Syntactic representations in the human brain: beyond effort-based metrics

Aniketh Janardhan Reddy[1] & Leila Wehbe [1,2,*]

[1] Machine Learning Department, Carnegie Mellon University
[2] Neuroscience Institute, Carnegie Mellon University
[*] to whom correspondance should be addressed: lwehbe@cmu.edu

Acknowledgments:
LW and AJR were supported by startup funds at Carnegie Mellon University. The authors would like to thank Jennifer Williams for useful feedback on the manuscript.

Author Contributions: AJR and LW designed the research. AJR wrote code and ran the analytical experiments. The data was previously collected and made public by LW. AJR and LW analysed the results. AJR and LW wrote the paper.

Data availability statement: The fMRI data used in this study has already been published as part of a previous manuscript. The version we use in this study is available at https://drive.google.com/file/d/1CtGpSrxsueilF0vTyv7PiCu3vN6AuJ3Z/view?usp=sharing. In a further revision, this version will be uploaded on KiltHub, a research repository provided by the Carnegie Mellon University Libraries.

Code availability statement: The code used to perform our analyses can be found at https://github.com/anikethjr/brain_syntactic_representations. The repository also contains detailed instructions on how to run our code.

Abstract

We are far from having a complete mechanistic understanding of the brain computations involved in language processing and of the role that syntax plays in those computations. Most language studies do not computationally model syntactic structure and most studies that do model syntactic processing use effort-based metrics. These metrics capture the effort needed to process the syntactic information given by every word. They can reveal where in the brain syntactic processing occurs, but not what features of syntax are processed by different brain regions. Here, we move beyond effort-based metrics and propose explicit features capturing the syntactic structure that is incrementally built while a sentence is being read. Using these features and functional Magnetic Resonance Imaging (fMRI) recordings of participants reading a natural text, we study the brain representation of syntax. We find that our syntactic structure-based features are better than effort-based metrics at predicting brain activity in various parts of the language system. We show evidence of the brain representation of complex syntactic information such as phrase and clause structures. We see that regions well-predicted by syntactic features are distributed in the language system and are not distinguishable from those processing semantics. Our results call for a shift in the approach used for studying syntactic processing.

Keywords: fMRI, syntax, encoding models, brain representations, naturalistic experiments

Neuroscientists have long been interested in how the brain processes syntax. To date, there is no consensus on which brain regions are involved in processing it. Classically, only a small number of regions in the left hemisphere were thought to be involved in language processing. More recently, the language system was proposed to involve a set of brain regions spanning the left and right hemisphere (Fedorenko & Thompson-Schill, 2014). Similarly, some findings show that syntax is constrained to specific brain regions (Friederici, 2011; Grodzinsky & Friederici, 2006), while other findings show syntax is distributed throughout the language system (Blank et al., 2016; Fedorenko et al., 2012; 2020).

The biological basis of syntax was first explored through studies of the impact of brain lesions on language comprehension or production (Grodzinsky, 2000) and later through non-invasive neuroimaging experiments that record brain activity while subjects perform language tasks, using methods such as functional Magnetic Resonance Imaging (fMRI) or electroencephalography (EEG). These experiments usually isolate syntactic processing by contrasting the activity between a difficult syntactic condition and an easier one and by identifying regions that increase in activity with syntactic effort (Friederici, 2011). An example of these conditions is reading a sentence with an object-relative clause (e.g. "The rat *that the cat chased* was tired"), which is more taxing than reading a sentence with a subject-relative clause (e.g. "The cat *that chased the rat* was tired"). In the past decade, this approach was extended to study syntactic processing in naturalistic settings such as when reading or listening to a story (Brennan et al., 2012; Hale et al., 2018; Willems et al., 2015). Because such complex material is not organized into conditions, neuroscientists have instead devised effort-based metrics capturing the word-by-word evolving syntactic demands required

to understand the material. Brain regions with activity correlated with those metrics are suggested to be involved in processing syntax.

We use the term effort-based metrics to refer to uni-dimensional measures capturing word-by-word syntactic demands. A standard approach for constructing a syntactic effort-based metric is to assume a sentence's syntactic representation and estimate the number of syntactic operations performed at each word. Node Count is popular such metric. It relies on constituency trees (structures that capture the hierarchical grammatical relationship between the words in a sentence). While traversing the words of the sentence in order, subtrees of the constituency tree get completed; Node Count refers to the number of such subtrees that get completed at each word, effectively capturing syntactic load or effort. Brennan et al. (2012) use Node Count to support the theory that the Anterior Temporal Lobe (ATL) is involved in syntactic processing. Another example of an effort-based metric is given by an EEG study by Hale et al. (2018). They show that parser action count (the number of possible actions a parser can take at each word) is predictive of the P600, a positive peak in the brain's electrical activity occurring around 600ms after word onset. The P600 is hypothesized to be driven by syntactic processing (to resolve incongruencies), and the results of Hale et al. (2018) align with this hypothesis.

Though effort-based metrics are a good proposal for capturing the effort involved in integrating a word into the syntactic structure of a sentence, they are not reflective of the entire syntactic information in play. Hence, these metrics cannot be used to study the brain representation of syntactic constructs such as nouns, verbs, relationships and dependencies between words, and the complex hierarchical structure of phrases and sentences.

Constituency trees and dependency trees are the two main structures that capture a sentence's syntactic structure. Constituency trees are derived using phrase structure grammars that encode valid phrase and clause structure (see Figure 1(A) for an example). Dependency trees encode relations between pairs of words such as subject-verb relationships. We use representations derived from both types of trees. We derive word-level dependency role (DEP) labels from dependency trees, and we focus on encoding the structural information given by constituency trees since we want to analyze if the brain builds hierarchical representations of phrase structure. We characterize the syntactic structure inherent in sentence constituency trees by computing an evolving vector representation of the syntactic structure processed at each word using the subgraph embedding algorithm by Adhikari et al. (2018).

We show that our subgraph embedding-based syntactic structure embeddings – along with other simpler syntactic structure embeddings built using conventional syntactic features such as part-of-speech (POS) tags and DEP tags – are better than effort-based metrics at predicting the fMRI data of subjects reading text. This indicates that representations of syntax, and not just syntactic effort, can be observed in fMRI. Thus, we conclude that such syntactic structure embeddings can complement effort-based metrics in future studies.

We also address the important question of whether regions that are predicted by syntactic features are selective for syntax, meaning they are only responsive to syntax and not to

other language properties such as semantics. To answer this question, we model the semantic properties of words using a contextual word embedding space (Devlin et al., 2018). We find that regions that are predicted by syntactic features are also predicted by semantic features and thus are not selective for syntax.

1 Results

In order to uncover the brain areas that process syntax, we build many feature spaces. Then, we analyze their effectiveness in predicting brain activity recorded using fMRI while subjects read a book chapter one word at a time. These features are incremental, meaning that they are computed per word. We start with a simple *punctuation* feature space that indicates the type of punctuation shown along with a word (if any). While punctuation is seldom considered a syntactic feature, sentence boundaries or pauses within a sentence are highly correlated with changes in working memory load. These changes are bound to be a great source of variability in the fMRI signal (as we will observe later). Hence, this feature space is used to measure the baseline level of prediction performance. Failing to account for sentence boundaries and working memory might be a source of confounding that has been ignored in the literature. Then, we compute different *effort-based metrics* by relying on previous work to identify important variables. These metrics include Node Count, Syntactic Surprisal, Word Frequency and Word Length (see methods for a discussion). In order to capture the actual syntactic information inherent in sentences, beyond effort, we then construct various syntactic structure embeddings. We first construct a simple syntactic structure embedding that encodes the *part-of-speech and dependency role tags* of each word.

