The Predictive Brain: Neural Correlates of Word Expectancy Align with Large Language Model Prediction Probabilities

Nikola Kölbl[1,2], Konstantin Tziridis[1], Andreas Maier[3], Thomas Kinfe[4], Ricardo Chavarriaga[5], Achim Schilling[1,2,4,*], Patrick Krauss[1,2,*]

[1]Neuroscience Lab, University Hospital Erlangen, Germany
[2]CCN Group, Pattern Recognition Lab, FAU Erlangen-Nürnberg, Germany
[3]Pattern Recognition Lab, FAU Erlangen-Nürnberg, Germany
[4]Neuromodulation and Neuroprosthetics, University Hospital Mannheim, University Heidelberg, Germany
[5]ZHAW Zürich, Switzerland
*both authors contributed equally

Abstract

Predictive coding theory suggests that the brain continuously anticipates upcoming words to optimize language processing, but the neural mechanisms remain unclear, particularly in naturalistic speech. Here, we simultaneously recorded EEG and MEG data from 29 participants while they listened to an audio book and assigned predictability scores to nouns using the BERT language model. Our results show that higher predictability is associated with reduced neural responses during word recognition, as reflected in lower N400 amplitudes, and with increased anticipatory activity before word onset. EEG data revealed increased pre-activation in left fronto-temporal regions, while MEG showed a tendency for greater sensorimotor engagement in response to low-predictability words, suggesting a possible motor-related component to linguistic anticipation. These findings provide new evidence that the brain dynamically integrates top-down predictions with bottom-up sensory input to facilitate language comprehension. To our knowledge, this is the first study to demonstrate these effects using naturalistic speech stimuli, bridging computational language models with neurophysiological data. Our findings provide novel insights for cognitive computational neuroscience, advancing the understanding of predictive processing in language and inspiring the development of neuroscience-inspired AI. Future research should explore the role of prediction and sensory precision in shaping neural responses and further refine models of language processing.

Language; Transformer; Large Language Models; MEG; EEG; BERT; N400; predictive coding:

1 Introduction

The human brain is a prediction machine, constantly anticipating upcoming sensory inputs, words, events, and in general future states [FK09; SBF21; Sch+23a]. In language processing, this predictive mechanism enables fast and efficient comprehension by minimizing

surprise [WES24]. When predictions fail, the brain updates its internal model to better match the incoming input, ensuring adaptive and flexible processing [SK24; FK09]. Although predictive coding theories are well established in perception, how this framework applies to language, particularly semantic processing, remains largely unknown [CGK23]. Despite decades of research, the neural mechanisms underlying the extraction and representation of meaning are still not fully understood [Pul13].

Understanding how the brain efficiently processes language is not only a fundamental question in cognitive neuroscience, but could also provide insights for improving artificial intelligence (AI) [Has+17]. In recent years, transformer-based language models such as BERT, Llama and GPT-4o have revolutionized natural language processing (NLP) by using context to predict upcoming or masked words [Tou+23b; Tou+23a; AIM24; Kra+24a; KT19; Wol+20; RSK24; KSK24; Ope22]. These models provide a computational framework that potentially can parallel how the human brain processes linguistic input [KD18; AIK+24; Cos+24; Sch+22]. However, it remains unclear whether and to what extent neural responses in natural language comprehension reflect such statistical predictability. Investigating this relationship could bridge the gap between biological and AI, and shed light on common principles of efficient information processing [Has+17; Che+22; Sto+24].

To unravel the mechanisms of language prediction in the brain, both experimental neuroscience and computational modeling approaches are needed (see e.g. [Sur+23; Sto+22; Sto+23]). Advances in neuroimaging, particularly EEG and MEG, allow the tracking of brain responses to linguistic stimuli with high temporal resolution. In particular, the N400 component - a well-established neural marker of semantic processing - has been linked to predictability and surprise effects [Mic+24; MB22; Fra+13]. However, many studies investigation predictability with neuroimaging techniques use artificial language stimuli such as isolated words, constructions or sentences [GMP17; GBP24], despite the fact that these experiments do not generalize well across different stimulus protocols [Ber17]. Thus, nowadays these artificial stimuli are replaced by natural stimuli and continuous speech such as audio books [Sch+21; KSK23; Koe+24; Köl+24; Gar+22; Sch+23b; Sch+24].

In the present study, we used EEG/MEG measurements of 29 participants stimulated with a German audio book and compared event-related fields (ERF) and event-related potentials (ERP) with predictions of the BERT language model to test two main hypotheses (see [Koe+24; Kra+24b]). First, we hypothesized that when a word is highly predictable within continuous speech, the neural response associated with its processing - particularly in the N400 time window - should show reduced amplitude compared to less predictable words (analogously to [MB22], but using a continuous audio book rather than controlled, visually presented sentences). The second hypothesis states that, if predictive mechanisms shape language comprehension, neural activity reflecting anticipation should be more pronounced

for highly predictable words even before their onset (analogously to [GTP21]).

We examined the relationship between BERT-based predictability scores and neural activity, focusing on both pre-word onset signals and N400 responses. Our results showed that as word predictability increased and consequently surprisal was reduced, N400 amplitudes decreased, in line with predictive coding theories. Moreover, highly predictable words indeed elicited stronger pre-activation of neural activity, indicating that the brain actively engages in anticipatory processing when predictability is high — a finding consistent with previous studies on predictive coding in language comprehension [GMP17; PG20; MB22].

2 Methods

2.1 Data

As a natural language stimulus, we presented participants with approximately 50 minutes of the science fiction audio book Vakuum by Phillip P. Peterson, narrated by Uve Teschner (Argon Hörbuch). The audio book has several story lines, two of which were selected and divided into eight alternating chapters of approximately seven minutes each. To maintain engagement and assess comprehension, participants answered three multiple choice questions after each chapter.

We recorded brain activity of 29 participants (15 ♀, mean age: 22.8 $\pm$ 3 years) while they listened to the audio book. All participants were right-handed, native German speakers with normal hearing, no history of neurological disorders, and no use of substances. Neural responses were captured simultaneously using 248-channel magnetoencephalography (MEG) (Magnes 3600WH, 4D-Neuroimaging) and 64-channel electroencephalography (EEG) (ANT Neuro) with additional electrooculogram (EOG) and electrocardiogram (ECG) (MEG: sampling frequency = 1017.25 Hz, EEG: sampling frequency = 2000.0 Hz). To prevent interference from electronic components, the audio signal was delivered via an air tube from external loudspeakers into the MEG chamber, where it was played through headphones. Volume levels were adjusted individually to ensure optimal intelligibility and comfort. To minimize eye and muscle artifacts, participants were instructed to fixate on a central cross and remain as still as possible while lying down. The study was approved by the Ethics Committee of the University Hospital Erlangen (Approval No: 22-361-2, PK).

