continuous surprisal signal $s(t)$ over time lags $\tau \in [\tau_{min}..\tau_{max}]$:

$$r(t,n) = \sum_{\tau} w(\tau,n)s(t-\tau) + \epsilon(t,n) \quad (1)$$

Where $w(\tau,n)$ are the TRF weights that indicate how a change in surprisal at time lag τ affects the EEG signal after τ milliseconds. We used the TRF model implementation from the Python library mTRFpy[4]. For each dyad, we divided the EEG and surprisal signals into training (first 90% of the signals) and testing (last 10% of the signals) segments. The time lag window was set between $\tau_{min} = -400$ ms and $\tau_{max} = 1000$ ms, resulting in 358 timestamps (signal frequency = 256 Hz). The TRF weights were estimated by minimizing the mean squared error (MSE) between the observed EEG signal $r(t,n)$ and the predicted signal $\hat{r}(t,n)$. To prevent overfitting, we trained the TRF model for each dyad using ridge regression (Tikhonov, 1977), which introduces a regularization term λ in the estimation of the TRF weights matrix W. The optimal λ was selected through k-fold cross-validation ($k = 5$) where we tested 100 values of λ on a logarithmic scale (between 0.1 and 10^7). Model performance was evaluated using the Pearson correlation metric between the predicted and actual EEG signals ($r(t,n)$ and $\hat{r}(t,n)$). The same approach was followed to train and evaluate the TRF models for the word-onset feature.

Statistical Significance

We used a statistical hypothesis test method to determine if the estimated TRF models were statistically significant. Specifically, we compared the TRF models derived from the actual EEG and surprisal/word onset signals to TRF models trained using the same EEG signals paired with shuffled surprisal/word onset signals.

For the word onset, the conveyed information corresponds to the moment when a word begins to be distinguished as a unique entity -Word Boundary-. Therefore, to construct a shuffled model for the word onset, we disrupted this information by permuting the temporal markers, thereby eliminating the temporal reference points of the words. Following the same reasoning, the conveyed information in the surprisal signal is the surprisal of the different words, represented with spikes at word onset. We constructed a shuffled surprisal signal by keeping the spikes at word onset, as in the original signal, but with shuffled surprisal values, resulting in spikes at the same timestamps but with random amplitudes.

For each dyad, 100 random models, i.e. models predicting the neural response based on a shuffled surprisal/word onset signal, were trained following the same approach as for the actual model and evaluated on the same test data, producing a distribution of the evaluation scores under the null hypothesis. The p-value was computed as the probability of obtaining a score better than that of the actual model. The estimated TRF model was then considered statistically significant if p-value < 0.05, otherwise, the model was not statistically significant.

Results

For each dyad, we trained two separate TRF models, one for each linguistic feature (surprisal and word onset). We then performed hypothesis testing on the resulting models using the approach described previously. Of the

[4] `https://github.com/powerfulbean/mTRFpy`

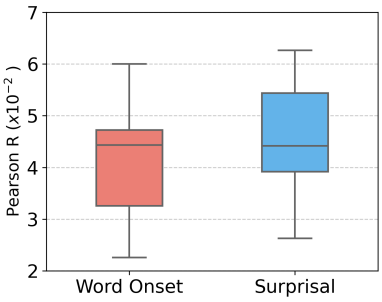

Figure 1: Box plots of EEG prediction accuracy for each predictor for significant dyads (middle bar: median, upper and lower bars: third and first quartiles, upper and lower whiskers: maximum and minimum, p-value < 0.05)

17 dyads, only eight resulted in statistically significant models (p-value < 0.05).

We began the analysis by comparing the ability of the models in reconstructing the EEG recordings from the predictor. This was carried out by evaluating their performance on unseen data using the Pearson correlation score. The results in Table 1 indicate that surprisal is a better predictor of brain activity than word onset, as it achieves higher accuracy. This observation is further supported by the plot in Figure 1, which shows that the improved prediction accuracy is highly consistent across the different dyads.

Predictor	Word Onset	Surprisal
R	0.041	0.049

Table 1: EEG prediction accuracy evaluated with Pearson correlation score (mean, p-value < 0.05).

