Decoding EEG Signals to Explore Next-Word Predictability in the Human Brain

Boi Mai Quach
mai.quach3@mail.dcu.ie
School of Computing, ML-Labs,
Dublin City University, Ireland

Binh T. Nguyen
ngtbinh@hcmus.edu.vn
Department of Computer Science,
VNUHCM – University of Science

Cathal Gurrin
cathal.gurrin@dcu.ie
School of Computing, ADAPT Centre,
Dublin City University, Ireland

Graham Healy
graham.healy@dcu.ie
School of Computing, ADAPT Centre,
Dublin City University, Ireland

Abstract

Humans invented reading and have passed down this complex skill across generations through language. This study provides empirical evidence of the neural mechanisms underlying bottom-up (related to high-order linguistic structure) and top-down (related to next-word predictability) processes, which interact to guide comprehension during reading. While previous studies have focused on either the N400 effects of predictability or lexical categories, research on how predictability influences N400 responses across different lexical categories is limited, mainly due to constraints in publicly available datasets. Here, we examine how predictability influences brain responses, recorded at millisecond resolution using electroencephalography (EEG), with a focus on the N400 time window (300-500 ms post-stimulus) across different lexical and grammatical categories. Our results indicate that significant differences in N400 responses between high and low cloze probability levels were more pronounced for content words than function words. Among the two primary content categories, verbs exhibited greater N400 differences than nouns, while nouns carried more distinct information about their predictability than verbs. Moreover, we demonstrate that the decoding technique is more effective than the event-related potential (ERP) traditional analysis in capturing more detailed and distinct representations of cognitive processes over time.

Keywords: Cognitive Neuroscience; Reading Comprehension; Machine learning; Electroencephalography (EEG)

Introduction

Reading comprehension involves two interconnected processes: bottom-up and top-down processing (Perfetti & Hart, 2001). While bottom-up processing is often considered an automatic outcome of accurate word recognition (MacDonald, Pearlmutter, & Seidenberg, 1994; Ngabut, 2015), top-down processing uses cues to disambiguate and predict upcoming words (Federmeier, 2007; Kuperberg & Jaeger, 2016). For example, we can infer that the statement "I drink my coffee with cream and ..." is likely to end with "sugar.". Readers access linguistic structures (e.g., lexical groups, grammatical categories) via a bottom-up process to predict that the next word is probably a noun. Simultaneously, they rely on top-down cues to anticipate the upcoming word with its correct meaning before it is fully revealed.

The N400 ERP component is a negative-going deflection peaking around 400 ms post-stimulus word onset, well known for its sensitivity to semantic complexity (Kutas & Federmeier, 2011). Cloze probability (Taylor, 1953), which estimates how likely a particular word is in a given sentence and also reflects the top-down prediction process, has long been considered the primary predictor of N400 amplitude. The graded attenuation of the N400 as a function of cloze probability is one of the most widely replicated phenomena in reading comprehension using EEG, supporting the conclusion that words with higher cloze value elicit a smaller (less negative) N400 response compared to words with lower cloze value (Federmeier & Kutas, 1999; Block & Baldwin, 2010; Michaelov, Coulson, & Bergen, 2022).

Basically, words are grouped into two primary types: content and function groups. Content words, including nouns, verbs, adjectives, and most adverbs, convey specific semantic information about objects and events. Function words, on the other hand, comprise pronouns, determiners, auxiliaries, and adpositions and serve as structural elements in phrase construction. Evidence from ERP recordings during reading processing supports a lexical-based analysis guided by a bottom-up process. Results from pioneering studies indicate that content words elicit a larger N400 amplitude compared to function words (Neville, Mills, & Lawson, 1992). This finding was later supported by observations showing that both word classes could elicit N400 responses, but function words, especially high-frequency ones, show significantly smaller amplitudes (Münte et al., 2001; He, Boudewyn, Kiat, Sagae, & Luck, 2022). Grammatical categories, particularly nouns and verbs, also influence N400 amplitude, with larger amplitudes observed for verbs compared to nouns (Federmeier, Segal, Lombrozo, & Kutas, 2000; Lee & Federmeier, 2006).