2.2 Predictability Scores

To quantify the predictability of each noun in the text - the likelihood of its occurrence in a given context - we used BERT (Bidirectional Encoder Representations from Transformers), a transformer-based language model. BERT processes words bidirectionally, capturing

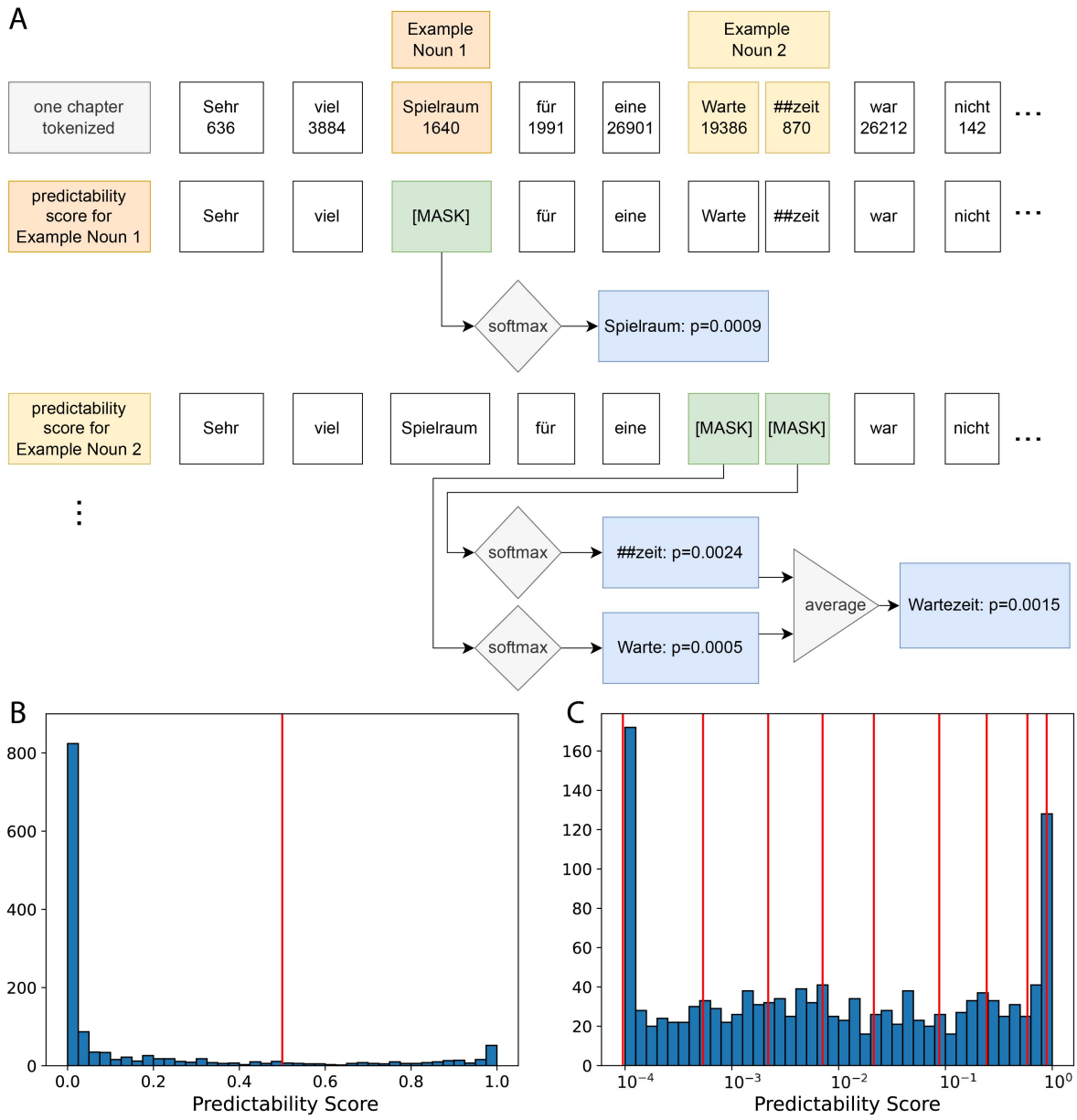

Figure 1: (A): Scheme for computing predictability scores for individual nouns in the audio book. Each target noun (and its corresponding sub-words) is masked ([MASK]), and the resulting BERT output is processed through a softmax function to determine the probability of the original word. For multi-sub-word tokens, the final predictability score is obtained by averaging the probabilities across all sub-words.
(B): Histogram of predictability scores of all nouns (red vertical line: predictability of 0.5 used to divide low and high predictable words). (C): Semi-logarithmic plot of the histogram including thresholds for dividing scores into 10 equal splits used for correlation analysis (red lines).

contextual dependencies across entire sentences [KT19]. For our analysis, we used the pre-trained German model 'bert-base-german-cased' via the Hugging Face Transformers library

[Wol+20]. The text was first pre-processed by removing all punctuation. Next, each word token (along with any associated sub-tokens) was iteratively masked - one at a time - before being passed to BERT for prediction. To ensure that the model remained in inference mode, gradient computations were disabled using PyTorch, as we were not interested in fine-tuning the model [Pas+19]. The modified text containing the masked token was then fed into BERT, and the model's output logits were normalized using a softmax function to obtain probability values between 0 and 1 for each word (see Fig. 1 A for process diagram). For words that were split into multiple sub-tokens, the final predictability score was calculated as the average probability across all sub-word-components (Fig. 1 B). This process was applied to all the words in the audio book, after which we extracted predictability scores only for nouns.

The distribution of predictability scores for all nouns in the audio book shows a clear imbalance, with the majority of nouns having lower predictability. To facilitate the analysis, the predictability scores were binned into two categories: low predictability (scores < 0.5) and high predictability (scores > 0.5, see Fig. 1 B, threshold in red). A significant class imbalance was observed, with 1,182 trials in the low predictability category compared to only 194 trials in the high predictability group. To ensure a balanced comparison, a subset of low-predictability trials was randomly selected to match the number of high-predictability trials, thus creating an equal distribution across conditions. For further analysis, we applied a semi-logarithmic scaling approach, correlating neural responses with the logarithm of predictability scores (Fig. 1 C). This transformation reverses the softmax operation of predictability scores and allows for a more refined assessment of the relationship between linguistic predictability (probabilities) and brain activity. To explore potential systematic patterns in brain activity correlated with predictability scores while maintaining consistent signal-to-noise ratios across categories, we divided the scores into 10 equal-sized bins, each containing approximately 137 trials (Fig. 1 C).

