Uncovering Linguistic Representations in MEG Data Using Deep Learning

Nikola Kölbl (nikola.koelbl@fau.de)
Neuroscience Lab, University Hospital Erlangen, Waldstrasse 1
91054 Erlangen, Germany

Abhinav Singh (abhinav.singh@fau.de)
CCN Group, Pattern Recognition Lab, Immerwahrstrasse 2a
91058 Erlangen, Germany

Patrick Krauss (patrick.krauss@fau.de)
CCN Group, Pattern Recognition Lab, Immerwahrstrasse 2a
91058 Erlangen, Germany

Achim Schilling (achim.schilling@fau.de)
Neuromodulation and Neuroprosthetics, University Hospital Mannheim, University Heidelberg, Theodor-Kutzer-Ufer 1-3
68167 Mannheim, Germany

Abstract

The ability to use complex language is uniquely human and underpins abstract thought, cultural transmission, and the structure of society. Understanding its neural basis in naturalistic settings remains a major challenge. In this study, we investigate whether word classes can be decoded from low-dimensional MEG data recorded during audio book listening. Using a minimalist modeling approach, we trained neural networks on individual MEG channels and identified peak classification performance over left frontal sensors, consistent with the involvement of Broca's area in grammar and predictive processing. As a proof of concept, we applied sequential deep dreaming to reveal prototypical neural patterns for nouns and verbs. While the results demonstrate feasibility, limitations due to data sparsity, class imbalance and single-subject design highlight the need for broader validation. Our approach represents a first step towards interpretable decoding of linguistic structure from MEG during natural, continuous speech comprehension.

Keywords: MEG; language; ERF; word classes; AI; sequential deep dreaming; frontal cortex

Introduction

Deciphering how the human brain processes language in real-world settings remains a fundamental challenge in cognitive neuroscience. Magnetoencephalography (MEG), with its millisecond-level temporal resolution, offers a powerful window into the rapid neural dynamics that underlie natural speech comprehension (Hansen, Kringelbach, & Salmelin, 2010; Tavabi, Obleser, Dobel, & Pantev, 2007). However, naturalistic stimuli such as audio books pose analytical difficulties. Unlike isolated word presentations, continuous speech elicits temporally overlapping neural responses, complicating the attribution of activity to specific linguistic units (Schilling et al., 2021; Koelbl et al., 2024; Garibyan, Schilling, Boehm, Zankl, & Krauss, 2022). Another difficulty lies in the limited number of occurrences of certain linguistic categories, such as word classes, in natural speech. This data sparsity is a particular problem for MEG, where a large number of features can easily lead to overfitting and poor model generalization. In addition, MEG has a low signal-to-noise ratio and is particularly sensitive to extraneous noise and physiological artifacts (Burgess, 2020; Hansen et al., 2010), making it more difficult to detect consistent neural patterns. While classification of MEG data can demonstrate that linguistic categories such as word classes are decodable, it provides limited insight into how these categories are actually represented and processed in the brain. To go beyond decoding accuracy, we use a generative approach - sequential deep dreaming (Schlegel, Keim, & Sutter, 2024) - to explore the internal representations learned by the model. Sequential deep dreaming is a method of iteratively generating input patterns that maximally activate a trained neural network, allowing researchers to uncover prototypical features associated with particular classes (Ellis, Sendi, Miller, & Calhoun, 2021). Thus, we use this technique to identify prototypical neural response patterns associated with different word classes, providing a window into the brain's encoding of linguistic structure.

In this first study, we explore the feasibility of decoding word-class information from MEG recordings collected during naturalistic audio book listening in a single participant. To address the challenges of overlapping neural responses and data sparsity, we employ a streamlined modeling approach: individual MEG channels are used to train a fully connected neural network, with both spatial and temporal resolution deliberately reduced to minimize complexity and overfitting. As a proof of concept, we also apply a sequential deep-dreaming approach to the trained model to uncover prototypical neural patterns associated with word classes. Together, these methods lay the groundwork for extracting structured linguistic representations from low-dimensional MEG data, while highlighting the preliminary nature of the results.