To further assess the neural response to the different features, we visualized the TRF weights of the models for the surprisal as well as the word onset, averaged over the eight significant dyads. Figure 2 reports the resulting weights for each feature, along with those of TRF models trained using shuffled surprisal values, which serve as a reference. The results show that the TRF weights from the reference models were random and exhibited no clear neural response to surprisal at any time lag. In contrast, TRF weights from models incorporating surprisal and word onset signals revealed distinct temporal patterns (see Figure 2), with strong neural responses to surprisal and word onset in two temporal windows: Around 200 ms and around 800 ms. We observed a strong neural response to surprisal in an early interval spanning 50 to 400 ms, peaking near 300 ms, and a later interval between 700 to 1000 ms, with a maximal response around 900 ms. A similar pattern was found for the TRF weights derived from word onsets, although the neural response appeared to occur earlier than for surprisal. Specifically,

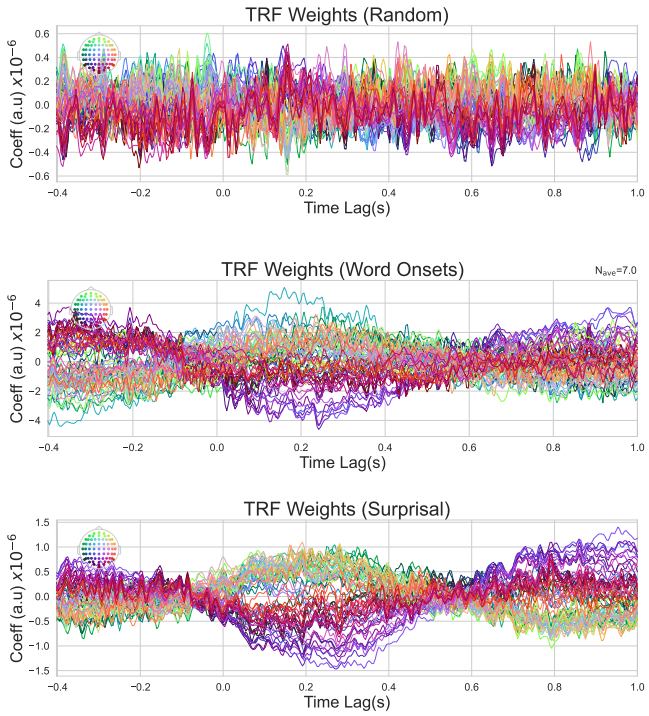

Figure 2: TRF Weights averaged over significant dyads for randomized signal (i.e. shuffled surprisal values), word onset signal and surprisal signal (p-value < 0.05).

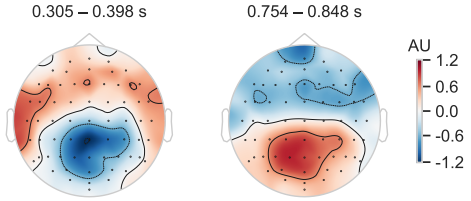

Figure 3: Topographic plots from the TRF models trained using the surprisal signal for significant dyads (p-value < 0.05).

the early response ranged from 0 to 350 ms, with peaks at approximately 150 ms and 250 ms, while the later interval spanned 650 to 1000 ms, with lower amplitude than the earlier interval and peaking around 900 ms. Additionally, we observed polarity differences in TRF weights across scalp regions for both features (see Figure 3). The left-frontal regions exhibited a strong positive wave of TRF weights in the first interval, followed by a pronounced negative wave in the second interval. In contrast, the right-posterior regions showed the opposite pattern, with a negative polarity in the first interval followed by a positive polarity in the second interval.

Discussion

The analysis of the TRF weights revealed the existence of neural tracking for both surprisal and word onset

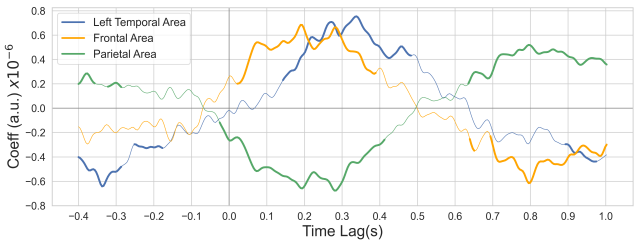

Figure 4: Averaged TRF weights using surprisal signal for significant dyads per scalp area with significant weights highlighted in bold.

during spontaneous conversation in two time windows. This cortical tracking effect has been observed in previous studies using multiple regression (Brennan & Hale, 2019) and TRFs (Heilbron et al., 2022; Weissbart et al., 2020; Weissbart & Martin, 2024; Gillis et al., 2021). Our findings align with the literature, extending these effects from passive listening and reading tasks to the more interactive context of spontaneous conversations.

