MEG Encoding using Word Context Semantics in Listening Stories

Subba Reddy Oota, Nathan Trouvain, Frederic Alexandre, Xavier Hinaut

MEG Encoding using Word Context Semantics in Listening Stories

Subba Reddy Oota[1], *Nathan Trouvain*[1], *Frederic Alexandre*[1], *Xavier Hinaut*[1]

[1]Inria Bordeaux, France,

{subba-reddy.oota,nathan.trouvain,frederic.alexandre,xavier.hinaut}@inria.fr

Abstract

Brain encoding is the process of mapping stimuli to brain activity. There is a vast literature on linguistic brain encoding for functional MRI (fMRI) related to syntactic and semantic representations. Magnetoencephalography (MEG), with higher temporal resolution than fMRI, enables us to look more precisely at the timing of linguistic feature processing. Unlike MEG decoding, few studies on MEG encoding using natural stimuli exist. Existing ones on story listening focus on phoneme and simple word-based features, ignoring more abstract features such as context, syntactic and semantic aspects. Inspired by previous fMRI studies, we study MEG brain encoding using basic syntactic and semantic features, with various context lengths and directions (past vs. future), for a dataset of 8 subjects listening to stories. We find that BERT representations predict MEG significantly but not other syntactic features or word embeddings (e.g. GloVe), allowing us to encode MEG in a distributed way across auditory and language regions in time. In particular, past context is crucial in obtaining significant results.

Index Terms: brain encoding, human-computer interaction, MEG, syntax, semantics, context length

1. Introduction

Over the past decade, Brain-Computer Interface (BCI) helped to make significant progress in understanding language processing in the brain using a popular computational paradigm: Brain encoding, the process aiming to map stimuli features to brain activity. The central aim of brain encoding for language processing analysis is to unravel how the brain represents linguistic knowledge (i.e. semantic and syntactic properties) and carries out sentence-processing information [1, 2, 3, 4, 5] by modeling the effect of such information on brain recordings. For instance, using functional Magnetic Resonance Imaging (fMRI) brain recordings, a number of previous studies have investigated the alignment between text stimuli representations extracted from language models (e.g. Bi-directional Encoder Representation Transformer (BERT) [6]) and brain recordings of people comprehending language [7, 8, 9, 10, 11].

While a large part of brain encoding literature uses fMRI brain recordings to study linguistic contrasts involved in language processing, the low temporal resolution of fMRI makes it difficult to link brain activation to specific linguistic processes. Magnetoencephalograph (MEG) recordings, on the other hand, have a better temporal resolution (generally understood as the smallest time period of brain activation that can be distinguished), and allow us to better understand the neural dynamics of the underlying language processing network. However, few studies use MEG to study how word embeddings such as BERT can be related to the brain activity of subjects reading one word at a time from a story [7]. We propose to uncover the insights of humans sentence processing during a naturalistic story listening task.

Studies using word embedding representations and fMRI have revealed that syntactic features are distributively represented across brain language networks and largely overlapped with semantic networks [4, 5]. Despite the great strides in learning sentence comprehension at a functional level, there are still many problems that could benefit from further improvements in understanding sentence structure and meaning at the temporal level. Therefore, investigating how the brain encodes semantic and fine-grained syntactic features of words using MEG recordings seems crucial to understand the timing of language comprehension mechanisms. Some important questions remain to be explored: (1) How much context is maintained through time to process words? (2) Is the direction of context important (past context vs. future context)? The main objective of this work is to address these questions using MEG activity, in time at different sensor locations, for both syntactic and semantic representations during naturalistic story listening.

Brain Regions of Interest (ROIs) for sentence processing: Several MEG studies report an evidence from well-formed natural language expressions for a role of the left posterior temporal lobe (PTL) in incremental syntactic processing. Similarly, postnominal adjectives were relayed to the inferior frontal gyrus (IFG) and influence of semantic type in the left anterior temporal lobe (ATL) [12, 13, 14]. Further, Toneva *et al.* [15] conclude that the involvement of a language network with task-specific settings (e.g. question-answering task) is localized to the frontal and the left temporal lobes. These findings correspond to many fMRI studies [16, 17, 18, 19, 20, 21, 22]. However, the time at which different brain regions are sensitive to distinct syntactic and semantic properties remains unclear.