There is a lack of evidence in lexical-based analysis using next-word predictability due to the limited availability of publicly accessible EEG datasets for semantic-level information (Deniz, Nunez-Elizalde, Huth, & Gallant, 2019). To gain deeper theoretical insights, this study aims to test two hypotheses: The first hypothesis states that content words elicit a larger N400 difference effect between high and low cloze levels compared to function words. The second hypothesis suggests that the N400 predictability difference between high and low cloze is more pronounced for verbs than nouns.

Moreover, we not only explore the nature of top-down processing and its relationship to bottom-up information processing during reading comprehension, but we also provide evidence that multivariate decoding outperforms traditional ERP analysis in capturing distinct neural information across differ-

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

ent conditions. By leveraging univariate and decoding techniques to investigate our hypotheses, we further demonstrate the advantages of multivariate decoding in this context.

Methodology

EEG Recording and Data Preparation

To examine the N400 response to the predictability across different lexical classes, we utilised the DERCo dataset (Quach, Gurrin, & Healy, 2024), including EEG data recorded from 22 native English speakers while they were reading The Grimm Brothers' Fairy Tales. Two participants "QPF42" and "USQ95" were excluded from data analysis due to excessive eye movements in their data. Additionally, word-by-word cloze values were collected using a cloze procedure via Mechanical Turk crowdsourcing platform. High-density EEG data was recorded using electrode 32-channels placed according to the international 10–20 system (Klem, 1999).

Parts of speech were identified using the Python library IPA [1], which allows us to extract and categories the grammatical function of each word within a sentence. For example, in the sentence "She quickly reviewed some detailed reports with her team," the parts of speech are categorised as follows: She (pronoun), quickly (adverb), reviewed (verb), some (determiner), detailed (adjective), reports (noun), with (adposition), her (pronoun), team (noun). In terms of cloze probability, we utilised the same two distinct groups (high versus low) as in the DERCo dataset. This decision was based on their validation, which indicated significant differences in N400s across most electrodes.

After preprocessing, the number of trials for high and low cloze probability levels in terms of lexical classes varied between participants. To standardise the trial numbers for analysis, we identified the participant with the fewest trials at each cloze probability level and set these as a baseline. We then randomly selected an equivalent number of trials from other participants, ensuring their cloze value distributions closely matched the baseline. This random selection process was guided by the Kolmogorov-Smirnov (KS) test, iterating up to 5000 times to find the subset with the lowest KS statistic, indicating the best distribution match. As mentioned, our analysis focused on binary classifications such as lexical groups (content versus function words) and content word categories (nouns versus verbs). The number of trials used as a baseline for each group per condition is detailed in Table 1.

Univariate Analysis

Traditional univariate analysis involves calculating a difference wave between two or more conditions for a given ERP component. The most common statistical approach to dealing with this issue uses data clustering combined with permutation testing (Maris & Oostenveld, 2007). By examining data across all electrode sites, these analyses ensure comprehensive spatial coverage, increase the statistical power, control for multiple comparisons, and allow for the identification of a cluster of electrodes where the component of interest is largest (Maris & Oostenveld, 2007; Groppe, Urbach, & Kutas, 2011; Luck & Gaspelin, 2017). In the standard univariate approach, the difference in ERP amplitude between the two conditions is quantified at multiple time points. For the mass univariate analyses, we have a significantly larger set of (channel, time) pairs, also referred to as samples, wherein we aim to examine the N400 effect.

In our analysis, each word-level EEG epoch will have three dimensions: 32 electrodes x 20 participants x 600 time points (ranging from -100 to 500 ms after word onset), requiring multiple comparisons. In comparison to single-sensor analyses, the multiple comparisons problem is significantly more pronounced in this context: with 32 channels and 600 temporal points, we generate 19,200 t-values. The cluster-based permutation test (10,000 times) used a point-wise independent samples t-test to identify clusters with data points below the alpha level ($p < 0.05$). For multi-channel analyses, the approach is similar to single-channel analyses but differs in clustering: rather than clustering based solely on temporal adjacency, and we now cluster the selected (channel, time) samples based on both spatial and temporal adjacency.

A major advantage of this analysis is its capability to identify the optimal group of electrode sites within a specific time window. Once clusters of (channel, time) samples are identified, those with p-values exceeding the critical two-sided alpha-level (0.05) are considered insignificant. The false discovery rate (FDR) is calculated and used to correct for multiple comparisons (Riffenburgh & Gillen, 2020). Each significant cluster represents a contiguous block of activity across time points and potentially across electrode channels. By averaging the difference waves for all participants within an optimal cluster of electrode sites, we can approximate the electrode locations that best account for the ERP effect. Note that if the approach does not yield any optimal electrode sites using FDR correction, we will select optimal channels without applying the correction. Table 1 shows the results of the optimal clusters of electrode sites in the N400 time window.