The results revealed similar patterns for the neural response to the two linguistic features, although word onset appeared to elicit an earlier response. The observed effects for word onset likely result from neural entrainment to the speech envelope, which has been hypothesized to arise from a superposition of discrete onset-related brain responses (Ding & Simon, 2014). Our findings are consistent with prior research (Brodbeck, Hong, & Simon, 2018; Broderick et al., 2018; Weissbart et al., 2020; Gillis et al., 2021), showing an amplitude increase around 50 ms after word onset, peaking at approximately 150 ms. As noted in Gillis et al. (2021), we also observed early frontotemporal activation. This initial effect likely reflects low-level processing of the acoustic signal, particularly word segmentation based on acoustic and phonotactic cues. It may correspond to early stages of word recognition before higher-level linguistic processing occurs. A second, weaker effect appears between 650 and 1000 ms, which we interpret as the integration of the word into the broader context.

For surprisal, although the response patterns resemble those of word onset potentially suggesting that the observed tracking could result from low-level acoustic features rather than the surprisal itself, we argue that the surprisal effect is not merely a byproduct of word onset. While the neural response to surprisal may be partially linked to word onset (since surprisal values were added at word onsets), the semantic and syntactic information embedded in the surprisal plays a crucial role in the observed response. This hypothesis is supported first by the prediction accuracy results (Figure 1), which showed that surprisal yielded higher accuracy than word onset, suggesting that surprisal captures additional linguistically relevant information that contributes to neural responses during speech processing. Additionally, the temporal patterns of neural responses to surprisal reveal peaks around 350 ms and 800 ms, two time markers associated with semantic and syntactic aspects of language processing. The peak at approximately 350 ms can relate to the N400 component, which is linked to semantic and lexical processing (Friederici, 2002) and which have been shown to reflect word predictability (Kutas & Hillyard, 1984; Dambacher et al., 2006; Frank et al., 2015; Wang et al., 2023). Furthermore, the peak around 350 ms appears to be centered in the left temporal region (see Figure 3 and Figure 4), consistent with findings of Wang et al. (2023), which showed that unexpected words (words that would have a higher surprisal) elicit an increased neural response between 300 and 500 ms in the left temporal lobe. The later response around 800 ms may correspond to the P600 component, which has been linked to syntactic violations or the processing of unexpected words (Kaan, 2007). Unexpected words will tend to have a higher surprisal values leading to an increased neural response. Moreover, our results indicate that this late response is centered in the frontal and parietal lobes (see Figure 3 and Figure 4), further aligning with studies showing that implausible or unexpected words elicit late responses in frontal and posterior regions (Wang et al., 2023). However, in our case, no definitive conclusions can be drawn about the precise localization of neural responses, as higher spatial-resolution techniques such as MEG would be needed for a more detailed analysis.

Overall, while our results suggest support for the surprisal theory, we do not claim to provide definitive proof. Our study has several limitations, mainly the number of significant dyads (8 out of 17), which makes us very cautious in drawing general conclusions about the surprisal theory in interactive contexts. We hypothesize that these limitations may be due to noise in the EEG signal, which we found very challenging to clean. This motivated our choice to focus on the narrative task, which helped reduce noise in the data, as it includes less back-and-forth interaction compared to free conversations. Nevertheless, we still consider the narrative task to be an interactive condition, since the listener produced, on average, one word for every five words spoken by the speaker.

Conclusion

This study aims to advance our understanding of predictive processing in natural interactions, particularly in relation to the surprisal theory. Our results demonstrate cortical tracking of word onset and surprisal at varying time lags, across different brain regions associated with language processing. While further studies are required to confirm our findings, these results provide promising evidence supporting the plausibility of the surprisal theory in the context of face-to-face conversation.

