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

Aniketh Janardhan Reddy
Machine Learning Department
Carnegie Mellon University
ajreddy@cs.cmu.edu

Leila Wehbe
Machine Learning Department
Neuroscience Institute
Carnegie Mellon University
lwehbe@cmu.edu

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 processing, 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 areas. In this paper, we move beyond effort-based metrics and propose explicit features capturing the syntactic structure that is incrementally built while a sentence is read one word at a time. 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. Our results suggest that the brain represents 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 that process semantics. Our results call for a shift in the approach used for studying syntactic processing.

1 Introduction

How the brain processes syntax has long been a question of interest for neuroscientists. To date, there is no consensus on which brain areas are involved in syntax. 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 [1]. Similarly, some findings show that syntax is constrained to specific brain regions [2, 3], while other findings show syntax is distributed through the language system [4, 5].

The biological basis of syntax was first explored through studies of the impact of brain lesions on language comprehension or production [6] 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 neuroimaging experiments usually isolate syntactic processing by contrasting the activity between a more difficult syntactic condition and an easier one and identifying brain regions that increase in activity with syntactic effort [3]. 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 how syntax is processed in naturalistic settings such as when subjects read or listen to a story [7–9]. Because such natural 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 areas with activity that correlates with those metrics are suggested to be involved in syntactic processing.

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. [7] 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. [8]. 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 [8] 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 complex hierarchical structure underlying phrases and sentences.

In this paper, we propose to characterize the syntactic structure inherent in sentences by computing an evolving representation of the syntactic structure processed at each word. We propose a method for embedding the syntactic structure of sentences and phrases in a vector space. Our method relies on constituency trees and adapts a subgraph embedding algorithm by Adhikari et al. [10]. We show that our syntactic structure embeddings – along with other simpler syntactic structure embeddings built using conventional syntactic features such as part-of-speech (POS) tags and dependency (DEP) tags – are better than Node Count at predicting the fMRI data of subjects reading text. This indicates that representations of syntax, and not just syntactic effort, can be observed in the fMRI signal. Additionally, we 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 the words using a contextual word embedding space [11]. We find that regions that are predicted by syntactic features are also predicted by semantic features and thus are not selective for syntax.

1.1 Scientific questions

We ask three main questions:

- How can we construct syntactic structure embeddings that capture the syntactic structure inherent in phrases and sentences?
- Are these syntactic structure embeddings better at predicting brain activity compared to effort-based metrics such as Node Count?
- Which brain regions are involved in syntactic processing and are they different from regions involved in semantic processing?

1.2 Contributions

We make four main contributions:

- We propose a novel subgraph embeddings-based method to model the syntactic structure inherent in phrases and sentences.
- We show that effort-based metrics (specifically Node Count) can be complemented by syntactic structure embeddings.
- Using our syntactic structure embeddings, we find some evidence indicating that our brain processes and represents complex syntactic information such as phrase and clause structure.
- We find that syntactic processing appears to be distributed in the language network in areas that are not selective for syntax.

2 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.

2.1 Node Count as an effort-based metric

As Node Count is a unidimensional effort-based metric popular in neuroscience. We use it as a representative of effort-based metrics. Node Count relies on constituency trees, one of two main types of structures that capture a sentence's syntax, the other being dependency trees. 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 focus here on encoding constituency trees since we want to analyze if the brain builds hierarchical representations of phrase structure. To compute Node Count, we obtain the constituency tree of each sentence using the self-attentive encoder-based constituency parser by Kitaev and Klein [12]. We compute Node Count for each word as the number of subtrees that are completed by incorporating this word into its sentence.

2.2 Constituency tree-based Graph Embeddings (ConTreGE)

Constituency trees are a rich source of syntactic information. They are computed using the entire sentence; however, subjects read sentences one at a time. To account for this, we build embeddings using only the completed subtrees for a given word (see Figure 1(B)). The largest subtree which is completed upon incorporating a word into a sentence is representative of the implicit syntactic information given by the word. Given that Node Count reduces all of the information present in these completed 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 constituency 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. Hence, there is a need for better ways to encode the syntactic information given by them.

Further, we hypothesize that the brain not only processes structure seen thus far but also predicts structure it has not seen from structure it already knows. To test this hypothesis, we also construct embeddings 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 higher level sentence structure before the full sentence is read (see Figure 1). 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.