Next, we propose three new syntactic structure embeddings intended to capture the hierarchical syntactic structure of a sentence and how multiple words are composed into phrases and clauses. These features are computed as the subgraph embeddings of various subtrees of the constituency tree that contains a given word. The embeddings are computed using the techniques described by Adhikari et al. (2018). Each of the three embeddings capture different aspects of the evolving syntactic structure that is built at each word. The first set of embeddings, which we call *ConTreGE Comp* vectors (Constituency Tree-based Graph Embeddings, Complete), embed the largest subtree that is completed upon incorporating a given word into its sentence's constituency tree. This subtree contains the implicit syntactic information that is given upon reading a word. The second set of embeddings, referred to as *ConTreGE Incomp* vectors (Constituency Tree-based Graph Embeddings, Incomplete), encode incomplete subtrees of the sentence's constituency tree. These subtrees are built by retaining all the phrase structure grammar productions that are necessary to derive the words seen till then and thus contain non-terminal symbols as leaves. Encoding these subtrees in the ConTreGE vectors endows them with future syntactic information that has not yet been seen. For the third set of embeddings, we use the incremental top-down parser by Roark (2001) to generate the most probable partial parse trees considering the words seen so far, and use these to produce what we refer to as *InConTreGE* vectors (Incremental Constituency Tree-based Graph Embeddings). We collectively refer to these three syntactic structure embeddings as

the ConTreGE class of vectors. Figure 1 illustrates the different types of subtrees.

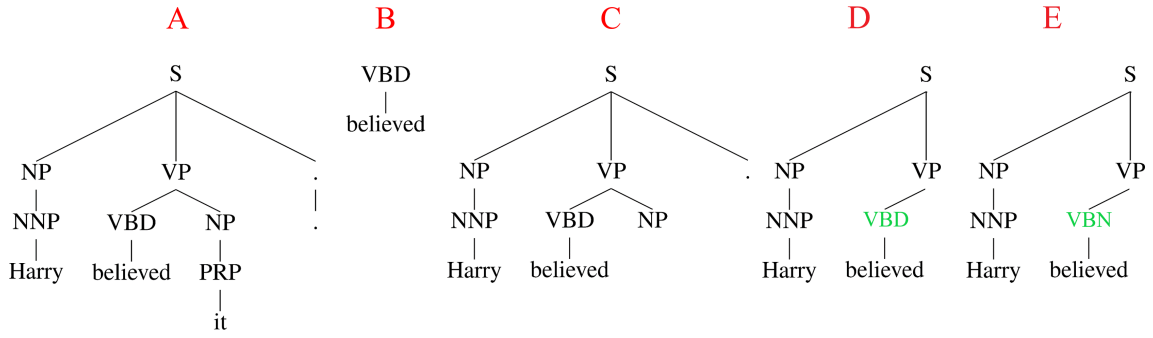


Figure 1: Example of complete and incomplete subtrees and two possible partial parses: Part A shows a sentence's constituency tree generated by a self-attentive encoder-based constituency parser (Kitaev & Klein, 2018) using all of its words. The largest completed subtree for "believed" is shown in part B and the incomplete subtree generated till "believed" is shown in part C. Incomplete subtrees are generally much deeper than complete ones. In parts D and E, we can see two possible partial parses generated by an incremental top-down parser (Roark, 2001) only using the words till "believed". We see that the POS tag assigned to "believed" is different in the two parses.

The ConTreGE vectors can be used to test different hypotheses. Brain regions that are sensitive to ConTreGE Comp could be involved in processing more local (i.e. more word specific) syntactic information. If the ConTreGE Incomp vectors are predictive of brain activity, this could mean that the brain anticipates some amount of future syntactic information. Finally, if the InConTreGE vectors are good at predicting the data, it might be the case that the brain constructs multiple possible partial parses while reading a sentence.

Finally, in order to determine if syntax and semantics are processed in similar regions, we use contextual word embeddings from *BERT*, a transformer-based language model (Devlin et al., 2018). We then use different sets of feature spaces as inputs to encoding models that predict the activity in each fMRI voxel (see methods for details). Many of our feature spaces have overlapping information. POS and DEP tags include punctuation, BERT vectors have been shown to encode syntactic information (Hewitt & Manning, 2019) and ConTreGE vectors, built from constituency trees, encode some POS tags information. To detect brain regions sensitive to the distinct information given by a feature space, we build hierarchical feature groups in increasing order of syntactic information and test for significant differences in prediction performance between two consecutive groups. We start with the simplest feature – punctuation, and then add more complex features in order: the effort-based metrics, POS and DEP tags, one of the ConTreGE vectors and the vectors derived from BERT (which can be thought of as a super-set of semantics and syntax). At each step, we test if the introduction of the new feature space leads to significantly larger than chance improvement in R^2. Figures 2 and 3 summarize our results (Appendix A has the raw prediction results and Appendix C contains detailed plots for every ROI showing the subject-specific values).

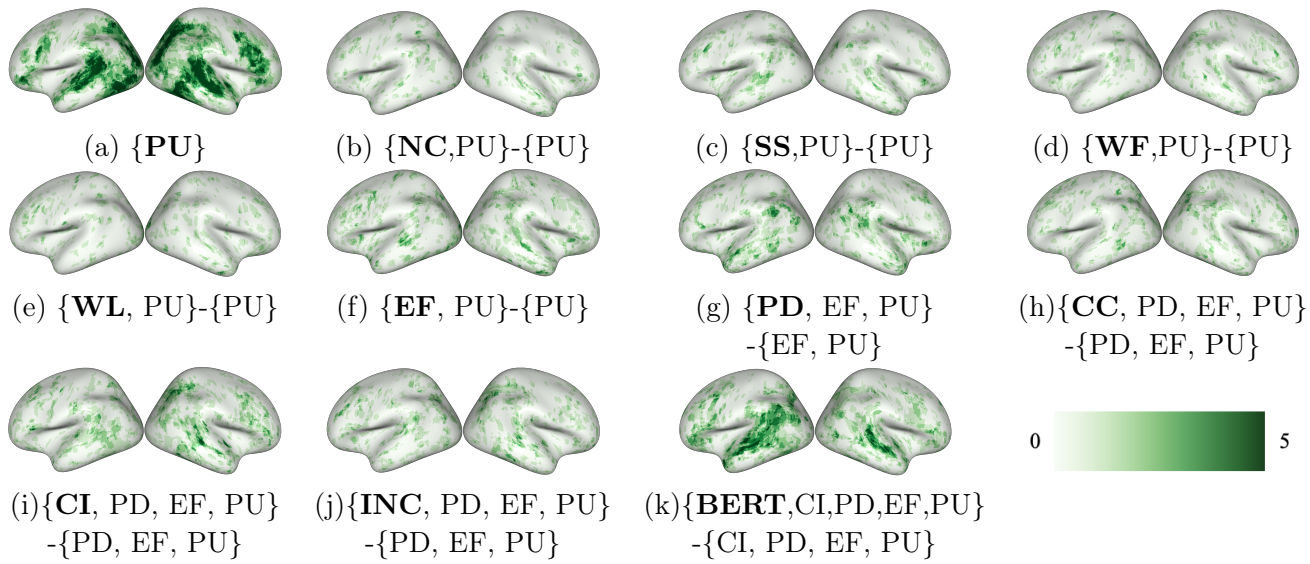


Figure 2: The first plot shows the number of subjects for which a given voxel is significantly predicted by punctuation ($p \leq 0.05$). The others show the number of subjects for which the difference in R^2 scores between two feature groups is significant ($p \leq 0.05$). Here, PU = Punctuation, NC = Node Count, SS = Syntactic Surprisal, WF = Word Frequency, WL = Word Length, EF = All effort-based metrics, PD = POS and DEP Tags, CC = ConTreGE Comp, CI = ConTreGE Incomp, INC = InConTreGE, BERT = BERT embeddings and '{,}' indicates that these features were concatenated in order to make the predictions. '-' indicates a hypothesis test for the difference in R^2 scores between the two feature groups being larger than 0. The distinct information given by syntactic structure-based features is more predictive of brain activity than that given by effort-based metrics. The semantic vectors are also very predictive and many well-predicted regions overlap with those that are predicted by syntax.

1.1 Syntactic structure embeddings are more predictive of brain activity than effort-based metrics

Figures 2 (b)-(e) show that there are a small number of voxels that are predicted by the effort based metrics when taken in isolation. Figures 2 (f)-(j) indicate that although the information provided by the effort metrics combined is predictive of brain activity to some degree (when controlling for punctuation), there is still a considerable amount of structural information that is contained in the POS and DEP tags and in ConTreGE Incomp that predict additional portions of the activity. These results are made even clearer by Figure 3. Many voxels have significant increase in the R^2 scores (above what is predicted by the effort metrics) after including POS and DEP tags and ConTreGE Incomp. We also notice that ConTreGE Comp is not as predictive as ConTreGE Incomp, hinting that future syntactic information helps in predicting current brain activity. Additionally, InConTreGE is not as predictive as ConTreGE Incomp, suggesting that the top down parser might be generating partial parses that are not reflective of brain representations.