Group	**Baseline Trials**		**Optimal Electrode Cluster**
	High	**Low**	
Content	193	365	P3, P7, CP1, CP2, Pz, P4, Fp2, F7, F3, Fz, F4
Function	318	155	FT10, FC1, FC2, C3, Cz, C4, Fp2, F7, F3, Fz, F4
Noun	125	93	P3, P7, CP1, CP2, Pz, P4, F7, F3, Fz, F4
Verb	40	165	T8, CP6, FT9, P3, P7, CP1, CP2, Pz, P4, F7, F3, Fz, F4

Table 1: Baseline trial counts and optimal electrode clusters in the N400 time window for each condition. **High** and **Low** refer to high and low cloze conditions.

[1] https://github.com/Riccorl/ipa

Decoding Analysis

Compared to other Machine Learning (ML) models, such as linear discriminant analysis and random forests, support vector machine (SVM) has demonstrated superior performance in examining N400 effects of prediction and semantic relatedness (Trammel, Khodayari, Luck, Traxler, & Swaab, 2023). SVM is particularly effective in decoding EEG/ERP data due to its ease of implementation, strong performance with small training sets, and especially its ability to handle non-linear relationships in high-dimensional spaces (Carrasco, Bahle, Simmons, & Luck, 2024) by using kernel functions.

The flowchart in Figure 1 illustrates EEG data decoding using an SVM with an RBF kernel. The process begins with the collection of neural data from multiple subjects, where each participant's data is organised into a 3-dimensional array corresponding to trials, electrode sites, and time points. To reduce the potential bias, the trials are randomly shuffled, mitigating order effects and ensuring better generalisation. Single-trial EEG epochs are often too noisy to effectively decode subtle stimulus classes. An increased signal-to-noise ratio (SNR) can be achieved by randomly dividing the data into M sets of approximately a certain number of trials each, creating an averaged ERP for each set (Isik, Meyers, Leibo, & Poggio, 2014; Carrasco et al., 2024). Many previous studies (Isik et al., 2014; Grootswagers, Wardle, & Carlson, 2017; Carrasco et al., 2024) have used multiple sets of 10–20 trials and achieved the highest or near-highest performance in classifier accuracy. Therefore, an average of 15 trials was utilised in this study. For example, 80 high- and 80 low-cloze for the content group could be averaged into five sets of 16 trials per condition. This approach also captures within-subject variability and enhances decoding reliability.

The averaged ERPs are then subjected to stratified K-fold cross-validation (Bishop & Nasrabadi, 2006), a technique that addresses the unbalanced distribution of classes and minimizes overfitting risks (Bae & Luck, 2018; Carrasco et al., 2024), ensuring each class receives fair representation in both training and validation phases. In our study, the SVM-based decoding was repeated five times. For each round, an SVM classifier with an RBF kernel was trained on four folds and tested on the remaining fold. As in the main decoding procedure, this procedure was applied to each time point independently. Decoding accuracy is defined as the proportion of test cases that are correctly classified. To increase the resolution of the decoding accuracy, the entire process is iterated 100 times (L) for each participant. For each iteration, we re-randomised the assignment of trials to averages and trained new SVMs.

If the neural patterns encode distinct lexical class information, decoding accuracy should be greater than the 0.5 chance level for binary classifications. To compare decoding accuracy to the chance at each time point while controlling for multiple comparisons, we used a cluster-based permutation technique similar to that in the univariate analysis. We used a one-sample t test to compare the mean accuracy across participants to chance. However, one-tailed tests were used for the SVM decoding accuracies instead of two-tailed tests in the univariate technique because the SVM classifier should not produce meaningful below-chance decoding.

For a given score (i.e., a voltage in a univariate difference wave or a decoding accuracy), we selected a random sample of n of the N participants (sampling from the set of N participants with replacement) and computed the effect size for this random sample. We then conducted 5,000 iterations of the testing procedure, making it possible to construct the null distribution of the maximum cluster-level t_{max} with a cluster p-value.