2.3 Data Preparation

To improve the quality of the EEG and MEG data, we applied a standard pre-processing pipeline [Fer+22]. Data processing was performed using MNE-Python (v1.8.0), starting with the identification and interpolation of bad sensors and electrodes, defined as those with flat or excessively noisy signals using Maxwell filtering [Gra+13]. We then applied a 1-20 Hz bandpass filter to remove irrelevant frequency components. For computational efficiency, the data was downsampled to 200 Hz before performing independent component analysis (ICA) for artifact rejection. To remove artifacts related to eye movements and cardiac activity, we excluded the first two independent components (ICs) with the highest variance,

along with any additional ICs that correlated with simultaneously recorded electrooculogram (EOG) or electrocardiogram (ECG) signals [KSK23].

To segment the continuous MEG signal, we employed Forced Alignment using *Web-MAUS* [Sch15], aligning the audio files with their corresponding transcripts to extract precise word onset timestamps from the audio book. The audio signal was simultaneously recorded on a separate stimulus channel during playback, allowing for precise synchronization. Using these word onset markers, we segmented MEG and EEG data into epochs spanning from -1.0 to +2.0 s relative to word onset, with baseline correction applied from -1.0 to 0.0 s. Since our primary focus was on the differential processing of highly-predictable vs. low-predictable nouns, we used the Natural Language Processing (NLP) software spaCy (model: de_core_news_sm) to classify words in the audio book by their part-of-speech tags [Hon+20]. We analyzed ERPs and ERFs for nouns and identified peak responses in the time window from -1.0 s until 2.0 s and extracted the topographic distribution at the time point of the strongest negative responses (Figure 2 A and B). For EEG data, we selected parietal electrodes: CP2, CPz, CP1, P2, Pz, P1. For MEG data, we focused on left frontal sensors: A229, A212, A178, A154, A126, A230, A213, A179, A155, A127, A177, A153, A125. We divided the nouns first into low and high predictable and then into the ten bins mentioned in Chapter 2.2 and calculated corresponding ERPs and ERFs.

2.4 Statistical Tests

To identify time windows in which low- and high-predictable nouns elicited significantly different neural responses, we performed a paired Wilcoxon signed-rank test with 5,000 randomizations with Brainstorm [Pan+05; Tad+11]. A false discovery rate (FDR) correction was applied across signal, time and frequency dimensions to control for multiple comparisons.

For calculating correlations we extracted the mean activities in the significant time intervals over the selected channels and sensors for each of the ten binned nouns ERPs/ERFs. We then fit a linear regression on these values using sklearn.linear_model and calculated the p-values and r^2-values using the libraries numpy and scipy to quantify the relationship between predictability and brain activity [Ped+11; Har+20; Vir+20].

2.5 Source Reconstruction

For source reconstruction, we used the open source software Brainstorm [Tad+11] with the standard ICBM 152 brain anatomy from the MNI database to generate a head model [Fon+09]. For MEG data we used the overlapping spheres method to compute a cortical surface head model, while for EEG data we used a boundary element model (BEM) with

OpenMEEG [HML99; Gra+10]. The noise covariance was estimated from a one minute silent baseline recorded immediately before the audio book was played. For source estimation, we used minimum norm imaging with the sLORETA method and constraint dipole orientations [Pas+02]. Source reconstructions were first computed for participant-level averages of low- and high-predictability noun ERPs and ERFs, followed by a grand average across subjects.

3 Results

3.1 ERF/ERP Analysis

To investigate the effect of predictability on neural processing, we compared brain responses to low (predictability score < 0.5) and high (predictability score > 0.5) predictable nouns (Fig. 2 A, B: ERPs and ERFs for all nouns (high and low predictability); Fig. 2 C, D: ERF/ERP comparison between high/low predictable nouns). Cluster-based permutation statistics were used to identify significant time windows while controlling for multiple comparisons (Fig. 2 C, D) [Pan+05]. In the EEG data, differences emerged 300-450 ms after word onset, consistent with the well-established N400 component typically associated with semantic processing (Fig. 2 C). In the MEG data, significant effects were observed both before word onset (-350 to -250 ms) and after word onset (500-650 ms).

To confirm that these effects were not due to chance, we analyzed two additional low predictable data subsets for both EEG and MEG. The EEG second subset showed a trend toward significance from 320 to 520 ms ($p = 0.11$), while the third subset exhibited a significant effect from 310 to 460 ms ($p < 0.05$). Both MEG subsets revealed consistent significant differences: subset 2 from -330 to -270 ms, 520 to 700 ms, and 820 to 920 ms; subset 3 from -310 to -250 ms, 590 to 740 ms, and 820 to 960 ms. These matching patterns across subsets suggest the presence of consistent and reproducible effects.

While the early effect suggests an anticipatory process in response to less predictable words, the later effect may reflect N400-related activity, similar to the EEG findings (Fig. 2 C). In both data sets, high-predictability nouns elicited lower amplitude responses than low-predictability nouns, supporting the idea that when word expectancy is high, the brain expends less effort during processing [MB22]. These results are consistent with the hypothesis that the brain is more engaged when encountering less predictable words, likely reflecting increased processing demands during lexical access and integration.