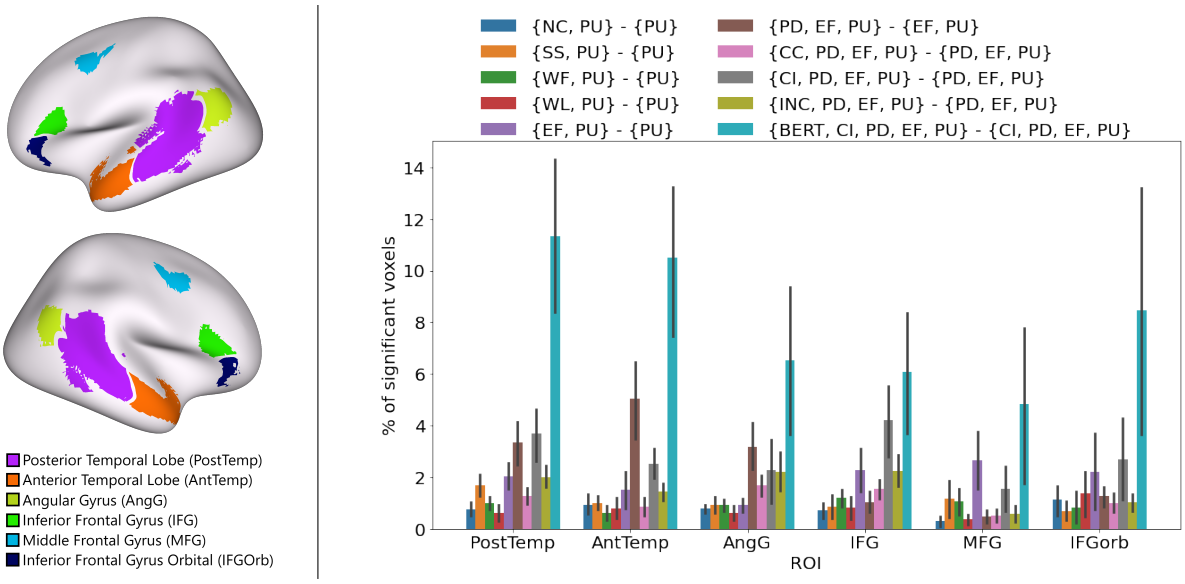


Figure 3: Region of Interest (ROI) analysis of the prediction performance. [Left] Language system ROIs by Fedorenko et al. (2010) from (3). [Right] Percentage of significantly predicted ROI voxels. Each bar represents the average percentage across subjects and the error bars show the standard error across subjects. We use the same abbreviations as in Figure 2 and see the same trends across ROIs.

1.2 ConTreGE Incomp results suggest that complex syntactic information is encoded in the brain

In this section we analyze the information in ConTreGE Incomp to interpret its brain prediction performance. We estimate how much of the constituency tree is captured by each feature by using it to predict the level N ancestor of a word (in its constituency tree). We vary N from 2 to 9 and train a logistic regression model for each N. Since POS tags are the level 1 ancestors of words, we start the analysis at N=2. Because there are many phrase labels, we group them into 7 larger buckets - noun phrases, verb phrases, adverb phrases, adjective phrases, prepositional phrases, clauses and other miscellaneous labels. Also, if a word's depth in its tree is less than N, the root is considered its level N ancestor.

Table 1 shows the results of this analysis. We use the constituency trees generated by the Kitaev & Klein (2018) parser. Given the skewed label distribution, the optimal strategy for a predictor that takes random noise as input is to always output the majority class ancestor at that level. Chance performance is thus equal to the frequency of the majority label. The effort-based metrics are not as predictive as ConTreGE Incomp at any level. POS and DEP tags are predictive of labels at all levels and produce the highest accuracies for lower levels. The InConTreGE vectors are not as predictive as ConTreGE Incomp or ConTreGE Comp, hinting that the top down parser might not be very accurate. ConTreGE Incomp is the best predictor of higher level ancestors but ConTreGE Comp is better than ConTreGE Incomp at predicting lower level ancestors. This may be because graph embeddings of a

Feature	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9
Most Popular Label %	51	38.76	54.42	64.05	73.44	78.25	82.38	85.82
Node Count	51	42.6*	55.45*	64.01	73.4	78.23	82.4	85.82
Syntactic Surprisal	50.21	41.48*	54.42	64.05	73.44	78.25	82.38	85.82
Word Frequency	51	41.56*	57.13	66.58	76*	80.68	84.78	88.16
Word Length	50.99	39.18	54.42	64.22	73.44	78.28	83.31	88.68*
POS and DEP tags	**92.23***	**71.27***	**67.06***	**70***	**77.51***	82.09*	86.26*	89.59*
ConTreGE Comp	66.76*	51.01*	59.52*	67.95*	77.28*	82.08*	86.26*	89.68*
ConTreGE Incomp	52.91	45.42*	57.37*	67.29	76.87*	**82.3***	**86.51***	**90.39***
InConTreGE	52.46	45.59*	57.61*	66.75	76.13	81.08	85.27*	89.1*
BERT Embeddings	52.18	45.73*	58.62*	66.79	75.68	80.31	84.72	88.72

Table 1: 10-fold cross validation accuracies in predicting the ancestors of a given word. * denotes accuracies significantly above chance (tested using Wilcoxon signed-rank test, $p \leq 0.01$). POS and DEP tags best predict lower level ancestors while ConTreGE Incomp vectors best predict higher level ones.

tree tend to capture more of the information near the tree's root (a random walk through a somewhat balanced tree is likely to contain more occurrences of nodes near the root and random walks are encoded in the subgraph embedding generation process). ConTreGE Comp vectors, created from shallow complete trees, likely over-represent lower level ancestors while ConTreGE Incomp vectors, created from relatively deeper trees, likely over-represent higher level ancestors. Given that ConTreGE Incomp is predictive of brain activity and contains information about the higher level ancestors of a word, this suggests that the brain represents complex hierarchical syntactic information such as phrase and clause structure.

1.3 Syntax and semantics are processed in a distributed way in overlapping regions across the language system

Our results indicate that syntactic and semantic information are processed in a distributed fashion across the language network. Most of the regions in the language system are better predicted by the BERT embeddings after controlling for all our other feature spaces, and these regions overlap with the regions that are predicted by the syntactic feature spaces. While the BERT embeddings include both semantic and syntactic information, it is likely that the semantic information is at least partially predictive of the brain activity, given that we have already controlled for a lot of syntactic information.

2 Discussion and Related Work

2.1 Syntactic representations

Apart from Brennan et al. (2012) and Hale et al. (2018), many others (Boston et al., 2008; Brennan et al., 2016; Frank et al., 2015; Henderson et al., 2016; Willems et al., 2015) use effort-based metrics to study syntactic processing during natural reading or listening. However, a few studies do explicitly encode syntactic structure: Wehbe et al. (2014) find that POS and DEP tags are the most predictive out of a set of word, sentence and discourse-level features. Moving away from popular approaches that are dependent on effort-based metrics, we extended the work of Wehbe et al. (2014) by developing a novel graph embeddings-based approach to explicitly capture the syntactic information provided by constituency trees. Our results show that these explicit features have substantially more information that is predictive of brain activity than effort based metrics. Given these results, we believe that future work in this area should supplement effort-based metrics with features that explicitly encode syntactic structure.

2.2 Syntax in the brain

Traditionally, studies have associated a small number of brain regions, usually in the left hemisphere, with syntactic processing. These include parts of the inferior frontal gyrus (IFG), ATL and Posterior Temporal Lobe (PTL) (Friederici, 2011; Friederici et al., 2003; Grodzinsky & Friederici, 2006; Matchin & Hickok, 2020). However, some works point to syntactic processing being distributed across the language system. Blank et al. (2016) shows that significant differences in the activities of most of the language system are greater when reading hard to parse sentences than easier phrases.Wehbe et al. (2014) use POS and DEP tags to arrive at similar conclusions.