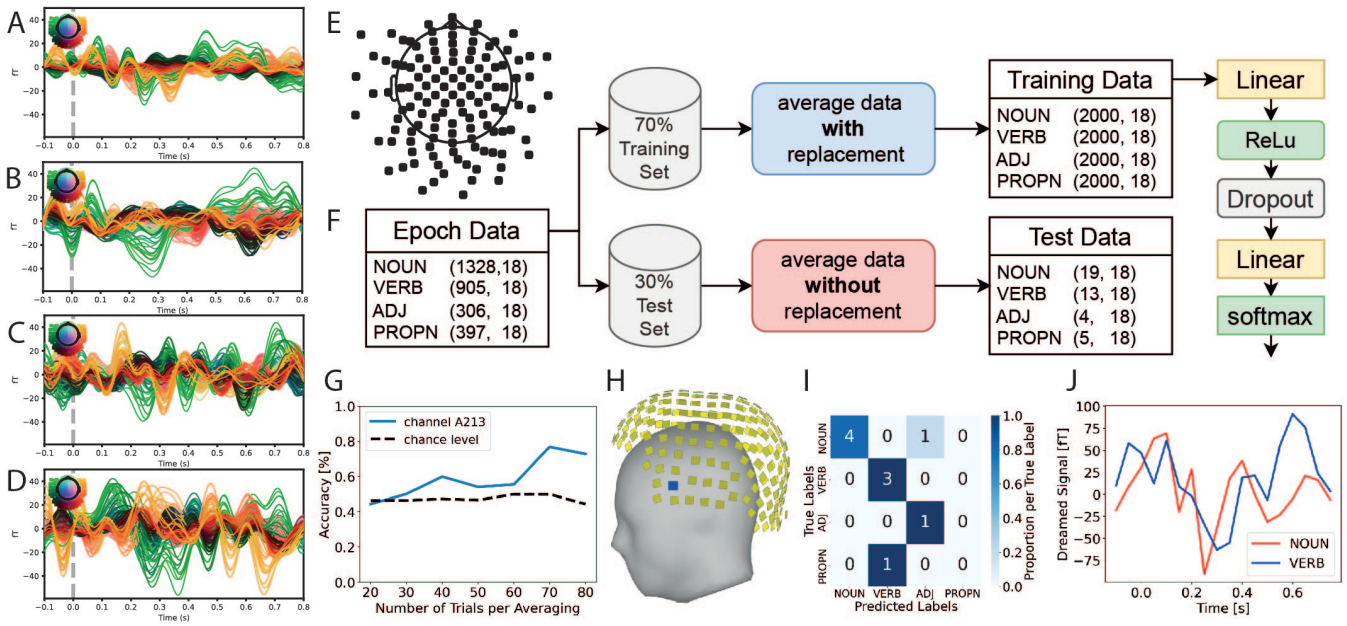

Figure 1: Average brain activity of one subject for nouns (A), verbs (B), adjectives (C) and proper nouns (D). (E): Sensor positions; (F): Workflow of data splitting, sampling, and training the classifier network; (G): Classification accuracies for best sensor A213 (location in H). I: Confusion matrices for the classification results of 70 trials; J: Prototypical ERFs signals generated through sequential deep dreaming using the weights of corresponding classifier network.

Methods

In this study, brain activity was recorded while a participant listened to approximately 50 minutes of the German audio book "Vakuum" by Philip P. Peterson (Argon Verlag). Magnetoencephalography (MEG) data were acquired using a 248-channel system (Magnes 3600WH, 4D-Neuroimaging). The study was approved by the Ethics Committee of the University Hospital Erlangen (Approval No. 22-361-2, PK). Pre-processing was performed using MNE-Python (v1.7.1) (Gramfort et al., 2013), including detection and interpolation of bad sensors, band-pass filtering (1–20 Hz), and downsampling to 200 Hz. Independent component analysis (ICA) was applied to further reduce artifacts (Ferrante et al., 2022; Koelbl, Schilling, & Krauss, 2023). To simplify the data for classification, we applied an additional low-pass filter (1–10 Hz) and further downsampled the signal to 20 Hz. Linguistic annotation of the audio book transcript was carried out using spaCy (Honnibal, Montani, Van Landeghem, & Boyd, 2020), and four word classes were selected for analysis: nouns (n = 1328), verbs (n = 905), adjectives (n = 306), and proper nouns (n = 397). MEG data were segmented from 100 ms before to 800 ms after word onset for each instance (Figure 1 A–D: Average event-related fields (ERFs) for the four word classes).

Given the imbalanced distribution of word classes in our dataset (ranging from 306 to 1328 samples) and the overall limited sample size, we implemented tailored sampling strategies for training and testing (Fig. 1 F). To improve the signal-to-noise ratio, we averaged multiple trials to generate class-specific ERFs. The training set (70% of the data) was upsampled to 2,000 instances per class using sampling with replacement, while the test set (30%) was sampled without replacement to preserve the original class distribution. Additionally, we addressed the effect of trial averaging on classification performance by generating ERFs with varying numbers of trials, from 20 to 80 in steps of 10. To reduce data dimensionality, we trained models on every second MEG channel individually (channels A1–A248, step size 2; Fig. 1 E). For classification, we used a compact fully connected neural network consisting of two linear layers with dropout (p = 0.3) for regularization, implemented in PyTorch (Paszke et al., 2019) (layer sized specified in Fig. 1 F).