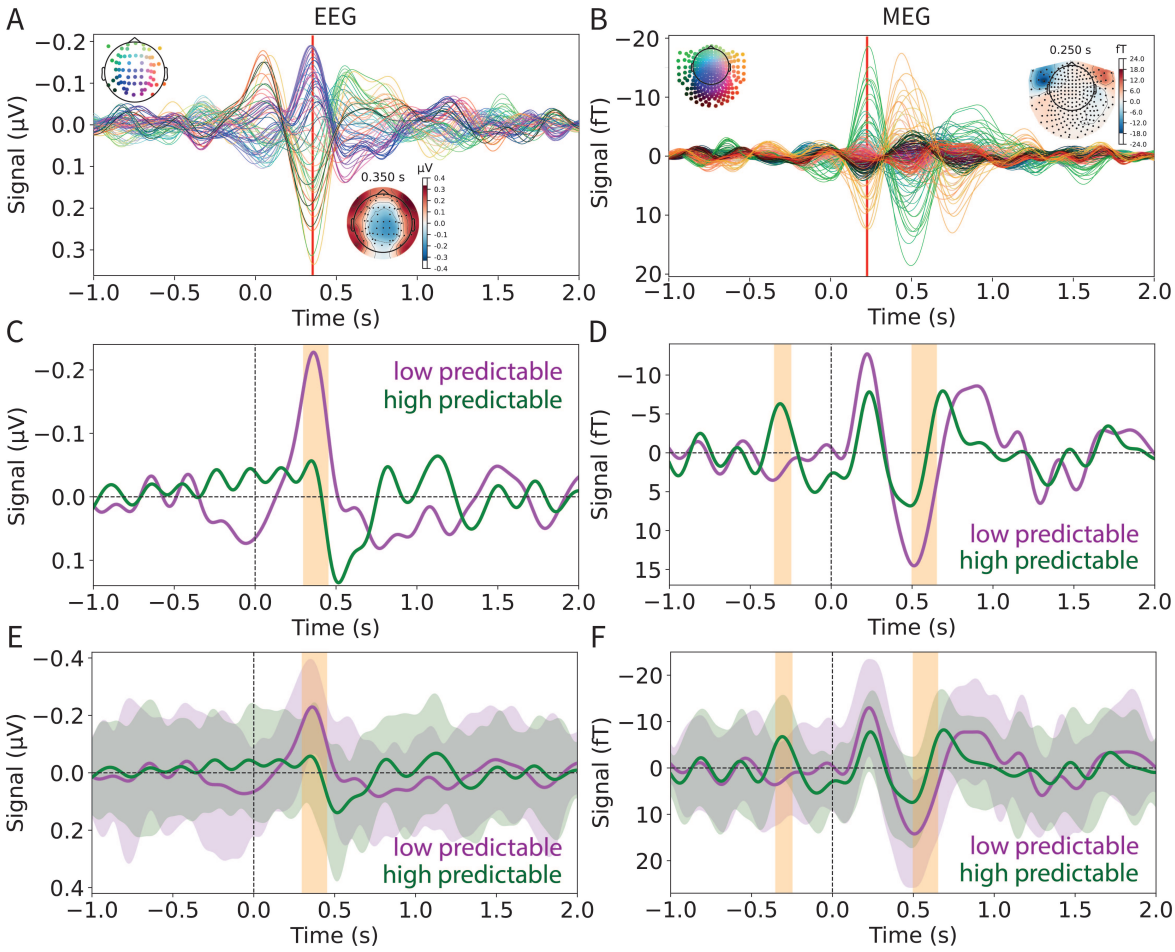

Figure 2: A: Grand average nouns ERP (EEG) including topographic map at peak 350 ms after word onsets. (B): Grand average nouns ERF (MEG) with topographic map at 250 ms after word onsets. C, D: Comparison of ERPs and ERFs of high (green) and low (purple) predictable nouns; ERPs averaged across parietal channels (CP2, CPz, CP1, P2, Pz, P1) and ERFs averaged across left frontal channels (A229, A212, A178, A154, A126, A230, A213, A179, A155, A127, A177, A153, A125). Orange background colors highlight significant p-values ($p<0.05$; FDR corrected, permutation paired test statistic: Wilcoxon signed-rank test). E, F: same as C, D but with variance.

3.2 Source Space Analysis

To specify the brain regions involved in semantic processing, we examined the source reconstruction of event-related fields (ERFs, Fig. 3 B, D) and event-related potentials (ERPs, Fig. 3 A, C) for low- and high-predictability nouns. Consistent with the observed differences in sensor-level data, source estimates revealed that low-predictability nouns elicited stronger neural responses and engaged more extensive cortical regions compared to high-predictability nouns after word onset. These effects were evident within the significant time windows identified previously, with increased activation in EEG (300-450 ms after on-