References

Bhattasali, S., & Resnik, P. (2021). Using surprisal and fmri to map the neural bases of broad and local contextual prediction during natural language comprehension. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 3786–3798).

Bigi, B. (2012). Sppas: a tool for the phonetic segmentations of speech. In *Proceedings of the eighth international conference on language resources and evaluation (lrec'12)* (pp. 1748–1755). Istanbul, Turkey.

Bigi, B. (2015). Sppas-multi-lingual approaches to the automatic annotation of speech. *The Phonetician. J. Int. Soc. Phonet. Sci.*, *111*, 54–69.

Blache, P., Bertrand, R., Ferré, G., Pallaud, B., Prévot, L., & Rauzy, S. (2017). The corpus of interactional data: A large multimodal annotated resource. *Handbook of linguistic annotation*, 1323–1356.

Boudin, A., Bertrand, R., Rauzy, S., Houlès, M., Legou, T., Ochs, M., & Blache, P. (2023). Smyle: A new multimodal resource of talk-in-interaction including neuro-physiological signal. In *Companion publication of the 25th international conference on multimodal interaction* (pp. 344–352).

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, *14*(1), e0207741.

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., ... Simon, J. Z. (2023). Eelbrain, a python toolkit for time-continuous analysis with temporal response functions. *Elife*, *12*, e85012.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.

Brodbeck, C., & Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Physiology*, *18*, 25-31.

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.

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*(1), 4963.

Chalehchaleh, A., Winchester, M., & Di Liberto, G. M. (2025). Robust assessment of the cortical encoding of word-level expectations using the temporal response function. *Journal of Neural Engineering*, *22*(1).

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.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain research*, *1084*(1), 89–103.

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, *107*(1), 78–89.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience*, *8*, 311.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11.

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, *6*(2), 78–84.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815–836.

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural markers of speech comprehension: measuring eeg tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*, *41*(50), 10316–10329.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (cmcl 2018)* (pp. 10–18).

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... others (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, *7*, 267.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Haller, P., Bolliger, L. S., & Jäger, L. A. (2024). On language models' cognitive biases in reading time prediction. In *Icml 2024 workshop on llms and cognition.*

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Heilbron, M., Ehinger, B., Hagoort, P., & De Lange, F. P. (2019). Tracking naturalistic linguistic predictions with deep neural language models. *arXiv preprint*

arXiv:1909.04400.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... others (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, *10*(3), 626–634.

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and linguistics compass*, *1*(6), 571–591.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General*, *135*(1), 12.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189-193.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological science*, *14*(6), 648–652.

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*, *5*(1), 107–135.

Minaee, S., et al. (2024, feb). Large language models: A survey. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/2402.06196`

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 398–408).

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, 777963.

Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological bulletin*, *144*(10), 1002.

Rao, R. P., & 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.

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: what is next? *Trends in Cognitive Sciences*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Simoulin, A., & Crabbé, B. (2021). Un modèle transformer génératif pré-entrainé pour le _ français. In *Traitement automatique des langues naturelles* (pp. 246–255).

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415–433.

Tikhonov, A. N. (1977). Solutions of ill-posed problems. *VH Winston and Sons*.

Wang, L., Schoot, L., Brothers, T., Alexander, E., Warnke, L., Kim, M., ... Kuperberg, G. R. (2023). Predictive coding across the left fronto-temporal hierarchy during language comprehension. *Cerebral Cortex*, *33*(8), 4478–4497.

Watson-Gegeo, K. A. (1981). Wallace l. chafe (ed.), the pear stories: Cognitive, cultural, and linguistic aspects of narrative production (advances in discourse processes, vol. iii). norwood, n.j.: Ablex, 1980. pp. 323. *Language in Society*, *10*(3), 451–453. doi: 10.1017/S0047404500008897

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of cognitive neuroscience*, *32*(1), 155–166.

Weissbart, H., & Martin, A. E. (2024). The structure and statistics of language jointly shape cross-frequency neural dynamics during spoken language comprehension. *Nature Communications*, *15*(1), 8850.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral cortex*, *26*(6), 2506–2516.

Wong, D. D., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., & Cheveigné, A. d. (2018). A comparison of temporal response function estimation methods for auditory attention decoding. *Biorxiv*, 281345.