Word stimulus representations for brain encoding: Several studies have used basic syntactic features such as part-of-speech (POS), dependency relations (DEP), complexity metrics [23, 4, 24], and semantic word embeddings [25, 26, 27, 7, 28, 10] to represent words for fMRI brain encoding with text stimulus. However, modeling these basic syntactic and semantic features for MEG recordings is still unexplored. In this paper, to understand when the brain processes linguistic structure in sentences, we leverage text representations using basic syntactic features as well as semantic features, with various context lengths, directions (past vs. future), and within-context relative importance.

Overall, our main contributions are as follows. (1) We explore: (a) basic syntactic features, (b) GloVe embeddings, and (c) semantic BERT embeddings for MEG brain encoding. We found that only BERT embeddings were predictive of MEG activity. (2) We find that prediction of the MEG activity using BERT is in regions such as the bilateral temporal lobes, frontal

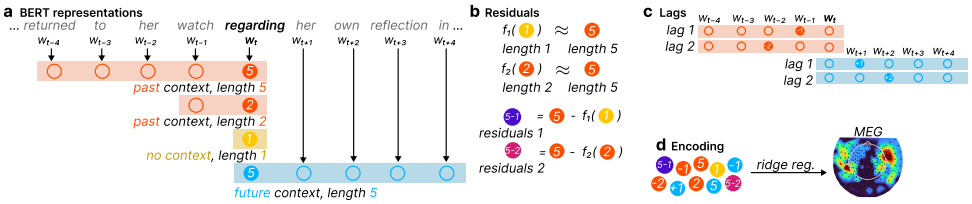

Figure 1: Global schema of the study. *Each circle is a vector embedding for a particular word. Here, "*regarding*" is the current word w_t encoded.* **(a)** *BERT representations are computed with varying context lengths and directions (past vs. future). "*Past (future) context of length n*" means that word w_t is encoded with its n preceding (following) words $w_{t-1}, \dots w_{t-n-1}$ ($w_{t+1}, \dots w_{t+n-1}$). (Red) Past contexts (lengths 5 and 2). (Blue) Future context 5. (Yellow) Absence of context (context 1: static representation of the word).* **(b)** *For a given past (future) context around a word w_t, we name "*residuals n*" the results of filtering information of the n nearest words from the representation of current word w_t (e.g. past residuals 2 are the result of removing information of past context 2 $w_t^2 = [w_t, w_{t-1}]$ from past context 5 w_t^5). Filtering is performed by first fitting a length n sub-context w_t^n to the 5 context w_t^5, and then computing residuals between estimated and real 5 context.* **(c)** *For a given past (future) context around a word w_t, we name "*lag l*" the representation of the word w_{t-l} (w_{t+l}).* **(d)** *For each word, all of these representations are used to predict the subject's MEG activity at word onset in the story using ridge regression.*

lobe and parietal lobe between 250ms to 750 ms (word onset is at 200ms). (3) We report that past context has greater predictive power than future context. When dealing with past context, $R^2 scores$ are proportional to context length.

2. Feature Representations

We used different features computed per word to simultaneously test different syntactic and semantic representations.

(1) Basic Syntactic Features: Similar to [29, 4, 5], we use various multi-dimensional syntactic features such as Complexity Metrics (Node Count (NC), Word Length (WL), Word Frequency (WF)), Part-of-speech (POS) and Dependency tags (DEP), described briefly below. **Node Count (NC)** The node count for each word is the number of subtrees that are completed by incorporating each word into its sentence. **Word Length (WL)** Word length is the number of characters present in the word. **Word Frequency (WF)** Word frequency reports log base-10 of the number of occurrences per billion of a given word in a large text corpus. **Syntactic Surprisal (SS)** Word frequency reports log base-10 of the number of occurrences per billion of a given word in a large text corpus. **Part-of-speech (POS)** We use the Spacy English dependency parser [30] to extract the Part-of-speech (POS). We generate a one-hot vector for each word in which corresponding POS tag location is 1 and remaining tag values are 0. **Dependency tags (DEP)** We use the Spacy English dependency parser [30] to extract the dependency tags. In DEP, we generate a one-hot vector for each word and dependency tag in which corresponding dep tag location is 1 and remaining tag values are 0. **(2) Semantic Features** We use two semantic representations: (1) GloVe (distributed word representations) [31] and (2) BERT (contextualized representations) [6], described briefly below. **GloVe:** word vectors (each word is a 300-dimension vector) [31], and the model always represents unique embedding irrespective of the word appearing in different contexts.