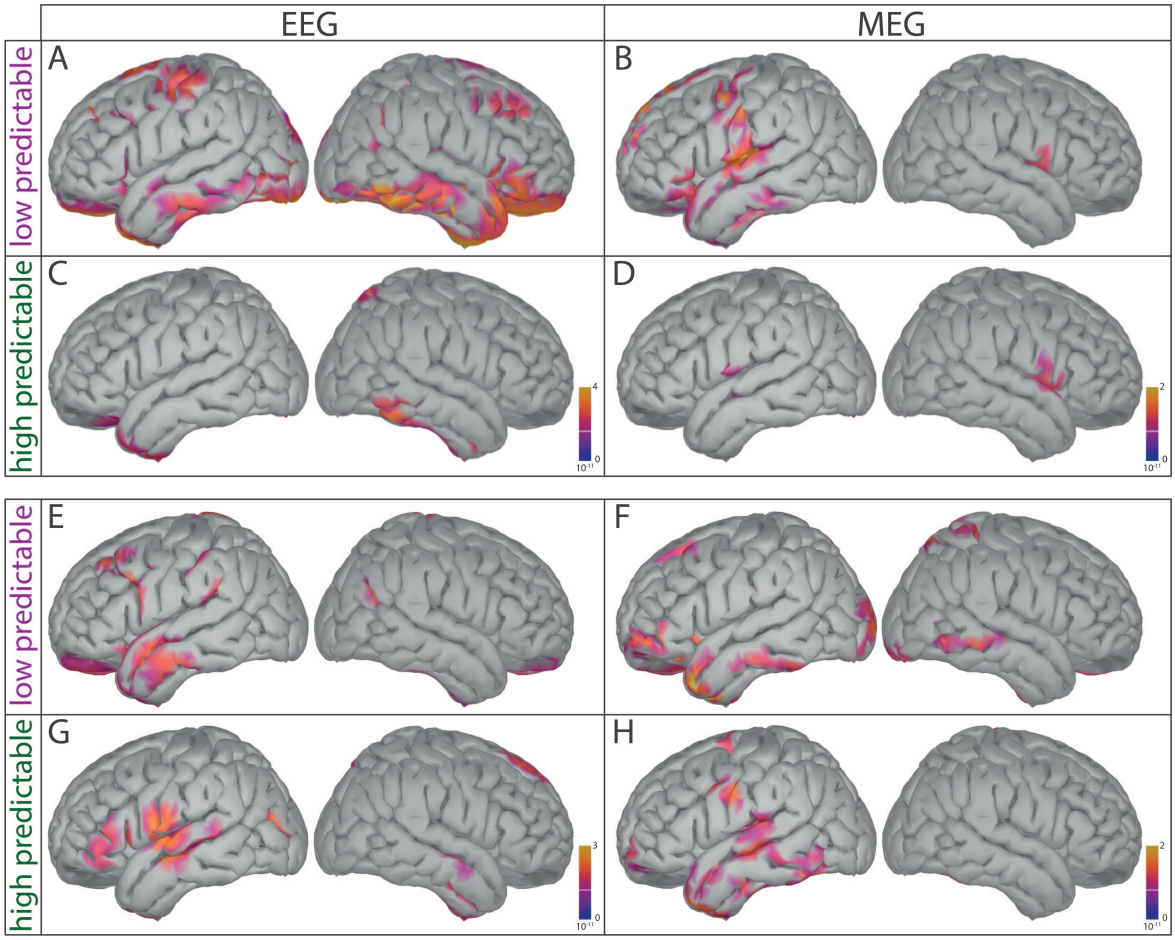

Figure 3: RMS amplitudes in source space for EEG data in the time interval 300 ms-450 ms for low predictable (A) and high predictable nouns (C). RMS amplitudes in source space for MEG data in the time interval 500 ms-650 ms for low predictable (B) and high predictable nouns (D); E-H: Predictive activity before word onset (time intervals: -100 ms-0 ms for EEG and -350 ms until -250 ms for MEG) for EEG (E,G) and MEG (F, H)

set) and MEG (500-650 ms after onset). In particular, significant activity was observed in parietal cortex and sensorimotor regions, suggesting that low-predictability nouns elicit broader cortical engagement, possibly reflecting increased processing demands for semantic integration and motor-related aspects of speech perception [Bon+22; Tia+23; Pul+05]. The widespread recruitment of these areas supports the idea that predictability modulates the neural effort required for language comprehension, with greater activation occurring when word expectancy is low [MB22]. To further investigate the anticipatory mechanisms involved in predictive language processing, we analyzed source activity prior to word onset (Fig. 3 E-H). Significant differences in activation patterns between low and high predictable nouns were observed within the pre-onset time windows (-100 to 0 ms for EEG and -350 to -250 ms for MEG). In the EEG data, highly predictable nouns were associated with greater

activation in left fronto-temporal regions (Fig. 3 G), showing the expected left-hemispheric lateralization characteristic of language processing. In contrast, slightly different effects were observed in the MEG data (Fig. 3 F, H); however, there was a tendency for increased pre-onset activity in sensorimotor regions for low-predictability nouns (Fig. 3 F), suggesting a possible involvement of motor-related processes in linguistic anticipation (see e.g. [Sin+24]).

3.3 Correlation Analysis

To gain a more detailed understanding of whether the effects described above are binary or continuous, we performed a correlation analysis between predictability scores and neural activity in the "N400" time windows 300 ms-450 ms for EEG and 500 ms-650 ms for MEG after word onset. To ensure the robustness of this relationship, we examined the mean amplitudes within the identified time windows across ten independent splits of the data over parietal channels in EEG and left frontal sensors in MEG (see Fig. 4 A for channel/sensor selection, B and C for exemplary EEG and MEG signals). This approach allowed us to assess how gradual changes in predictability affected neural signal strength. By averaging the predictability scores within each split, a clear trend emerged: higher predictability was consistently associated with lower neural response amplitudes (Fig. 4 D, G). Regression analysis revealed highly significant correlations in both EEG ($p = 0.0006$, $r^2 = 0.79$) and MEG ($p = 0.0013$, $r^2 = 0.75$) data, demonstrating a systematic and robust relationship between predictability and neural response amplitude (see Fig. 4 D, G). The fact that these correlations are highly significant further supports the idea that brain activity scales continuously with linguistic predictability, rather than operating in a binary fashion. This systematic reduction in signal amplitude reinforces the hypothesis that the brain expends more effort processing less predictable words. Building on these findings, we next investigated whether the reduction in neural activity during word processing could be driven by predictive mechanisms that occur prior to word onset. To test this, we analyzed pre-onset neural activity in both EEG and MEG (Fig. 4 E, H). For the MEG data, we focused on the previously identified significant time window (-350 to -250 ms before word onset), while in the EEG data, where no significant pre-onset effects were observed, we examined a pre-word interval at -100 ms to assess early anticipatory processes. Analysis revealed a significant relationship between pre-onset neural activity and predictability, although the effect was less pronounced than during word processing. Both EEG ($p = 0.028$, $r^2 = 0.47$) and MEG ($p = 0.0315$, $r^2 = 0.46$) data showed that higher predictability was associated with larger negativities. These results suggest that predictability modulates neural activity even before word onset, supporting the idea that the brain engages in anticipatory processing when