Comparison of Performance Between Methods

Both SVM-based decoding and univariate analysis showed significant N400 amplitude differences between lexical groups but could not quantify effect magnitude or directly compare ERP amplitude variations. A common solution is to use effect size or statistical power via Cohen's d_z (Cohen, 2013). Effect size measures the magnitude of the difference between outcomes. This approach allows researchers to present the magnitude of reported effects in a standardised metric, which can be understood regardless of the scale used to measure the dependent variable (Lakens, 2013).

This metric quantifies the ability of each approach to produce statistically significant results while accounting for variation in wave difference (in univariate analysis) and decoding accuracy (in SVM-based decoding analysis) across participants as well as the mean. A larger effect size reduces the Type II error rate, thereby increasing the proportion of significant effects (Ioannidis, 2005). Thus, the approach that generated a larger Cohen's d_z exhibited greater statistical power for detecting differences between experimental conditions in a within-subjects analysis. The Cohen's d_z is calculated as:

$$d_z = \frac{\bar{X} - \text{chance}}{\sigma_X}$$

In our analysis, regarding the SVM-based decoding approach, X represents the mean decoding accuracy across participants, with a chance level of 0.5, and σ_X denotes the standard deviation of these accuracy values. For the univariate analysis, X represents the mean wave difference between two conditions across participants, with a chance level of 0, and σ_X denotes the standard deviation of these difference values. Bootstrapping (10,000 iterations) is then used to estimate the standard error of the effect size.

Results

Content vs. Function Words

Traditional Univariate Analysis Figure 2 summarizes the performance of the N400 differences (measured in μV) between high and low cloze in both lexical classes. Results from the temporal permutation cluster test show that the N400 differences are statistically significant for content words from 150 to 500 ms after word onset ($p < .05$). In contrast, no

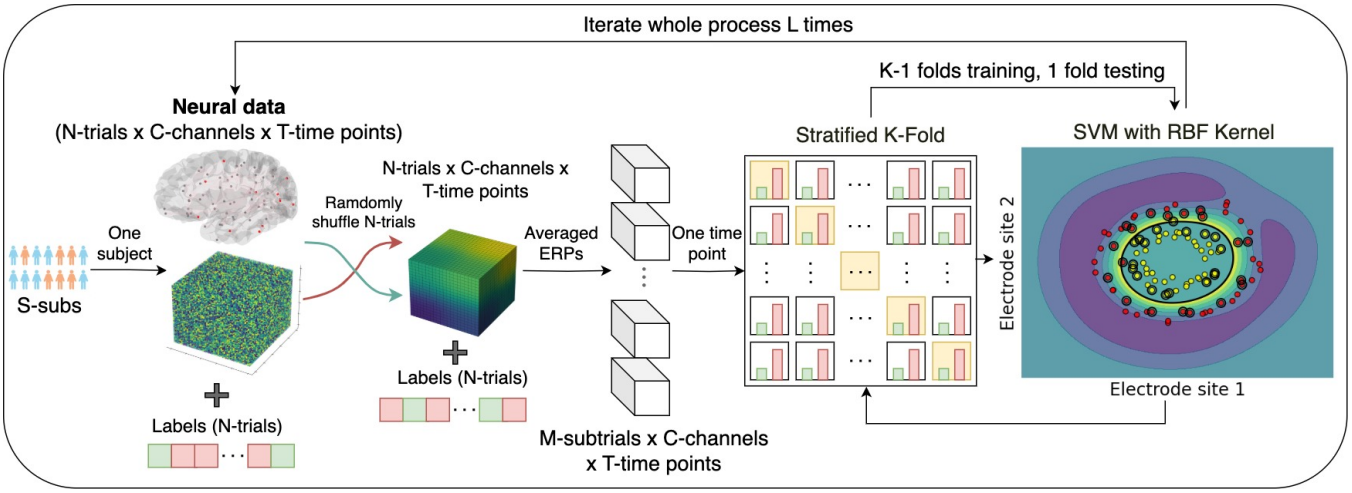

Figure 1: Flowchart of the EEG decoding process using an SVM with an RBF kernel. The process begins with EEG data from each participant, organized into 3D arrays by N trials, C electrodes, and T time points. Trials are randomly shuffled to reduce bias, and the SNR is increased by averaging certain EEG trials per set to create data with M sub-trials, C electrodes, and T time points. Stratified K-fold cross-validation is applied to balance training and testing data, followed by SVM training and testing on each fold. The process is repeated L times per participant, with a total of S participants.

significant difference in N400 predictability effects was observed for function words relative to word onset after correction for multiple comparisons. Indeed, function words have privileged access to prediction, especially when they are very frequent (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009), making their N400 responses less sensitive to cloze probability differences. Conversely, content words, being generally less frequent and more concrete in meaning (Bell et al., 2009), show greater sensitivity to such differences.