BERT: Given an input sentence, the pretrained BERT [6] outputs word representations at each layer. In this paper, we have used the pretrained BERT-base model. We have not performed any fine-tuning here. Since BERT embeds a rich hierarchy of linguistic signals: surface information at the bottom, syntactic information in the middle, semantic information at the middle to higher layers [32]; hence, we use the #words $\times$ 768D vector from the intermediate layer (layer-7) to obtain the embeddings.

(3) Varying the Context Length of BERT To extract the stimulus features at different context lengths (C = 1, 2, 3, 4, 5, 20), we constrained the model with maximum C words as context length (Fig. 1(a)). Since BERT model process whole sentence, we input all the C context-length words to the BERT model and use the representation of the last word for the past context and first word for the future context. For instance, given a story of M words and considering the context length of 5, while the third word's vector is computed by inputting the network with (w_1, w_2, w_3), the last word's vectors w_M is computed by inputting the network with ($w_{M-5}, \cdots, w_M$). Here, we extracted representations for both past and future contexts.

(4) Residuals To compute residuals from pretrained BERT representations at different context lengths, we use a ridge regression method in which the context w_M (M=1,2,3) as input and the context w_5 is the target vector (Fig. 1(b)). We compute the residuals by subtracting the predicted context from the actual context resulting in the (linear) removal of a particular context from context w_5 (see Fig. 1 for a schematic). Because the MEG brain prediction method is also a linear function (see Section 4), this linear removal limits the contribution of the word importance to the eventual brain prediction performance.

(5) Lags To extract lag l representations, we take as an embedding vector, for a given context length t, the vector of the word w_{t-l} for past context (or w_{t+l} for future context) (Fig. 1(c)). Contrary to residuals, these lag representations still contain information from the current word w_t. Encoding MEG using lag representations assesses how lag word information is correlated to current word MEG activity.

3. Dataset and Experiments

We used data from 8 subjects of the MEG-MASC dataset [33]. The activity from 208 MEG sensors was recorded while each subject listened to naturalistic spoken stories selected from the Open American National Corpus (*"Cable spool boy"*, *"LW1"*, *"Black willow"* and *"Easy money"*). **MEG preprocessing** We performed the minimal processing steps described in [33]. On raw MEG data and for each subject separately, using *MNE-Python* defaults parameters, we (i) bandpass filtered the MEG data between 0.5 and 30.0 Hz, (ii) temporally-decimated the data 10x, (iii) segmented these continuous signals between -200 ms and 600 ms after word onset, (iv) applied a baseline correction between -200 ms and 0 ms, and (v) clipped the MEG data between fifth and ninety-fifth percentile of the data across channels. **Word Processing** Since MEG data is sampled at a higher rate ($1000Hz$) than word presentation, epoching and downsampling yields, for each word, 81 times points recorded at 208 sen-

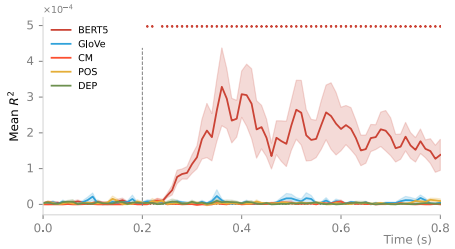

Figure 2: Only BERT representations significantly encode MEG activity. *Plain lines represent mean significant R^2 score (permutation test, $p < 0.05$, FDR correction) between predicted and real MEG activity, across sensors and subjects. Areas around lines represent standard error across subjects. Dots above the figure represent significant difference with 0, for all timestep (one-sample t-test, $p < 0.05$, FDR correction) (color is matching the legend). Word onset is at 200ms.*

sors. There are total of 8567 words across four stories. In our experiments, for each word, the model makes a prediction of MEG activity for all of these 16848 = 208 × 81 values. Here, each word is transformed into one of the feature representations described in section 2.