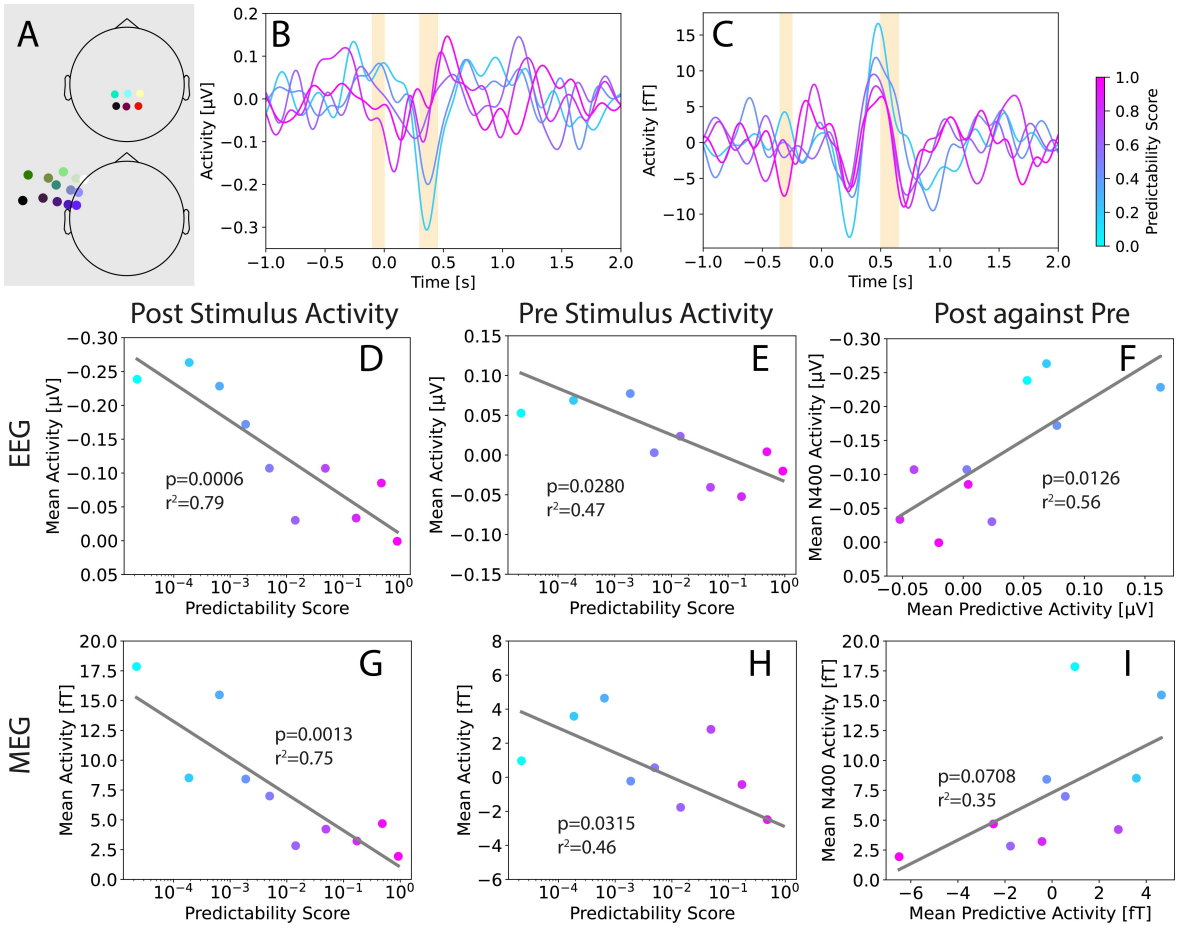

Figure 4: Correlation analysis between predictability scores of BERT and neural activity in ten independent splits. Mean activity over parietal channels (CP2, CPz, CP1, P2, Pz, P1) for EEG (A top) and over left frontal sensors (A229, A212, A178, A154, A126, A230, A213, A179, A155, A127, A177, A153, A125) for MEG (A bottom) of every second split in (B: EEG) and (C: MEG) color coded from cyan (low predictable) to magenta (high predictable). Regression analysis for post stimulus activities show high correlation with predictability scores with p=0.00006, r^2=0.79 for EEG data in time frame 300 ms-450 ms (D) and p=0.0013, r^2=0.75 for MEG data in time frame 500 ms-650 ms (G). Pre stimulus activities also reveal significant correlation for EEG (E, p=0.0280, r^2=0.47 in time frame -100 ms-0 ms) and MEG data (H, p=0.0315, r^2=0.46 in time frame -350 ms to -250 ms). Relation between mean post- and mean pre-stimulus activity show negative correlation (F for EEG data: p=0.0126, I for MEG data: p=0.0708, r^2=0.35).

the linguistic context allows for reliable predictions [GBP24]. This pre-stimulus activity is negatively correlated with the N400 amplitude (Fig. 4 F: (EEG) $p = 0.0126$, $r^2 = 0.56$, I: (MEG) $p = 0.0708$, $r^2 = 0.35$). The negative correlation between N400 amplitude and pre-stimulus activation supports predictive coding, where stronger anticipatory activity reduces processing demands at word onset, leading to a smaller N400 response when predictions

match input (see also [GTP21]).

4 Discussion

This study investigated how word predictability modulates neural activity during natural language processing, focusing on its effects during word recognition and anticipatory processing. EEG and MEG recordings were simultaneously collected from 29 participants listening to an audio book, with predictability scores assigned to nouns using BERT. Neural responses differed significantly between nouns with low and high predictability, with EEG effects in parietal electrodes and MEG effects in left frontal sensors. Detailed analysis confirmed a significant correlation between predictability and neural activity approximately 400 ms after word onset (N400 wave), showing that higher predictability was associated with reduced signal amplitude during word processing. We also found that pre-onset neural activity was associated with predictability, with greater fronto-temporal activation in EEG for highly predictable words. In MEG, low-predictability words showed a tendency for increased sensorimotor activity, suggesting a possible motor-related component to linguistic anticipation. The fact that motor planning and language anticipation / production are closely linked is highlighted by the fact that it is possible to build brain-computer interfaces (BCI) to control a mouse cursor based on an implant in the ventral precentral gyrus, which is related to tongue movement and thus speech production [Sin+24]. A detailed understanding on predictive coding, language understanding, and production can also have some impact on the development of universally applicable BCIs.

Our main finding that the N400 amplitude is inversely correlated with predictability scores, derived from a language model, is consistent with previous research on semantic processing. Goldstein et al. reported increased neural activity for unpredictable words around 400 ms after onset, consistent with the N400 effect [Gol+22]. Similarly, Maess et al. showed that more predictable nouns elicit a smaller N400 response compared to less predictable ones [Mae+16]. Importantly, our study extends these findings by using a combined MEG/EEG setup with more naturalistic linguistic stimuli, demonstrating that the relationship between N400 amplitude and predictability holds even in rich, continuous language contexts. The observed correlation between N400 amplitude and surprisal (negative log conditional probability from transformer networks [Mic+24]) highlights the role of predictability in semantic processing. Recent work has shown that surprisal estimates derived from large language models (LLMs) such as GPT-3 reliably predict N400 effects, strengthening the link between computational models of language prediction and neural responses [Mic+24; Mis+24]. Similar findings have been reported by Kuperberg and coworkers [KBW20], also based on reading and thus visual paradigms, highlighting

the need to investigate whether these effects generalize to naturalistic auditory language, resp. speech processing. Beyond predictability, N400 amplitudes may also be shaped by the precision of predictions and the strength of bottom-up sensory input, reflecting the brain's confidence in a given prediction [Lec+22]. This additional variability in prediction confidence provides an avenue for future research, offering a more nuanced perspective on how linguistic predictions modulate neural processing.

Our second main finding, that higher predictability is associated with greater pre-activation of neural representations prior to word onset, is consistent with previous research on anticipatory language processing. Predictable words have been shown to elicit enhanced pre-onset neural activity, likely reflecting context-driven lexical anticipation [Gol+22; GMP17; GBP24]. In this context, the so-called semantic prediction potential (SPP) has been introduced as new neural marker of predictive processing [GTP21]. We found that the SPP correlates positively with word predictability and negatively with the N400 wave, further linking pre-activation mechanisms to lexical integration. Thus, we could confirm the finding of Grisoni and co-workers, who showed similar correlations using more surrogate stimuli [GTP21]. However, further research is needed to fully understand the relationship between contextual predictability, SPP and N400. These effects are influenced by factors such as the precision (inverse variance, [Lec+22]) of the prediction (prior) and bottom-up sensory input (sensory precision), which also affect the explanatory power of LLM-derived surprise scores as predictors of ERP components [Lec+22; Kri+24]. In particular, the role of prediction precision, which can be interpreted as a confidence score, is supported by evidence that the amplitude of N400 correlates with subjective belief states [Lec+22]. Specifically, studies have shown that N400 responses to semantic violations in human-written texts are stronger than those generated by LLMs [RWC24]. Furthermore, when reading texts generated by LLM, the amplitude of N400 increases when individuals have greater confidence in the competency of the model [RWC24], suggesting that belief in the reliability of the language model may influence the accuracy of the prediction [Lec+22; Sch+23a].