Decoding Analysis Figure 3 shows that word class was decodable from the neural responses during the entire epoch ($p < .05$). Interestingly, the SVM decoding demonstrated strong performance, with accuracy consistently exceeding chance levels (0.5), except for a very few time points in the early time course where it slightly dropped below chance, likely accounted for by random noise rather than systematic misclassification.

Decoding accuracy for content words increased notably over time, peaking at 300-500 ms after word offset, with a maximum accuracy of around 0.65. Although function words were also encoded in neural activity, their decoding accuracy remained around 0.55. This difference between content and function words was possibly accounted for by cloze predictability of the upcoming word, in which neural activity increasingly differentiates between high and low cloze content words, while function words, which are more syntactically oriented and carry less semantic weight, generate less distinctive neural responses based on cloze probability.

Comparisons Both approaches effectively detect neural differences between high and low cloze probability words but offer different insights. The univariate analysis highlights the amplitude difference of the N400 component but struggles to differentiate between content and function words. In contrast, SVM-based decoding provides a more dynamic and precise measure of how the brain processes word predictability over time. The raincloud plots (rightmost) showing the absolute N400 amplitude (Figure 2) and the decoding accuracy (Figure 3) indicate that the distinction is easier to observe through decoding results than through traditional ERP analysis.

Nouns vs. Verbs

Traditional Univariate Analysis Figure 4 examines how nouns and verbs are processed under varying levels of predictability (high vs. low cloze probability) within their optimal electrode clusters. Using the permutation cluster approach at a threshold of $p < .05$, we did not find significant clusters for either nouns or verbs. However, at a more liberal threshold of $p < .1$, verbs showed a significant difference in the N400 time window, whereas nouns still showed no neural differences between predictability conditions. This finding suggests that verbs may be more sensitive to contextual predictability than nouns, likely due to the greater complexity and flexibility of verb semantics in varying contexts (Mätzig, Druks, Masterson, & Vigliocco, 2009; Earles & Kersten, 2017). To obtain more robust results, we plan to collect additional cloze data in the next version of the DERCo dataset, increasing statistical power for ERP analyses across grammatical categories.

Decoding Analysis The results in Figure 5 indicate that nouns acquired a higher decoding performance than verb, suggesting that neural signals for nouns carry more distinct information regarding their predictability across high and low

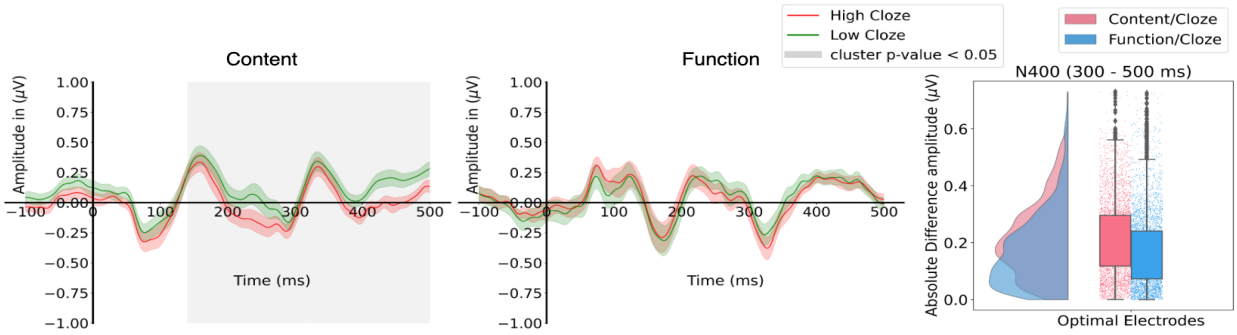

Figure 2: ERPs plotted for content words (leftmost) and function words (middle) with high cloze (red) and low cloze (green). Gray shading indicates significant time points (p-values < 0.05, two-tailed). The raincloud plots (rightmost) show the absolute N400 amplitude difference between content/cloze (pink) and function/cloze (blue) conditions across optimal electrodes.