Previous work generally did not use naturalistic stimuli to study syntax. Instead, subjects are usually presented with sentences or even short phrases that have subtle syntactic variations or violations. Regions with activity well correlated with the presentation of such variations/violations are thought to process syntax (Friederici, 2011). Observations from such studies have limited scope since these variations often cannot be representative of the wide range of variations seen in natural language. This is possibly why such studies report specific regions: it might be that the reported region is particularly sensitive to the exact conditions used. By using one type of stimulus which evokes only one aspect of syntactic processing, syntax might appear more localized than it really is. Our results support the hypothesis that it is instead processed in a distributed fashion across the language system. We believe that our results have a wider applicability since we use naturalistic stimuli and we leave for future work the study of whether different syntactic computations are delegated to different regions.

Some studies have also doubted the importance of syntactic composition for the brain.

Pylkkänen (2020) proposes that there is no conclusive evidence to indicate that the brain puts a lot of weight on syntactic composition, and that even though studies (some with effort-based metrics) have associated certain regions like the left ATL with syntactic processing, numerous studies have later shown that the left ATL might instead be involved in a more conceptually driven process. Gauthier & Levy (2019) showed that BERT embeddings which were fine-tuned on tasks that removed dependency tree-based syntactic information were more reflective of brain activity than those which contained this information. In contrast, our work uses purely syntactic embeddings to show that we can indeed significantly predict many regions of the language system. We attribute these differences in conclusions to our naturalistic stimuli and word-by-word evolving representations of syntax. Pylkkänen (2020)'s conclusions are mostly based on studies that present a phrase with just two words (like "red boat"). Gauthier & Levy (2019) use data averaged over entire sentences instead of modeling word-by-word comprehension. Since the syntactic structure of a sentence evolves with every word that is read, this approach is not necessarily adept at capturing such information.

Furthermore, our analysis of the syntactic information contained in various features highlighted that our ConTreGE Incomp vectors are good at encoding complex phrase or clause-level syntactic information whereas POS and DEP tags are good at encoding local word-level syntactic information. Several regions of the brain's language system were predicted by ConTreGE Incomp, hinting that the brain does indeed encode complex syntactic information. Another potentially interesting observation is that including ConTreGE Incomp increases prediction performance in the PTL and IFG by more than when we include POS and DEP tags (Figure 3) but not for the ATL and the Angular Gyrus (AG). These observations very loosely support the theory by Matchin & Hickok (2020) - that parts of the PTL are involved in hierarchical lexical-syntactic structure building, the ATL is a knowledge store of entities and the AG is a store of thematic relations between entities. This is because ConTreGE Incomp encodes hierarchical syntactic information and word-level POS and DEP tags are very indicative of the presence of various entities (various types of nouns) and the thematic relations between entities (verbs associated with noun pairs). This hypothesis should be tested more formally in future work.

We also observe that ConTreGE Incomp is more predictive than ConTreGE Comp and InConTreGE with the latter two being very weakly predictive. Thus, future syntactic information appears to be very useful while predicting BOLD signals, indicating that that the brain anticipates the eventual sentence structure while reading to a more accurate extent than an incremental top down parser.

2.3 Syntactic vs. semantic processing in the brain

Finally, our results support the theory that syntax processing is distributed throughout the language network in regions that also process semantics. This theory is supported by other studies (Blank et al., 2016; Fedorenko et al., 2012; 2020). However, Friederici et al. (2003) among others argue that syntax and semantics are processed in specific and distinct regions

by localizing the effects of semantic and syntactic violations.

3 Methods

We first describe the syntactic features used in this study and their generation. All of the features we use are incremental i.e. they are computed per word. We then describe our fMRI data analyses.

3.1 Effort-based metrics

We use four effort-based metrics in our analyses - Node Count, Syntactic Surprisal, word frequency and word length. Node Count is an effort-based metric popular in neuroscience. To compute it, we obtain the constituency tree of each sentence using the self-attentive encoder-based constituency parser by Kitaev & Klein (2018). We compute Node Count for each word as the number of subtrees that are completed by incorporating this word into its sentence. Syntactic Surprisal is another effort-based metric proposed by Roark et al. (2009) and is computed using an incremental top down parser (Roark, 2001). Both of these metrics aim to measure the amount of effort that is required to integrate a word into the syntactic structure of its sentence. The word frequency metric is computed using the wordfreq package (Speer et al., 2018) as the Zipf frequency of a word. This is the base-10 logarithm of the number of occurrences per billion of a given word in a large text corpus. Finally, word length is the number of characters in the presented word. The last two metrics approximate the amount of effort that is required to read a word.

3.2 Constituency Tree-based Graph Embeddings (ConTreGE)

Constituency trees are a rich source of syntactic information. We build three representations of these trees that encode this information:

1. The largest subtree which is completed upon incorporating a word into a sentence (see figure 1(B)) is representative of the implicit syntactic information given by the word. Given that Node Count reduces all of the information present in these subtrees to just one number, it is easy to see that it cannot effectively capture this information. POS tags (categorize words into nouns, verbs, adjectives, etc.) also capture some of the information present in these trees as they encode phrase structure to a certain extent. But, they are incapable of completely encoding their hierarchical structure and the parsing decisions which are made while generating them. In order to better encode their structure, we first build subgraph embeddings of these completed subtrees called ConTreGE Comp vectors.

2. We hypothesize that the brain not only processes structure seen thus far but also predicts future structure from structure it already knows. To test this, we construct embeddings, called ConTreGE Incomp vectors, using incomplete subtrees that are constructed by retaining all the phrase structure grammar productions that are required to derive the words seen till now, thereby allowing us to capture future sentence structure (in the form of future constituents) before the full sentence is read (see figure 1 (C)). These subtrees contain leaves that are non-terminal symbols unlike complete subtrees that only have terminal symbols (words and punctuation) as leaves. In this context, a non-terminal symbol is a symbol that can be derived further using some rule in the phrase structure grammar (ex. NP, VP, etc.). If incomplete subtrees are more representative of the brain's processes, it would mean that the brain expects certain phrase structures even before the entire phrase or sentence is read. ConTreGE Comp and ConTreGE Incomp vectors need to be built using accurate constituency trees constructed using the whole sentence. Thus, we reuse the trees generated to compute Node Count to build them.

3. Further, the brain could be computing several possible top down partial parses that can derive the words seen thus far (see figures 1 (D) and (E)) and modifying the list of possible parses as future words are read. To test this hypothesis, we designed Incremental ConTreGE (InConTreGE) vectors that are representative of the most probable parses so far. For a given word, its InConTreGE vector is computed as: $v = \sum_{i=1}^{5} e^{-s_i} W_i$ where W_i is the subgraph embedding of a partial parse tree built by an incremental top-down parser (Roark, 2001) after reading the word and s_i is the score assigned to this partial parse that is inversely proportional to the parser's confidence in this tree.

To effectively capture the structure of all subtrees, we encode them using the subgraph embeddings proposed by Adhikari et al. (2018) which preserve the neighbourhood properties of subgraphs. A long fixed length random walk on a subgraph is generated to compute its embedding. Since consecutive nodes in a random walk are neighbours, a long walk can effectively inform us about the neighbourhoods of nodes in the subgraph. Each node in a walk is identified using its unique ID. So, a random walk can be interpreted as a "paragraph" where the words are the node IDs. Finally, the subgraph's embedding is computed as the Paragraph Vector (Le & Mikolov, 2014) of this paragraph that is representative of the subgraph's structure. Note that all of the subtrees of a given type (complete, incomplete or partial parse) are encoded together. This ensures that all ConTreGE Comp vectors, all ConTreGE Incomp vectors and all InConTreGE vectors are in our own spaces.

Figure 4 illustrates the subtree encoding process. First, every unique non-terminal in the subtrees is mapped to a unique number (ex. S is mapped to 1, NP is mapped to 2, etc.) and every terminal is mapped to a unique number that is representative of the order in which they were presented (the first presented token is mapped to 10000, the second token is mapped to 10001 and so on). We did not map each unique terminal to a unique number (for instance, we did not map all instances of "Harry" to one number) because a random walk through the tree could give us word co-occurrence information and thus lead to the inclusion of some semantic information in the vectors.