4. Models and evaluations

Encoding Model To explore how and when syntactic and semantic specific features are represented in the brain when listening to stories, we extract different features describing each stimulus word and used them in an encoding model to predict brain responses (Fig. 1(d)). MEG encoder models attempt to predict brain responses associated with each MEG sensor and each time point when given audio stimuli (spoken words in our case). We trained a model per subject separately. Following the literature on brain encoding [1, 15, 3, 4], we chose to use a ridge regression as encoding model. The ridge regression objective function for the stimulus features is $f(X_s) = \min_{W_s} \|Y_b - X_s W_s\|_F^2 + \lambda \|W_s\|_F^2$. Here, X_s denotes the input stimuli representation, $W_s \in \mathbb{R}^{F_s \times LT}$ are the learnable weight parameters, F_s denotes the number of features in stimuli representation (768), L corresponds to number of MEG sensors (208), T represents the time dimension of the brain activity (81), s denotes the sample stimulus $s \in \mathbb{R}^{F_s}$, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight. λ was tuned on a small disjoint validation set obtained from the training set. **Cross-Validation** We follow 4-fold (K=4) cross-validation. All the data samples from K-1 folds (3 stories data) were used for training, and the model was tested on samples of the left-out fold (1 story data). **Evaluation Metrics** We compute the coefficient of determination R^2 [34] between real and predicted MEG activity to measure prediction performance for each sensor location and each timepoint within epochs. R^2 scores were then averaged over all epochs and across all folds. Along with R^2 score, we also use Root-Mean-Square (RMS) to measure the predicted evoked response, averaged across all MEG sensors, tasks and subjects. RMS scores are reported in supplementary materials. **Statistical Significance** We check R^2 scores statistical significance using a permutation test. We permute blocks of MEG predictions and compute R^2 scores between permuted predictions and real data 5000 times to estimate an empirical distribution of chance performance and corresponding p-values. Finally, the Benjamini-Hochberg False Discovery Rate (FDR) correction [35] is applied on all tests to control the type I error rate. **Implementation Details for Reproducibility** All experiments were conducted on a machine with 1 NVIDIA GEFORCE-GTX GPU with 4GB GPU RAM. We used ridge-regression with the following parameters: MSE loss function, and L2-decay (λ) varied from 10^{-1} to 10^{-3}.

5. Results

In order to assess the performance of MEG encoder models learned using syntactic and semantic representations, we computed the R^2-score between predicted MEG and ground-truth recordings of the evoked response at word onset, across all sensors, folds and subjects. Each figure reports the average R^2-scores of the different features, where all values are first filtered by significance for each time point (i.e. we set to 0 the score values for sensors where $p < 0.05$ after the permutation test and FDR correction procedure described in section 4).

Encoding Performance of Syntactic and Semantic Methods From Fig. 2, we make the following observations: (i) Only BERT-based feature representations significantly correlate to MEG activity, starting around 0.25s (0.05s after word onset). (ii) Basic syntactic (CM, POS and DEP), and non-contextual semantic features (GloVe) are, on average, not correlated with the considered window of MEG activity. These features poor performance may be explained by their overly simple nature or their limited contextual information. To better visualize the predicted MEG performance using these simple features, we report the RMS plot in supplementary. It is observed that the RMS plot for these methods is not closer to the original MEG in comparison to BERT.

Contextual BERT Embeddings: effect of length To assess whether the direction and length of context are important for predicting MEG activity during story listening, we report the R^2-score performance from both past and future BERT contextual representations in Fig. 4. From Fig. 4 (left), we observe that context length plays a crucial role in predicting MEG activity. The performance of this prediction is proportional to the length of the context. However, above a context length of 5, no significant improvement in MEG predictivity is noticeable. Moreover, the difference between context's performance is mainly observed between 300ms to 425ms (100–325ms from word onset). This suggests that MEG activity results from the integration of past auditory information on a short time horizon.