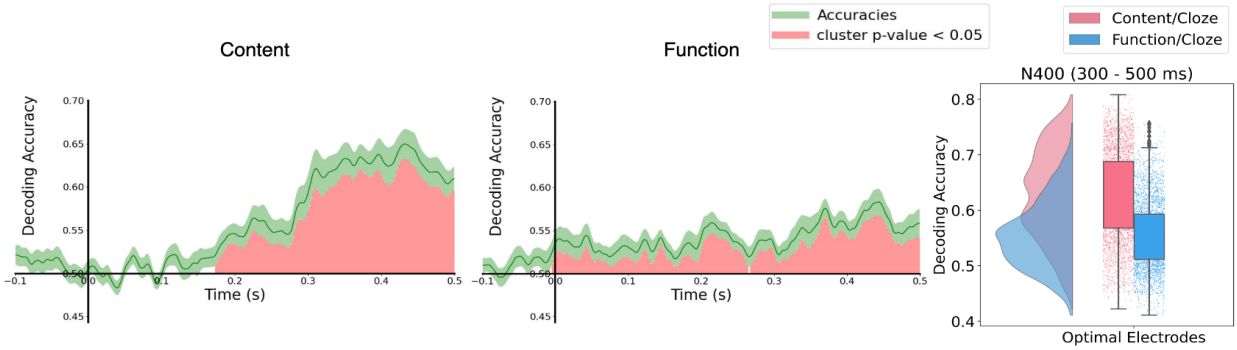

Figure 3: Decoding accuracy plotted for content words (leftmost), function words (middle) with high cloze (red) and low cloze (green). Red shading indicates significant time points (p-values < 0.05, one-tailed). The raincloud plots (rightmost) show the decoding accuracy for content/cloze condition (pink) and function/cloze condition (blue) across optimal electrodes.

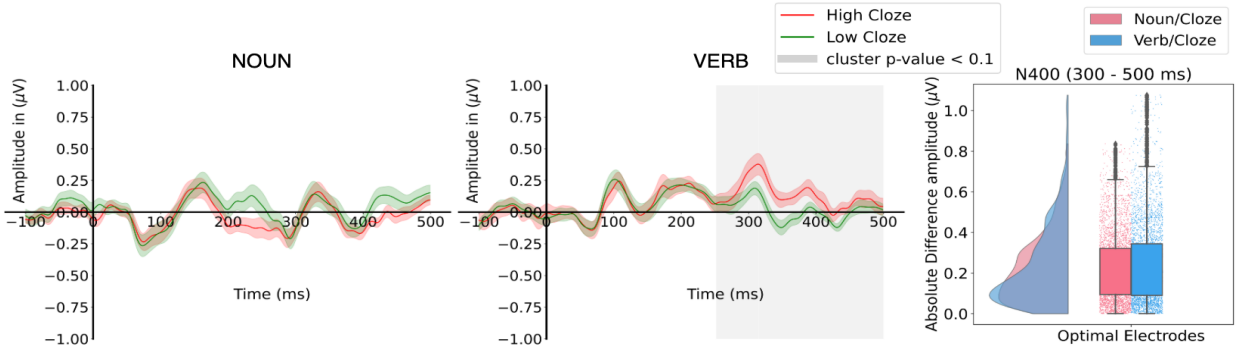

Figure 4: ERPs plotted for nouns (leftmost) and verbs (middle) with high cloze (red) and low cloze (green). Gray shading indicates significant time points (p-values < 0.1, two-tailed). The raincloud plots (rightmost) show the absolute N400 amplitude difference between nouns/cloze (pink) and verbs/cloze (blue) conditions across optimal electrodes.

cloze word groups, particularly during the N400 time window (accuracy $> 10\%$). This richer representation might be due to the generally more stable and predictable nature of nouns in language, making them easier to decode at the neural level (Vigliocco, Vinson, Druks, Barber, & Cappa, 2011). Verbs, however, show lower overall decoding accuracy, reflecting more complex and distributed processing required for integrating verbs into the sentence structure (Bird, Howard, & Franklin, 2000; Earles & Kersten, 2017).