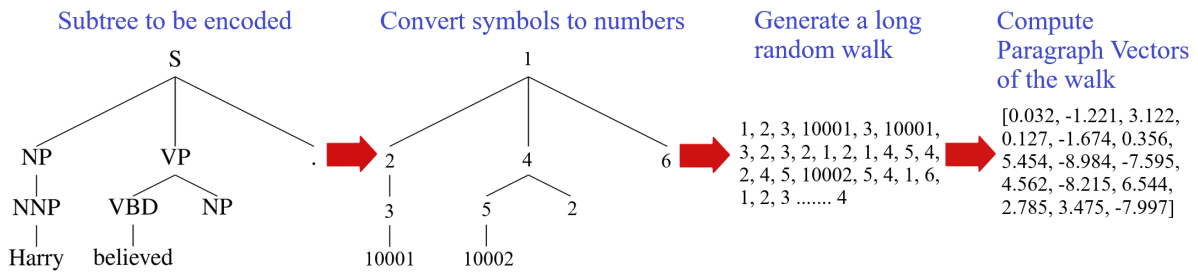


Figure 4: Steps for encoding subtrees.

Every tree node's label is then replaced by the number it was mapped to in the previous step. The edge lists of these subtrees are supplied to the subgraph embedding generation algorithm to finally obtain 15-dimensional vectors for every presented word. The length of the random walks is set to 100000 and we use an extension of the Distributed Bag of Nodes (DBON) model proposed by Le & Mikolov (2014) for generating Paragraph Vectors called Sub2Vec-DBON by Adhikari et al. (2018). The length of the sliding window is set to 5 and the model is trained for 20 epochs. Since ConTreGE Comp, ConTreGE Incomp and InConTreGE encode information about the neighbourhoods of all nodes in the constituency trees, they can capture their hierarchical structure. Thus, brain regions predicted by these vectors are likely to be involved in building and encoding hierarchical sentence structure.

3.3 Punctuation

We create one-hot binary vectors indicating the type of punctuation that was presented along with a word (e.g. **.** or **,**). For example, a sentence might have ended with "Malfoy.". In this punctuation-based feature space, the column corresponding to **.** will be set to 1 for this word.

3.4 Part-of-speech tags and dependency tags

We use two standard word-level syntactic features - POS and DEP tags. The POS tag of a word is read off previously generated constituency trees (those obtained using the Kitaev & Klein (2018) parser). The DEP tag of a word (ex. subject, object, etc.) correspond to its assigned role in the dependency trees of the presented sentences which were generated using the spaCy English dependency parser (2). We create one-hot binary vectors indicating the POS tag and the DEP tag of each word and concatenate them to create one feature space which we refer to as simple syntactic structure embeddings.

3.5 Semantic features

We adapt the vectors obtained from layer 12 of a pretrained (1) cased BERT-large model (Devlin et al., 2018) to identify regions that process semantics. We use layer 12 because of previous work showing that middle layers of sentence encoders are optimal for predicting brain activity (Jain & Huth, 2018; Toneva & Wehbe, 2019). We obtain the contextual embeddings for a word by running the pretrained model only on the words seen thus far, preventing the inclusion of future semantic information. Since a presented word can be broken up into multiple subtokens, we compute its embedding as the average of the subtokens' embeddings. Using principal component analysis (PCA), we reduce their dimensionality to 15 to match the ConTreGE vectors' dimensionality.

3.6 fMRI data

We use the fMRI data of 9 subjects reading chapter 9 of *Harry Potter and the Sorcerer's Stone* (Rowling, 2012), collected and made available by Wehbe et al. (2014). Participants gave their written informed consent and the study was approved by the Carnegie Mellon University Institutional Review Board. Words were presented one at a time at a rate of 0.5s each. All the brain plots shown here are averages over the 9 subjects in the Montreal Neurological Institute (MNI) space. Preprocessing details are in Appendix B.

3.7 Predicting brain activity

The applicability of a given syntactic feature in studying syntactic processing is determined by its efficacy in predicting the brain data described above. Ridge regression is used to perform these predictions and their coefficient of determination (R^2 score) measures the feature's efficacy. For each voxel of each subject, the regularization parameter is chosen independently. We use Ridge regression because of its computational efficiency and because of the Wehbe et al. (2015) results showing that with such fMRI data, as long as the regularization parameter is chosen by cross-validation for each voxel independently, different regularization techniques lead to similar results. Indeed, Ridge regression is a common regularization technique used for predictive fMRI models (Huth et al., 2016; Mitchell et al., 2008; Nishimoto et al., 2011; Wehbe et al., 2014).

For every voxel, a model is fit to predict the signals $Y = [y_1, y_2, \ldots, y_n]$ recorded in that voxel where n is the number of time points (TR, or time to repetition). The words are first grouped by the TR in which they were presented. Then, the features of words in every group are summed to form a sequence of features $X = [x_1, x_2, \ldots, x_n]$ aligned with the brain signals. The response measured by fMRI is an indirect consequence of brain activity that peaks about 6 seconds after stimulus onset. A common solution to account for this delay is to express brain activity as a function of the features of the preceding time points (Huth et al., 2016;

Nishimoto et al., 2011; Wehbe et al., 2014). Thus, we train our models to predict any y_i using $x_{i-1}, x_{i-2}, x_{i-3}$ and x_{i-4}.

We test the models in a cross-validation loop: the data is first split into 4 contiguous and equal sized folds. Each model uses three folds of the data for training and one fold for evaluation. We remove the data from the 5 TRs which either precede or follow the test fold from the training set of folds. This is done to avoid any unintentional data leaks since consecutive y_is are correlated with each other because of the lag and continuous nature of the fMRI signal. The brain signals and the word features which comprise the training and testing data for each model are individually Z-scored. After training we obtain the predictions for the validation fold. The predictions for all folds are concatenated (to form a prediction for the entire experiment in which each time point is predicted from a model trained without the data for that time point). Note that since all 3 ConTreGe vectors are stochastic, we construct them 5 times each, and learn a different model each time. The predictions of the 5 models are averaged together into a single prediction. The R^2 score is computed for every voxel using the predictions and the real signals.

We run a permutation test to test if R^2 scores are significantly higher than chance. We permute blocks of contiguous fMRI TRs, instead of individual TRs, to account for the slowness of the underlying hemodynamic response. We choose a common value of 10 TRs (Deniz et al., 2019). The predictions are permuted within fold 5000 times, and the resulting R^2 scores are used as an empirical distribution of chance performance, from which the p-value of the unpermuted performance is estimated. We also run a bootstrap test to test if a model has a higher R^2 score than another. The difference is that in each iteration, we permute (using the same indices) the predictions of both models and compute the difference of their R^2 and use the resulting distribution to estimate the p-value of the unpermuted difference. Finally, the Benjamni-Hochberg False Discovery Rate correction (Benjamini & Hochberg, 1995) is used for all tests (appropriate because fMRI data is considered to have positive dependence (Genovese, 2000)). The correction is performed by grouping together all the voxel-level p-values (i.e. across all subjects and feature groups) and choosing one threshold for all of our results. The correction is done in this way since we test multiple prediction models across multiple voxels and subjects. To compute Region of Interest (ROI) statistics, left-hemisphere ROI masks for the language system obtained from a “sentence vs. non-word" fMRI contrast (Fedorenko et al., 2010) are obtained from (3) and mirrored to obtain the right-hemisphere ROIs. v

References

Link 1. *BERT-Large, Cased: 24-layer, 1024-hidden, 16-heads, 340M parameters*. URL https://storage.googleapis.com/bert_models/2018_10_18/cased_L-24_H-1024_A-16.zip.

Link 2. *spaCy, en_core_web_sm model*. URL https://github.com/explosion/spacy-models/releases//tag/en_core_web_sm-2.2.5.

Link 3. *Group-level functional parcels.* URL https://evlab.mit.edu/funcloc/download-parcels.

Bijaya Adhikari, Yao Zhang, Naren Ramakrishnan, and B Aditya Prakash. Sub2vec: Feature learning for subgraphs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 170–182. Springer, 2018.