Contextual BERT Embeddings: effect of direction From Fig. 4 (right), we observe that there is a significant effect on the direction of context. All features created from future context display a low correlation with regard to features created using past context. Interestingly, this effect is inversely proportional to context length for future context, where BERT features extracted from a future context of length 5 achieve better R^2 scores than the same features created from a future context of length 20. This suggests that the MEG brain activity mostly correlates to past and current contexts. Inference of future context could not be detected, and if present, only on a short time horizon. Since length 5 context contains the current word, the relative importance of the current word could account for its relatively correct performance in its 5-word context, which is diluted in a 20-word context.

Contextual BERT Embeddings (Residuals vs. Lag) To investigate whether the removal of word-level information (current word, two nearest words, three nearest) from context has any effect in predicting MEG activity, we report the R^2-score performance of residuals in both past and future contexts, as shown in Fig. 3. We also report the lag representations performance,

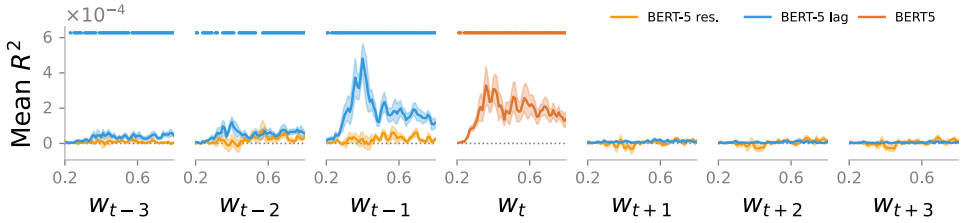

Figure 3: *R^2-score performance of encoding for different lag and residuals of BERT representations. Plots w_{t-n} report, for $n \in [3, 1]$, the performance of lag n and residuals n when encoding MEG activity at word w_t using its past context. Similarly, plots w_{t+n} reports performance of lag and residuals n using future context. Plot w_t displays performance of context 5. Lines, areas and dots figures same metrics as Fig. 2. Word onset is at 200ms.*

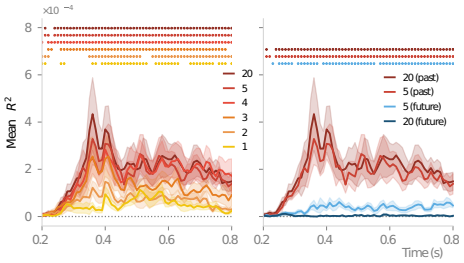

Figure 4: Long past contexts enable better encoding than future or short-scale present contexts. *(left) R^2 given BERT representation context length, from 1 (no context) to 20. (right) R^2 given BERT representation context direction (past vs. future) and length (5 vs. 20). Lines, areas and dots figures same metrics as Fig. 2. Word onset is at 200ms.*

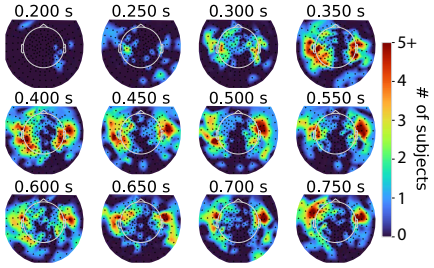

Figure 5: *Significantly predicted MEG activity for each timepoint and each sensor position (permutation test, $p < 0.05$, FDR correction) using BERT past context 5 word embeddings. Color denotes, for each sensor and timepoint, the number of subjects whose MEG activity was significantly predicted. Word onset is at 200ms.*

which represents the performance of the previous word representations in predicting the current word MEG. From Fig. 3, we make the following observations: (i) complete removal of current word information from past context through residual representations (i.e w_{t-n}) has a significant drop in R^2-score. (ii) Similarly, in the future context, the R^2-score performance of residuals is always zero or significantly below chance. (iii) In contrast to residuals, lag representations display significant performance only for lag 1 in the past context. Though, lag 2 and 3 in past contexts display significant above-chance performance. However, this performance is negligible compared to lag 1 performance. (iv) Similar to future context residuals, future context lag representations yield below-chance performance. From these results, we hypothesize that the current word MEG activity is the product of short-term past context and current word information. Both pieces of information are required to render MEG activity at a given word onset accurately. Future context information is not detectable in MEG activity. **Cognitive Insights** Fig. 5 reports MEG sensor locations which are significantly predicted across subjects by 5-context BERT representations (permutation tests, $p < 0.05$, FDR correction). Best brain MEG alignments are in the bilateral temporal and frontal regions between 250ms to 750ms (word onset is at 200ms).