Comparisons Nouns, despite showing a less pronounced N400 amplitude differences between high and low cloze in the univariate analysis, carry more distinct and accessible neural information than verbs, as outlined in decoding accuracy. This discrepancy arises because traditional ERP analy-

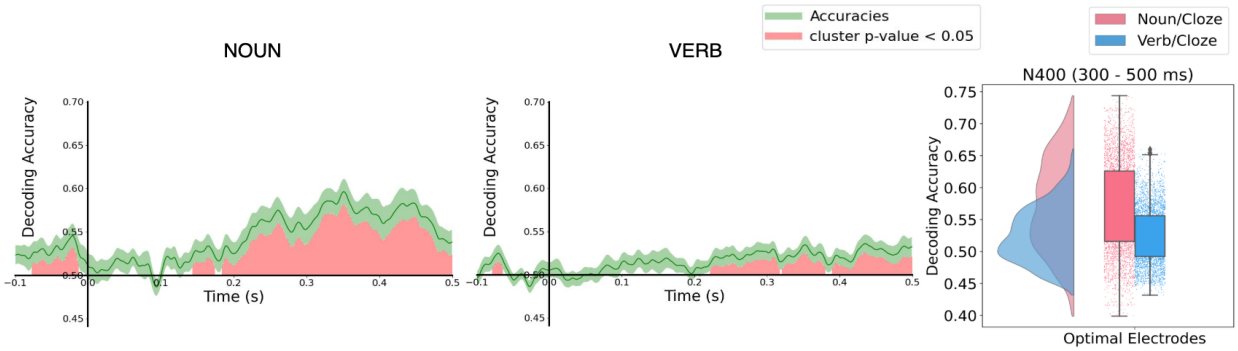

Figure 5: Decoding accuracy plotted for nouns (leftmost), verbs (middle) with high cloze (red) and low cloze (green). Red shading indicates significant time points (p-values < 0.05, one-tailed). The raincloud plots (rightmost) show the decoding accuracy for nouns/cloze condition (pink) and verbs/cloze condition (blue) across optimal electrodes.

ses quantify neural response differences between conditions while assuming a consistent scalp distribution across participants. In contrast, decoding analyses assess the amount of information about a condition present in the recorded signal for each participant individually without assuming uniform brain activity across participants. This approach allows decoding to uncover more aspects that might be missed by applying traditional ERP methods.

Results of Size Effects

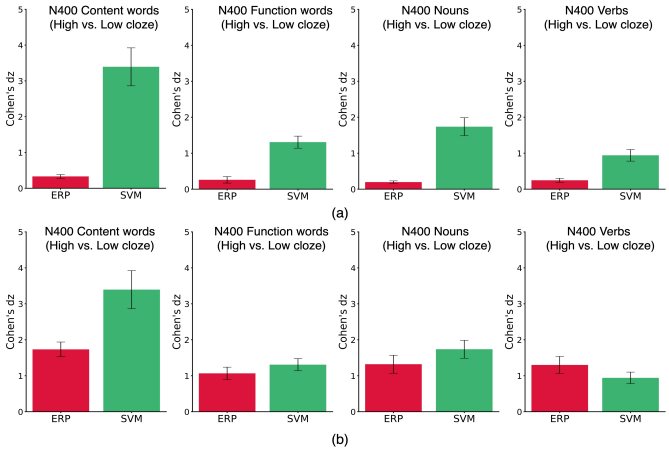

Figure 6: Effect size (Cohen's d_z) plotted on the same scale for each analysis in the N400 time window, comparing the traditional ERP analysis and the decoding technique, with (a) all electrodes and (b) optimal electrodes. The error bars show ± 1 standard error.

Figure 6 provides a comparative analysis of N400 effect sizes between SVM-based decoding and traditional ERP analysis across four cases: high versus low cloze probabilities for content words, function words, nouns, and verbs. In the univariate approach, the difference in ERP amplitude between the two conditions was quantified at a single electrode site. To increase effect size in conventional ERP analysis, we averaged the difference waves across 32 electrode sites.

In Figure 6 (a), SVM-based decoding consistently outperformed conventional ERP analysis in effect sizes across all cases. To increase the effect size of traditional ERP analysis, we used the optimal group of electrode sites that best accounts for the N400 effect (see Table 1). As shown in Figure 6 (b), the decoding technique still demonstrated superior sensitivity and effect sizes despite a significant improvement in ERP effect sizes. These results highlight the potential advantages of employing multivariate decoding techniques in neurocognitive language research.