Results

The most robust and consistent classification accuracy (Fig. 1 G) was observed for channel A213 (Fig. 1 H), located in the left prefrontal region, suggesting that this area may contain particularly discriminative information to distinguish word categories during naturalistic language processing (confusion matrix in Fig. 1 I). Despite the inherent variability in neural responses during continuous speech, driven by contextual, linguistic and attentional dynamics, the model achieved a meaningful classification of word types. As a first exploratory step, we applied sequential deep dreaming to the trained network, which generated distinct prototypical ERF patterns for nouns and verbs (Fig. 1 J). Given the limitations of the dataset, both in terms of size and balance, the findings from the deep dreaming analysis are exploratory in nature and should be interpreted with caution. The under-representation of certain word classes, particularly proper nouns and adjectives, limited both the statistical power of the study and the performance of

the classifier. In addition, inter-individual variability in MEG responses remains a critical factor, highlighting the need for larger, multi-participant datasets to validate and generalize these initial findings.

Discussion

The present study investigates the feasibility of decoding word categories from MEG data recorded during naturalistic audio book listening. Using a minimalist modeling approach - training a simple neural networks on individual MEG channels - we demonstrate that word class information can be extracted from low-dimensional data. In addition, we apply sequential deep dreaming to explore prototypical neural response patterns, providing preliminary insights into category-specific brain activity during continuous speech processing. The fact that a sensor placed above the left frontal cortex yielded the most robust classification performance suggests a critical role for this region in distinguishing between word classes during naturalistic language processing. This is consistent with the well-established involvement of Broca's area in grammatical structure processing and morphosyntactic integration (Sahin, Pinker, Cash, Schomer, & Halgren, 2009), as well as predictive coding accounts that implicate frontal regions in anticipating upcoming linguistic information based on contextual cues (Grisoni, Tomasello, & Pulvermüller, 2021; Grisoni, Miller, & Pulvermüller, 2017).

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): 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.

References

Burgess, R. C. (2020). Recognizing and correcting meg artifacts. *Journal of Clinical Neurophysiology*, *37*(6), 508–517.

Ellis, C. A., Sendi, M. S., Miller, R., & Calhoun, V. (2021). A novel activation maximization-based approach for insight into electrophysiology classifiers. In *2021 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 3358–3365).

Ferrante, O., Liu, L., Minarik, T., Gorska, U., Ghafari, T., Luo, H., & Jensen, O. (2022). Flux: A pipeline for meg analysis. *NeuroImage*, *253*, 119047.

Garibyan, A., Schilling, A., Boehm, C., Zankl, A., & Krauss, P. (2022). Neural correlates of linguistic collocations during continuous speech perception. *Frontiers in Psychology*, *13*, 1076339.

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.

Grisoni, L., Miller, T. M., & Pulvermüller, F. (2017). Neural correlates of semantic prediction and resolution in sentence processing. *Journal of Neuroscience*, *37*(18), 4848–4858.

Grisoni, L., Tomasello, R., & Pulvermüller, F. (2021). Correlated brain indexes of semantic prediction and prediction error: Brain localization and category specificity. *Cerebral Cortex*, *31*(3), 1553–1568.

Hansen, P., Kringelbach, M., & Salmelin, R. (2010). *Meg: An introduction to methods*. Oxford university press.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.
doi: 10.5281/zenodo.1212303

Koelbl, N., Mueller-Voggel, N., Rampp, S., Kaltenhaeuser, M., Tziridis, K., Krauss, P., & Schilling, A. (2024). Analyzing differences in processing nouns and verbs in the human brain using combined eeg and meg measurements. *bioRxiv*, 2024–12.

Koelbl, N., Schilling, A., & Krauss, P. (2023). Adaptive ica for speech eeg artifact removal. In *2023 5th international conference on bio-engineering for smart technologies (biosmart)* (pp. 1–4).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., & Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within broca's area. *Science*, *326*(5951), 445–449.

Schilling, A., Tomasello, R., Henningsen-Schomers, M. R., Zankl, A., Surendra, K., Haller, M., ... Krauss, P. (2021). Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods. *Language, Cognition and Neuroscience*, *36*(2), 167–186.

Schlegel, U., Keim, D. A., & Sutter, T. (2024). Finding the deepdream for time series: Activation maximization for univariate time series. *arXiv preprint arXiv:2408.10628*.

Tavabi, K., Obleser, J., Dobel, C., & Pantev, C. (2007). Auditory evoked fields differentially encode speech features: an meg investigation of the p50m and n100m time courses during syllable processing. *European journal of neuroscience*, *25*(10), 3155–3162.