J. Ashburner, CC Chen, G. Flandin, R. Henson, S. Kiebel, J. Kilner, V. Litvak, R. Moran, W. Penny, K. Stephan, et al. SPM8 manual. *Functional Imaging Laboratory, Institute of Neurology*, 2008.

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)*, pp. 289–300, 1995.

Idan Blank, Zuzanna Balewski, Kyle Mahowald, and Evelina Fedorenko. Syntactic processing is distributed across the language system. *Neuroimage*, 127:307–323, 2016.

Marisa Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *The Mind Research Repository (beta)*, (1), 2008.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, 120(2):163–173, 2012.

Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157:81–94, 2016.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. 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, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

E. Fedorenko, A. Nieto-Castanon, and N. Kanwisher. Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, 50(4):499–513, 2012.

Evelina Fedorenko and Sharon L Thompson-Schill. Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126, 2014.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

Evelina Fedorenko, Idan Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *bioRxiv*, pp. 477851, 2020.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11, 2015.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011.

Angela D Friederici, Shirley-Ann Rüschemeyer, Anja Hahne, and Christian J Fiebach. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral cortex*, 13(2):170–177, 2003.

James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, 9:23, 2015.

Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 529–539, 2019.

Christopher R. Genovese. A Bayesian time-course model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 95:691–703, 2000.

Yosef Grodzinsky. The neurology of syntax: Language use without broca's area. *Behavioral and brain sciences*, 23(1):1–21, 2000.

Yosef Grodzinsky and Angela D Friederici. Neuroimaging of syntax and syntactic processing. *Current opinion in neurobiology*, 16(2):240–246, 2006.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. Finding syntax in human encephalography with beam search. *arXiv preprint arXiv:1806.04127*, 2018.

John M Henderson, Wonil Choi, Matthew W Lowder, and Fernanda Ferreira. Language structure in the brain: A fixation-related fmri study of syntactic surprisal in reading. *Neuroimage*, 132:293–300, 2016.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://www.aclweb.org/anthology/N19-1419.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, Jack L Gallant, Wendy a De Heer, Thomas L Griffiths, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. doi: 10.1038/nature17637.Natural.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *Advances in neural information processing systems*, pp. 6628–6637, 2018.

Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014. URL http://arxiv.org/abs/1405.4053.

William Matchin and Gregory Hickok. The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498, 2020.

T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320 (5880):1191–1195, 2008.

S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 2011.

Liina Pylkkänen. Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B*, 375(1791): 20190299, 2020.

Brian Roark. Probabilistic top-down parsing and language modeling. *Computational linguistics*, 27(2):249–276, 2001.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 324–333, 2009.

J.K. Rowling. *Harry Potter and the Sorcerer's Stone*. Harry Potter US. Pottermore Limited, 2012. ISBN 9781781100271. URL http://books.google.com/books?id=wrOQLV6xB-wC.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. Luminosoinsight/wordfreq: v2.2, October 2018. URL https://doi.org/10.5281/zenodo.1443582.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pp. 14928–14938, 2019.

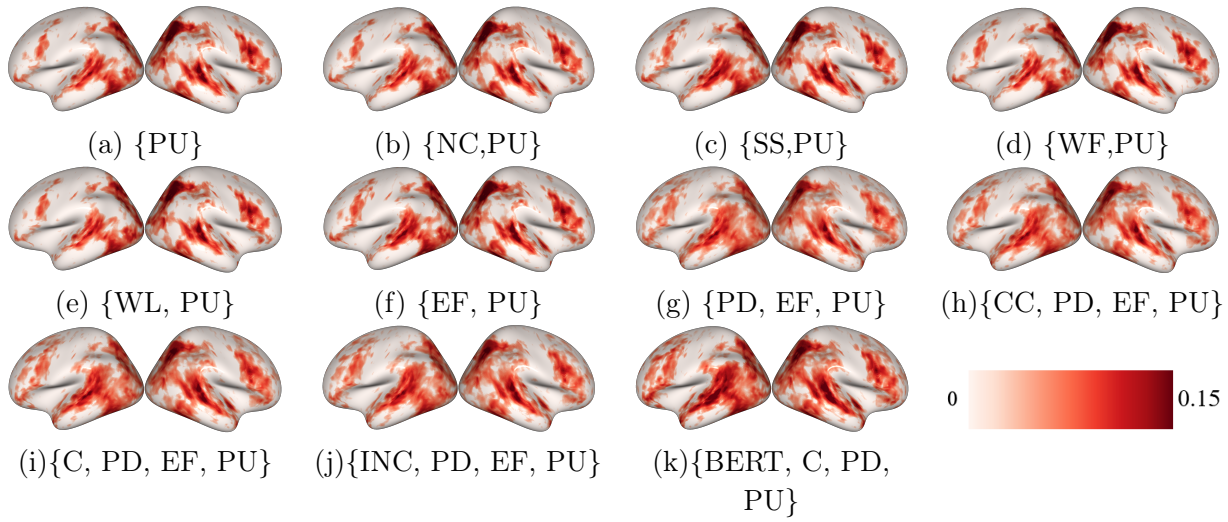
Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading Subprocesses. *PloS one*, 9(11):e112575, nov 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL http://dx.plos.org/10.1371/journal.pone.0112575.

Leila Wehbe, Aaditya Ramdas, Rebecca C Steorts, Cosma Rohilla Shalizi, et al. Regularized brain reading with shrinkage and smoothing. *The Annals of Applied Statistics*, 9(4): 1997–2022, 2015.

Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal Van den Bosch. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516, 2015.

A Raw prediction results

Figure 5 shows the prediction results obtained using each feature group. To be able to better judge different levels of accuracy, instead of looking at the R^2 scores, we compute R^{2+}, in which we replace the positive R^2 values by their squared root, making them easier to resolve visually, and the negative ones with 0.



(a) {PU} (b) {NC,PU} (c) {SS,PU} (d) {WF,PU}

(e) {WL, PU} (f) {EF, PU} (g) {PD, EF, PU} (h){CC, PD, EF, PU}

(i){C, PD, EF, PU} (j){INC, PD, EF, PU} (k){BERT, C, PD, PU}

Figure 5: Cross-subject prediction performance of all syntactic feature groups. The figures show cross-subject average R^{2+} scores. Here, PU = Punctuation, NC = Node Count, SS = Syntactic Surprisal, WF = Word Frequency, WL = Word Length, EF = All effort-based metrics, PD = POS and DEP Tags, CC = ConTreGE Comp, C = ConTreGE Incomp, INC = InConTreGE, BERT = BERT embeddings and '{,}' indicates that these features were concatenated in order to make the predictions.

B Acquiring and preprocessing the fMRI data

We obtained the raw data from Wehbe et al. (2014). This fMRI data is acquired at a rate of 2s per image and comprise $3 \times 3 \times 3mm$ voxels. The data for each subject is slice-time and motion corrected using SPM8 (Ashburner et al., 2008), then detrended and smoothed with an isotropic spherical Gaussian kernel with a standard deviation of $3mm$. The brain surface of each subject is reconstructed using Freesurfer (Fischl, 2012) and a grey matter mask is obtained. Pycortex (Gao et al., 2015) is used to handle and plot the data. All subject results are converted to MNI space using pycortex.

C ROI Analysis Plots

In this section, we show detailed plots from the ROI analysis of prediction performance. Each plot corresponds to one ROI from Figure 3. Like in Figure 3, each bar represents the average percentage of significantly predicted ROI voxels across subjects and the error bars show the standard error across subjects. The labels for the bars denote the feature that was added while performing the significance test. Thus, the labels used here correspond to the following labels used in Figures 2 and 3:

1. NC = {NC, PU} - {PU}
2. SS = {SS, PU} - {PU}
3. WF = {WF, PU} - {PU}
4. WL = {WL, PU} - {PU}
5. EF = {EF, PU} - {PU}
6. PD = {PD, EF, PU} - {EF, PU}
7. CC = {CC, PD, EF, PU} - {PD, EF, PU}
8. CI = {CI, PD, EF, PU} - {PD, EF, PU}
9. INC = {INC, PD, EF, PU} - {PD, EF, PU}
10. BERT = {BERT, CI, PD, EF, PU} - {CI, PD, EF, PU}

Each red point then shows the percentage of ROI voxels that were significantly predicted, according to the test specified by the bar label, in a given subject.