Discussion and Conclusion

By leveraging both traditional ERP analysis and decoding techniques, we have provided compelling evidence for the neural mechanisms of next-word prediction in reading through N400 analysis. Our results show that content words, carrying more semantic information, evoke stronger N400 effects than function words, which primarily aid in syntactic structuring, supporting prior research on the greater engagement of semantic processing in content words. Moreover, the decoding technique offers valuable theoretical insights, revealing that although verbs exhibit greater N400 differences than nouns, nouns carry more distinct predictability information. The comparison between two techniques highlights the advantages of multivariate techniques in capturing neural complexity. While univariate approaches identify broad patterns, they often miss subtle differences in brain activity.

While this study focused on the N400 component, our ongoing research explores additional ERP components, such as the P600, to gain further insights into syntactic processing (Gouvea, Phillips, Kazanina, & Poeppel, 2010) and provide a more holistic understanding of the neural mechanisms involved in reading comprehension. Likewise, future research should consider using additional decoding techniques to compare their results and potentially gain a more comprehensive understanding of the neural data (Trammel et al., 2023; Carrasco et al., 2024).

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183 and 13/RC/2106_P2. We would like to thank the anonymous reviewers for their helpful remarks.

Code Availability

All the scripts for analysis can be found at `https://github.com/Tayerquach/brain_decoding_model`.

References

Bae, G.-Y., & Luck, S. J. (2018). Dissociable decoding of spatial attention and working memory from eeg oscillations and sustained potentials. *Journal of Neuroscience*, *38*(2), 409–422.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111.

Bird, H., Howard, D., & Franklin, S. (2000). Why is a verb like an inanimate object? grammatical category and semantic category deficits. *Brain and language*, *72*(3), 246–309.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior research methods*, *42*(3), 665–670.

Carrasco, C. D., Bahle, B., Simmons, A. M., & Luck, S. J. (2024). Using multivariate pattern analysis to increase effect sizes for event-related potential analyses. *Psychophysiology*, e14570.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, *39*(39), 7722–7736.

Earles, J. L., & Kersten, A. W. (2017). Why are verbs so hard to remember? effects of semantic context on memory for verbs and nouns. *Cognitive science*, *41*, 780–807.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, *41*(4), 469–495.

Federmeier, K. D., Segal, J. B., Lombrozo, T., & Kutas, M. (2000). Brain responses to nouns, verbs and class-ambiguous words in context. *Brain*, *123*(12), 2552–2566.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the p600. *Language and cognitive processes*, *25*(2), 149–188.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of cognitive neuroscience*, *29*(4), 677–697.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields ii: Simulation studies. *Psychophysiology*, *48*(12), 1726–1737.

He, T., Boudewyn, M. A., Kiat, J. E., Sagae, K., & Luck, S. J. (2022). Neural correlates of word representation vectors in natural language processing models: Evidence from representational similarity analysis of event-related brain potentials. *Psychophysiology*, *59*(3), e13976.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of neurophysiology*, *111*(1), 91–102.

Klem, G. H. (1999). The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology. *Electroencephalogr. Clin. Neurophysiol. Suppl.*, *52*, 3–6.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59.

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.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, *4*, 863.

Lee, C.-l., & Federmeier, K. D. (2006). To mind the mind: An event-related potential study of word class and semantic ambiguity. *Brain Research*, *1081*(1), 191–202.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any erp experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

Mätzig, S., Druks, J., Masterson, J., & Vigliocco, G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex*, *45*(6), 738–758.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*, *15*(3), 1033–1042.

Münte, T. F., Wieringa, B. M., Weyerts, H., Szentkuti, A., Matzke, M., & Johannes, S. (2001). Differences in brain

potentials to open and closed class words: Class and frequency effects. *Neuropsychologia*, *39*(1), 91–102.

Neville, H. J., Mills, D. L., & Lawson, D. S. (1992). Fractionating language: Different neural subsystems with different sensitive periods. *Cerebral cortex*, *2*(3), 244–258.

Ngabut, M. N. (2015). Reading theories and reading comprehension. *Journal on English as a Foreign Language*, *5*(1), 25–36.

Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill.

Quach, B. M., Gurrin, C., & Healy, G. (2024). Derco: A dataset for human behaviour in reading comprehension using eeg. *Scientific Data*, *11*(1), 1104.

Riffenburgh, R. H., & Gillen, D. L. (2020). *Statistics in medicine*. Academic press.

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

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from eeg: A classification method comparison. *NeuroImage*, *277*, 120268.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426.