In summary, our study highlights the role of predictive coding in naturalistic language processing, emphasizing the integration of anticipatory and sensory signals. To our knowledge, this is the first study to demonstrate these effects using naturalistic speech stimuli, thereby bridging computational language models with neural measures in real-world contexts. This work advances cognitive computational neuroscience and neuroscience-inspired AI by opening new avenues for exploring the interplay between top-down predictions and bottom-up inputs [KD18; Has+17].

5 Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): grant TZ 100/2-1 (project number 510395418) to KT, grants KR 5148/2-1 (project number 436456810), KR 5148/3-1 (project number 510395418), KR 5148/5-1 (project number 542747151), and GRK 2839 (project number 468527017) to PK, and grant SCHI 1482/3-1 (project number 451810794) to AS. Furthermore, the research leading to these results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programmme (ERC Grant No. 810316) to AM.

6 Author Contributions

AS and PK developed the study protocol. AS, PK, KT, AM supervised the study. NK performed the measurements. NK developed the evaluation programs. All authors discussed the results. AS, PK, NK drafted the first version of the manuscript. AM and TK reviewed the first draft of the manuscript. All authors accept the final version of the manuscript.

References

[AIM24] AI@Meta. "Llama 3.2 Model Card". In: (2024). URL: https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

[AlK+24] Badr AlKhamissi et al. "The LLM Language Network: A Neuroscientific Approach for Identifying Causally Task-Relevant Units". In: *arXiv preprint arXiv:2411.02280* (2024).

[Ber17] Anna M Beres. "Time is of the essence: A review of electroencephalography (EEG) and event-related brain potentials (ERPs) in language research". In: *Applied psychophysiology and biofeedback* 42 (2017), pp. 247–255.

[Bon+22] Camille Bonnet et al. "Kinesthetic motor-imagery training improves performance on lexical-semantic access". In: *PloS one* 17.6 (2022), e0270352.

[CGK23] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. "Evidence of a predictive coding hierarchy in the human brain listening to speech". In: *Nature human behaviour* 7.3 (2023), pp. 430–441.

[Che+22] Yucan Chen et al. "How far is brain-inspired artificial intelligence away from brain?" In: *Frontiers in Neuroscience* 16 (2022), p. 1096737.

[Cos+24] Luiz Costa et al. "LLM-MRI Python module: a brain scanner for LLMs". In: *Simpósio Brasileiro de Banco de Dados (SBBD)*. SBC. 2024, pp. 125–130.

[Fer+22] Oscar Ferrante et al. "FLUX: A pipeline for MEG analysis". In: *NeuroImage* 253 (2022), p. 119047.

[FK09] Karl Friston and Stefan Kiebel. "Predictive coding under the free-energy principle". In: *Philosophical transactions of the Royal Society B: Biological sciences* 364.1521 (2009), pp. 1211–1221.

[Fon+09] Vladimir S Fonov et al. "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood". In: *NeuroImage* 47 (2009), S102.

[Fra+13] Stefan L Frank et al. "Word surprisal predicts N400 amplitude during reading". In: (2013).

[Gar+22] Armine Garibyan et al. "Neural correlates of linguistic collocations during continuous speech perception". In: *Frontiers in Psychology* 13 (2022), p. 1076339.

[GBP24] Luigi Grisoni, Isabella P Boux, and Friedemann Pulvermüller. "Predictive Brain Activity Shows Congruent Semantic Specificity in Language Comprehension and Production". In: *Journal of Neuroscience* 44.12 (2024).

[GMP17] Luigi Grisoni, Tally McCormick Miller, and Friedemann Pulvermüller. "Neural correlates of semantic prediction and resolution in sentence processing". In: *Journal of Neuroscience* 37.18 (2017), pp. 4848–4858.

[Gol+22] Ariel Goldstein et al. "Shared computational principles for language processing in humans and deep language models". In: *Nature neuroscience* 25.3 (2022), pp. 369–380.

[Gra+10] Alexandre Gramfort et al. "OpenMEEG: opensource software for quasistatic bioelectromagnetics". In: *Biomedical engineering online* 9 (2010), pp. 1–20.

[Gra+13] Alexandre Gramfort et al. "MEG and EEG data analysis with MNE-Python". In: *Frontiers in Neuroinformatics* 7 (2013), p. 267.

[GTP21] Luigi Grisoni, Rosario Tomasello, and Friedemann Pulvermüller. "Correlated brain indexes of semantic prediction and prediction error: Brain localization and category specificity". In: *Cerebral Cortex* 31.3 (2021), pp. 1553–1568.

[Har+20] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[Has+17] Demis Hassabis et al. "Neuroscience-inspired artificial intelligence". In: *Neuron* 95.2 (2017), pp. 245–258.

[HML99] MX Huang, John C Mosher, and RM Leahy. "A sensor-weighted overlapping-sphere head model and exhaustive head model comparison for MEG". In: *Physics in Medicine & Biology* 44.2 (1999), p. 423.

[Hon+20] Matthew Honnibal et al. "spaCy: Industrial-strength Natural Language Processing in Python". In: (2020). DOI: 10.5281/zenodo.1212303.

[KBW20] Gina R Kuperberg, Trevor Brothers, and Edward W Wlotko. "A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation". In: *Journal of cognitive neuroscience* 32.1 (2020), pp. 12–35.

[KD18] Nikolaus Kriegeskorte and Pamela K Douglas. "Cognitive computational neuroscience". In: *Nature neuroscience* 21.9 (2018), pp. 1148–1160.

[Koe+24] Nikola Koelbl et al. "Analyzing Differences in Processing Nouns and Verbs in the Human Brain using Combined EEG and MEG Measurements". In: *bioRxiv* (2024), pp. 2024–12.

[Köl+24] Nikola Kölbl et al. "Methodological Considerations in the Analysis of Acoustically Evoked Neural Signals: A Comparative Study of Active EEG, Passive EEG and MEG". In: *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2024, pp. 1–7.

[Kra+24a] Patrick Krauss et al. "Analyzing Narrative Processing in Large Language Models (LLMs): Using GPT4 to test BERT". In: *arXiv preprint arXiv:2405.02024* (2024).