6. Discussion & Conclusion

In this paper, we evaluated the alignment of basic syntactic, distributed word embeddings, and contextualized word representations (varying different context lengths, past vs. future context, residuals, and lags) with MEG brain responses in time. We showed that BERT representations, contrary to other features or GloVe, lead to a significant prediction in brain alignment across auditory and language regions between 50-550ms (250ms to 750ms with word onset at 200ms). Noteworthy, this prediction performance is a function of the amount of available past context, and only past context future or current word.

It is surprising that BERT current word representation alone w_t^1 (BERT-1) allows so weak predictions compared to $w_t^{past \geq 3}$ (BERT with contexts higher than 3) (Fig. 4). Moreover, lag results of Fig.3 shows that the previous BERT-5 word $w_{t-1}^{past3 \& future1}$ allows higher R^2 score than current word with low context $w_t^{past \leq 5}$. Additionally, it is surprising that near future context $w_t^{future5}$ which includes the current word is not relevant for MEG prediction, as if the brain was making no or very few predictions of future incoming words.

This suggests that the "word encoding center of mass" is few words behind the current word, as if the brain would wait for more future context before encoding "fully" the word, or similarly that the current representation of the incoming word is encoded in a transient representation that is changing until the next words come in. This is coherent with previous studies from *Gwilliams et al.* that showed that the several past phonemes information (with position and order in sequence) are kept in memory [36], and that current incoming word lexical information is retrieved in a context-sensitive manner (rather than using the most probable lexical category of the word) [37].

We hypothesise that such "encoding center of mass" lying in the past is also what is happening in the speaker's brain. Songbirds such as canaries need to keep track of long-time dependencies in the sequences of phrases performed in order to produce the next syllables at syntax branching points correctly [38]: the brain area managing these dependencies preferentially encodes past actions rather than future actions. Specific neuron populations preferentially encoding past actions were actually more active during the rare phrases that involve history-dependent transitions in song [38]. This is also coherent with the results of [36] where phoneme representations are sustained longer when linguistic identity is uncertain. Overall, it seems that the representations of past events or actions are kept in memory until they have been used to disambiguate future events/actions.

7. References

[1] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell, "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses," *PloS one*, vol. 9, no. 11, p. e112575, 2014.

[2] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[3] C. Caucheteux, A. Gramfort, and J.-R. King, "Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects," *arXiv:2110.06078*, 2021.

[4] A. J. Reddy and L. Wehbe, "Can fmri reveal the representation of syntactic structure in the brain?" *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[5] X. Zhang, S. Wang, N. Lin, J. Zhang, and C. Zong, "Probing word syntactic representations in the brain by a feature elimination method," in *AAAI*, 2022.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] M. Toneva and L. Wehbe, "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)," *arXiv:1905.11833*, 2019.

[8] M. Schrimpf, I. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. Tenenbaum, and E. Fedorenko, "The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing," *BioRxiv*, 2020.

[9] S. R. Oota, J. Arora, V. Agarwal, M. Marreddy, M. Gupta, and R. S. Bapi, "Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity?" in *NAACL*, 2022.

[10] S. R. Oota, F. Alexandre, and X. Hinaut, "Long-term plausibility of language models and neural dynamics during narrative listening," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, no. 44, 2022.

[11] S. R. Oota, M. Gupta, and M. Toneva, "Joint processing of linguistic properties in brains and language models," *arXiv preprint arXiv:2212.08094*, 2022.

[12] G. Flick and L. Pylkkänen, "Isolating syntax in natural language: Meg evidence for an early contribution of left posterior temporal cortex," *Cortex*, vol. 127, pp. 42–57, 2020.

[13] A. R. Kochari, A. G. Lewis, J.-M. Schoffelen, and H. Schriefers, "Semantic and syntactic composition of minimal adjective-noun phrases in dutch: An meg study," *Neuropsychologia*, vol. 155, p. 107754, 2021.