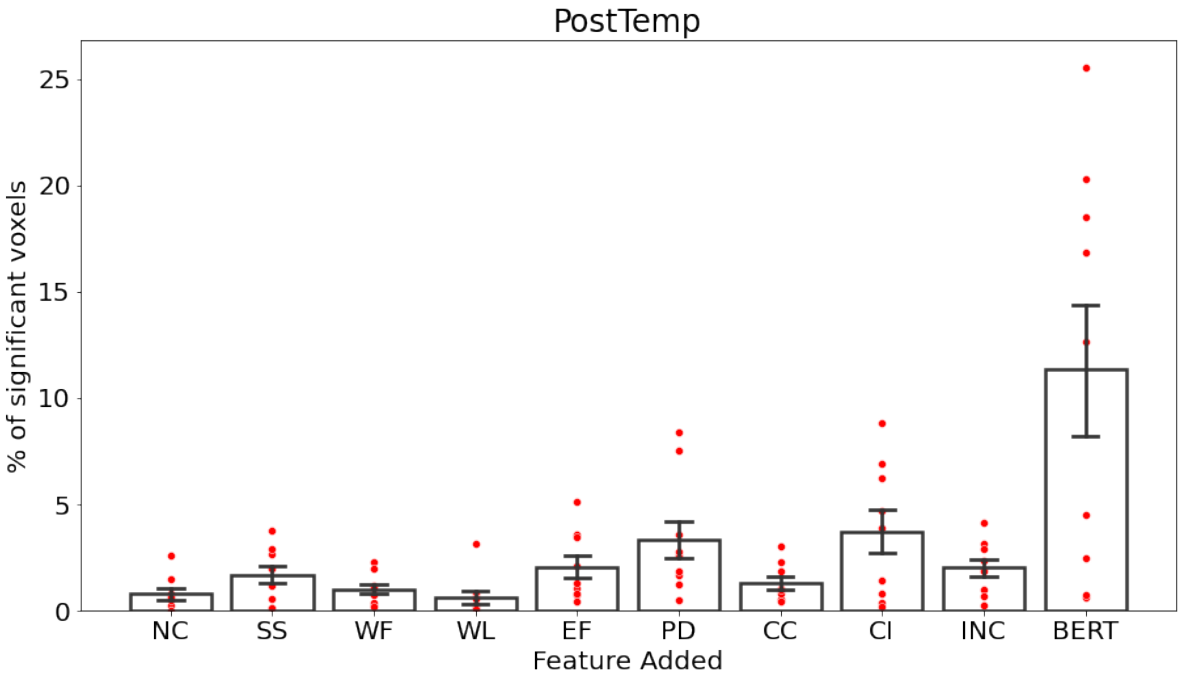


Figure 6: Results for the Posterior Temporal Lobe

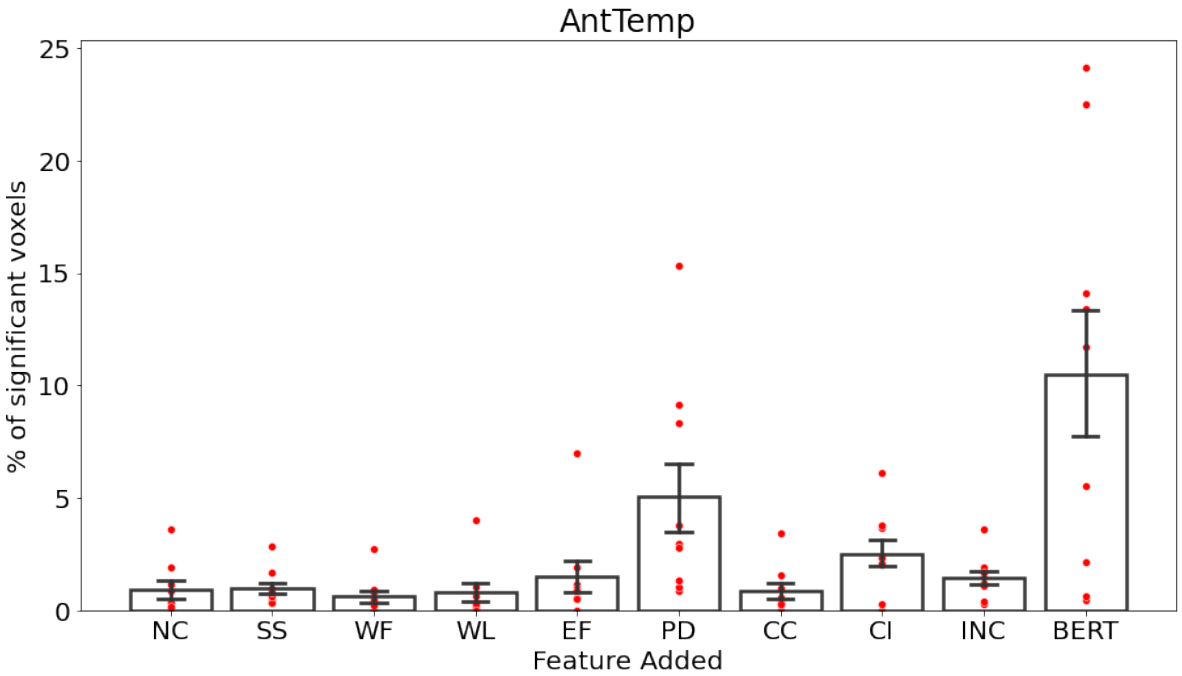


Figure 7: Results for the Anterior Temporal Lobe

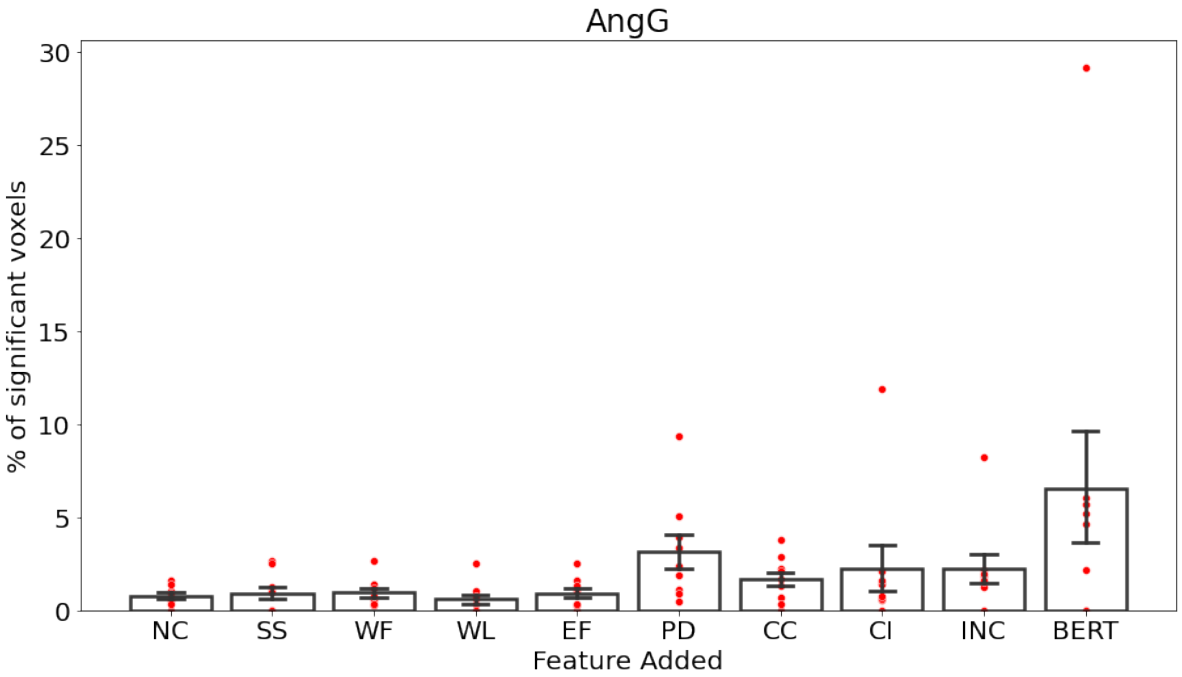


Figure 8: Results for the Angular Gyrus