[Kra+24b] Patrick Krauss et al. "Temporal and Hemispheric Dynamics in Neural Processing of Auditory and Speech Stimuli Across Linguistic Complexity: A MEG Source Space Study". In: *bioRxiv* (2024), pp. 2024–12.

[Kri+24] Benedict Krieger et al. "On the limits of LLM Surprisal as functional explanation of ERPs". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46. 2024.

[KSK23] Nikola Koelbl, Achim Schilling, and Patrick Krauss. "Adaptive ica for speech eeg artifact removal". In: *2023 5th International Conference on Bio-engineering for Smart Technologies (BioSMART)*. IEEE. 2023, pp. 1–4.

[KSK24] Hassane Kissane, Achim Schilling, and Patrick Krauss. "Analysis and Visualization of Linguistic Structures in Large Language Models: Neural Representations of Verb-Particle Constructions in BERT". In: *arXiv preprint arXiv:2412.14670* (2024).

[KT19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of naacL-HLT*. Vol. 1. Minneapolis, Minnesota. 2019, p. 2.

[Lec+22] Françoise Lecaignard et al. "Neurocomputational underpinnings of expected surprise". In: *Journal of Neuroscience* 42.3 (2022), pp. 474–486.

[Mae+16] Burkhard Maess et al. "Prediction signatures in the brain: semantic pre-activation during language comprehension". In: *Frontiers in Human Neuroscience* 10 (2016), p. 591.

[MB22] James Michaelov and Ben Bergen. "The more human-like the language model, the more surprisal is the best predictor of N400 amplitude". In: *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*. 2022.

[Mic+24] James A Michaelov et al. "Strong Prediction: Language model surprisal explains multiple N400 effects". In: *Neurobiology of language* 5.1 (2024), pp. 107–135.

[Mis+24] Gavin Mischler et al. "Contextual feature extraction hierarchies converge in large language models and the brain". In: *Nature Machine Intelligence* (2024), pp. 1–11.

[Ope22] TB OpenAI. *Chatgpt: Optimizing language models for dialogue. OpenAI*. 2022.

[Pan+05] Dimitrios Pantazis et al. "A comparison of random field theory and permutation methods for the statistical analysis of MEG data". In: *Neuroimage* 25.2 (2005), pp. 383–394.

[Pas+02] Roberto Domingo Pascual-Marqui et al. "Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details". In: *Methods Find Exp Clin Pharmacol* 24.Suppl D (2002), pp. 5–12.

[Pas+19] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[Ped+11] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[PG20] Friedemann Pulvermüller and Luigi Grisoni. "Semantic prediction in brain and mind". In: *Trends in cognitive sciences* 24.10 (2020), pp. 781–784.

[Pul+05] Friedemann Pulvermüller et al. "Functional links between motor and language systems". In: *European Journal of Neuroscience* 21.3 (2005), pp. 793–797.

[Pul13] Friedemann Pulvermüller. "How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics". In: *Trends in cognitive sciences* 17.9 (2013), pp. 458–470.

[RSK24] Pegah Ramezani, Achim Schilling, and Patrick Krauss. "Analysis of Argument Structure Constructions in the Large Language Model BERT". In: *arXiv preprint arXiv:2408.04270* (2024).

[RWC24] Xiaohui Rao, Hanlin Wu, and Zhenguang Garry Cai. "Comprehending semantic and syntactic anomalies in LLM-versus human-generated texts: An ERP study". In: (2024).

[SBF21] Ryan Smith, Paul Badcock, and Karl J Friston. "Recent advances in the application of predictive coding and active inference models within clinical neuroscience". In: *Psychiatry and Clinical Neurosciences* 75.1 (2021), pp. 3–13.

[Sch+21] Achim Schilling et al. "Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods". In: *Language, Cognition and Neuroscience* 36.2 (2021), pp. 167–186.

[Sch+22] Achim Schilling et al. "Intrinsic noise improves speech recognition in a computational model of the auditory pathway". In: *Frontiers in Neuroscience* 16 (2022), p. 908330.

[Sch+23a] Achim Schilling et al. "Predictive coding and stochastic resonance as fundamental principles of auditory phantom perception". In: *Brain* 146.12 (2023), pp. 4809–4825.

[Sch+23b] Alina Schüller et al. "Attentional modulation of the cortical contribution to the frequency-following response evoked by continuous speech". In: *Journal of Neuroscience* 43.44 (2023), pp. 7429–7440.

[Sch+24] Alina Schüller et al. "The early subcortical response at the fundamental frequency of speech is temporally separated from later cortical contributions". In: *Journal of Cognitive Neuroscience* 36.3 (2024), pp. 475–491.

[Sch15] Florian Schiel. "A statistical model for predicting pronunciation." In: *ICPhS*. 2015.

[Sin+24] Tyler Singer-Clark et al. "Speech motor cortex enables BCI cursor control and click". In: *bioRxiv* (2024), pp. 2024–11.

[SK24] Achim Schilling and Patrick Krauss. *The Bayesian brain: world models and conscious dimensions of auditory phantom perception*. 2024.

[Sto+22] Paul Stoewer et al. "Neural network based successor representations to form cognitive maps of space and language". In: *Scientific Reports* 12.1 (2022), p. 11233.

[Sto+23] Paul Stoewer et al. "Neural network based formation of cognitive maps of semantic spaces and the putative emergence of abstract concepts". In: *Scientific Reports* 13.1 (2023), p. 3644.

[Sto+24] Andreas Stoll et al. "Coincidence detection and integration behavior in spiking neural networks". In: *Cognitive Neurodynamics* 18.4 (2024), pp. 1753–1765.

[Sur+23] Kishore Surendra et al. "Word class representations spontaneously emerge in a deep neural network trained on next word prediction". In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2023, pp. 1481–1486.

[Tad+11] François Tadel et al. "Brainstorm: A user-friendly application for MEG/EEG analysis". In: *Computational intelligence and neuroscience* 2011.1 (2011), p. 879716.

[Tia+23] Lili Tian et al. "Spatiotemporal dynamics of activation in motor and language areas suggest a compensatory role of the motor cortex in second language processing". In: *Neurobiology of Language* 4.1 (2023), pp. 178–197.

[Tou+23a] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[Tou+23b] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[Vir+20] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[WES24] Franz Wurm, Benjamin Ernst, and Marco Steinhauser. "Surprise-minimization as a solution to the structural credit assignment problem". In: *PLOS Computational Biology* 20.5 (2024), e1012175.

[Wol+20] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.