[14] R. Law and L. Pylkkänen, "Lists with and without syntax: A new approach to measuring the neural processing of syntax," *Journal of Neuroscience*, vol. 41, no. 10, pp. 2186–2196, 2021.

[15] M. Toneva, O. Stretcu, B. Póczos, L. Wehbe, and T. M. Mitchell, "Modeling task effects on meaning representation in the brain via zero-shot meg prediction," *NeurIPS*, vol. 33, 2020.

[16] A. Caramazza and E. B. Zurif, "Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia," *Brain and language*, vol. 3, no. 4, pp. 572–582, 1976.

[17] A. D. Friederici, C. J. Fiebach, M. Schlesewsky, I. D. Bornkessel, and D. Y. Von Cramon, "Processing linguistic complexity and grammaticality in the left frontal cortex," *Cerebral Cortex*, vol. 16, no. 12, pp. 1709–1717, 2006.

[18] A. D. Friederici, "The brain basis of language processing: from structure to function," *Physiological reviews*, vol. 91, no. 4, pp. 1357–1392, 2011.

[19] E. Zaccarella and A. D. Friederici, "Merge in the human brain: A sub-region based functional investigation in the left pars opercularis," *Frontiers in psychology*, vol. 6, p. 1818, 2015.

[20] C. Humphries, J. R. Binder, D. A. Medler, and E. Liebenthal, "Syntactic and semantic modulation of neural activity during auditory sentence comprehension," *Journal of cognitive neuroscience*, vol. 18, no. 4, pp. 665–679, 2006.

[21] C. Rogalsky and G. Hickok, "Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex," *Cerebral Cortex*, vol. 19, no. 4, pp. 786–796, 2009.

[22] D. K. Bemis and L. Pylkkänen, "Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases," *Journal of Neuroscience*, vol. 31, no. 8, pp. 2801–2814, 2011.

[23] C. Caucheteux, A. Gramfort, and J.-R. King, "Disentangling syntax and semantics in the brain with deep networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1336–1348.

[24] S. R. Oota, M. Marreddy, M. Gupta, and B. R. Surampud, "Syntactic structure processing in the brain while listening," *arXiv preprint arXiv:2302.08589*, 2023.

[25] S. R. Oota, N. Manwani, and R. S. Bapi, "fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings," in *ICONIP*. Springer, 2018, pp. 3–15.

[26] S. Jain and A. G. Huth, "Incorporating context into language encoding models for fmri," in *NIPS*, 2018, pp. 6629–6638.

[27] N. Hollenstein, A. de la Torre, N. Langer, and C. Zhang, "Cognival: A framework for cognitive word embedding evaluation," in *CoNLL*, 2019, pp. 538–549.

[28] A. R. Vaidya, S. Jain, and A. G. Huth, "Self-supervised models of audio effectively explain human cortical responses to speech," 2022. [Online]. Available: https://arxiv.org/abs/2205.14252

[29] S. Wang, J. Zhang, N. Lin, and C. Zong, "Probing brain activation patterns by dissociating semantics and syntax in sentences," in *AAAI*, vol. 34, 2020, pp. 9201–9208.

[30] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[31] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[32] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[33] L. Gwilliams, G. Flick, A. Marantz, L. Pylkkanen, D. Poeppel, and J.-R. King, "Meg-masc: a high-quality magneto-encephalography dataset for evaluating natural speech processing," *arXiv preprint arXiv:2208.11488*, 2022.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[36] L. Gwilliams, J.-R. King, A. Marantz, and D. Poeppel, "Neural dynamics of phoneme sequences reveal position-invariant code for content and order," *Nature communications*, vol. 13, no. 1, p. 6606, 2022.

[37] L. Gwilliams, A. Marantz, D. Poeppel, and J.-R. King, "Top-down information shapes lexical processing when listening to continuous speech," *Language, Cognition and Neuroscience*, pp. 1–14, 2023.

[38] Y. Cohen, J. Shen, D. Semu, D. P. Leman, W. A. Liberti III, L. N. Perkins, D. C. Liberti, D. N. Kotton, and T. J. Gardner, "Hidden neural states underlie canary song syntax," *Nature*, vol. 582, no. 7813, pp. 539–544, 2020.