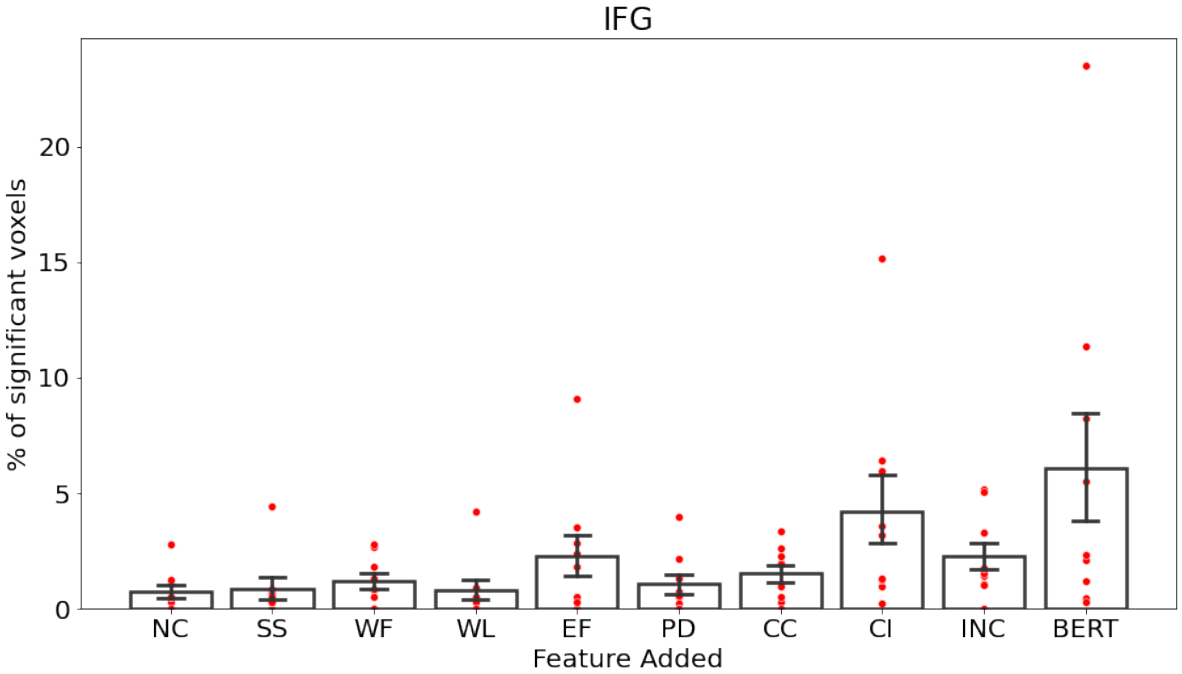


Figure 9: Results for the Inferior Frontal Gyrus

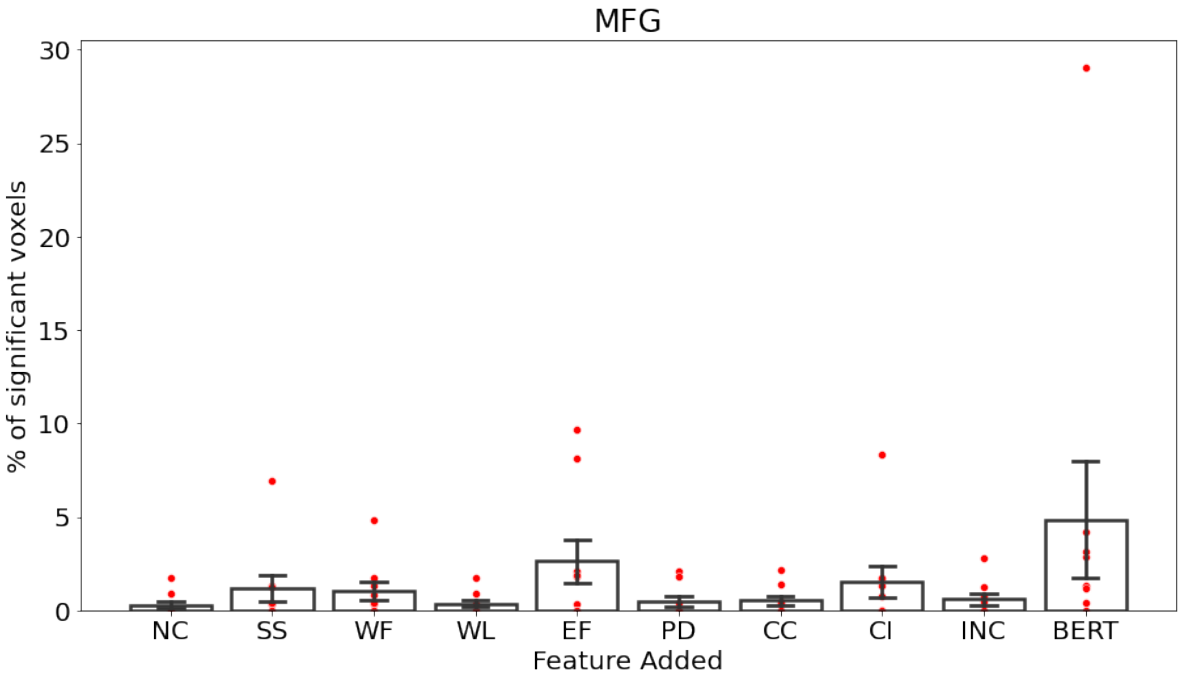


Figure 10: Results for the Middle Frontal Gyrus

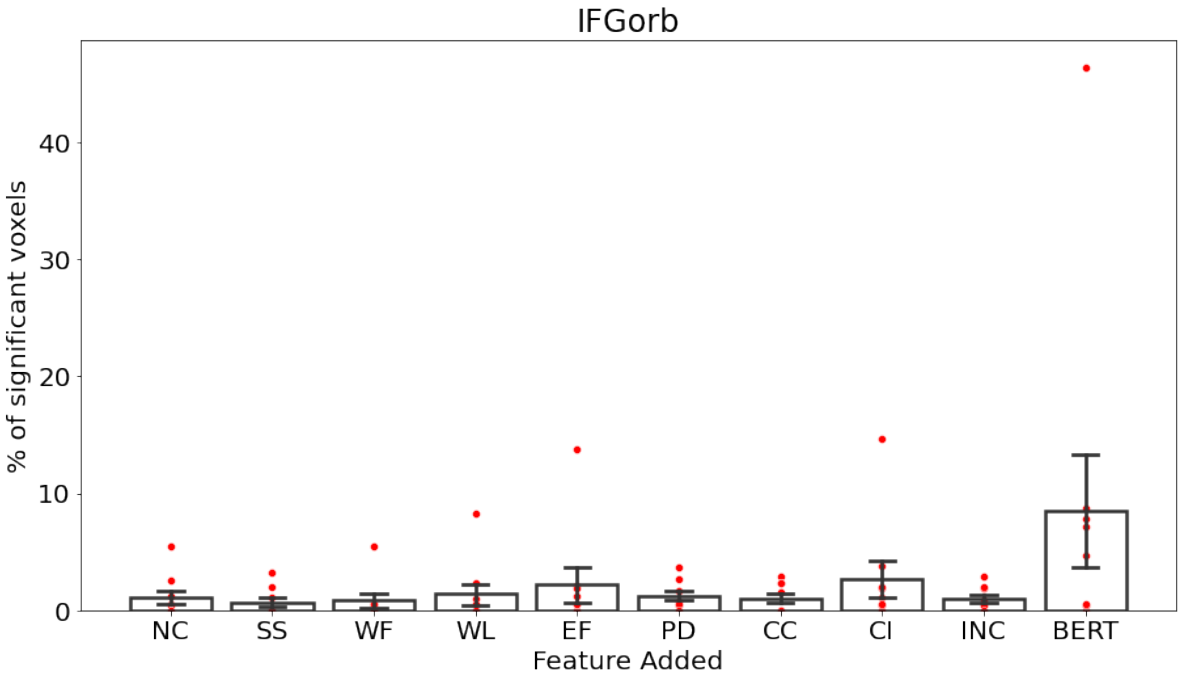


Figure 11: Results for the Inferior Frontal Gyrus Orbital