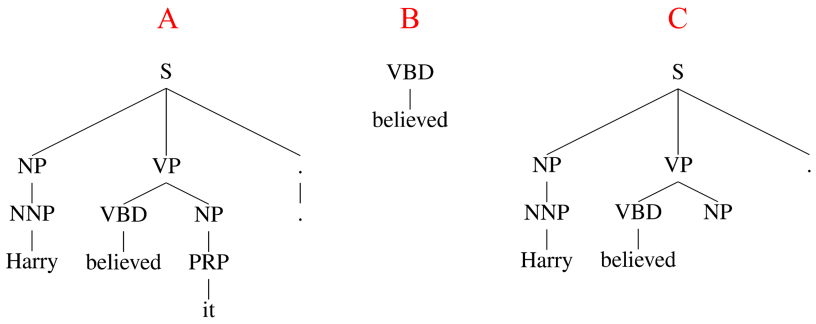


Figure 1: Example of complete and incomplete subtrees: Part A shows a sentence's constituency tree. The largest completed subtree for "believed" is shown in part B and the incomplete subtree generated till "believed" is shown in part C. Incomplete subtrees are generally much deeper than complete ones.

To effectively capture the syntactic structure inherent in constituency trees we use the subgraph embeddings proposed by Adhikari et al. [10] 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 [13] of this paragraph that is representative of the subgraph's structure.

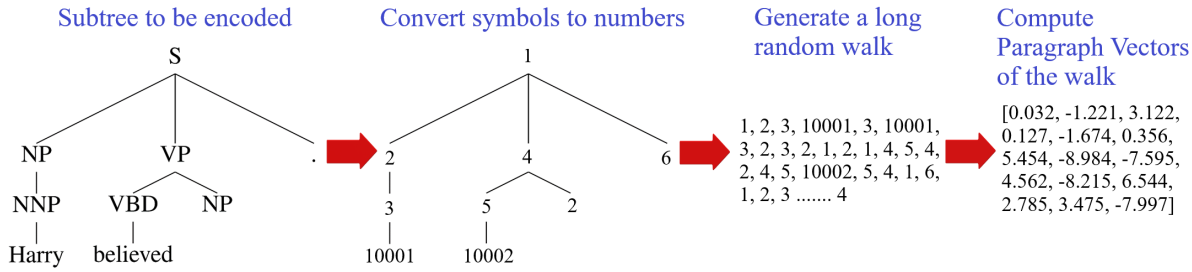


Figure 2: Steps for generating ConTreGE vectors.

We reuse the constituency trees generated to compute Node Count to build ConTreGE Comp (from complete trees) and ConTreGE (from incomplete trees). Figure 2 illustrates the generation process. To generate ConTreGE Comp, for every presented word, we extract the largest subtree that is completed by incorporating the word into it. After performing this extraction, every unique non-terminal in these extracted 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.

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 then 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 and Mikolov[13] for generating Paragraph Vectors called Sub2Vec-DBON by Adhikari et al. [10]. The length of the sliding window is set to 5 and model is trained for 20 epochs. The same process is applied to the incomplete subtrees to get the ConTreGE vectors. Since ConTreGE and ConTreGE Comp encode information about the neighbourhoods of all the nodes in the constituency trees, they are capable of capturing the hierarchical structure of these trees and the parsing decisions which led to their generation. Thus, the regions of the brain that are well predicted by these vectors are likely to be involved in building and encoding hierarchical sentence structure.

2.3 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. While punctuation is seldom considered a syntactic feature, sentence boundaries 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). Failing to account for sentence boundaries and working memory might be a source of confounding that has been ignored in the literature.

2.4 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. 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 [14]. 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.

2.5 Semantic features

We adapt the vectors obtained from layer 16 of a pretrained [15] cased BERT-large model [11] to identify regions of the brain that process semantics. We use layer 16 of this model because of previous work showing that middle layers of sentence encoders are optimal for predicting brain activity [16, 17], and that layer is also particularly good at encoding dependency tree-based syntactic information [18]. We obtain the contextual embeddings by running the pretrained model on each presented sentence. Since a presented word can be broken up into multiple subtokens, we compute its embedding as the average of the subtoken's embeddings. Finally, we perform principal component analysis (PCA) to reduce their dimensionality to 15 so that their size is comparable to those of our other features. The dimensionality reduction did not seem to affect performance.

2.6 fMRI data

We use the fMRI data of 9 subjects reading chapter 9 of *Harry Potter and the Sorcerer's Stone* [19], collected and made available online by Wehbe et al. [20]. Words are presented one at a time at a rate of 0.5s each. All the brain plots shown in this paper are averages over the 9 subjects in the Montreal Neurological Institute (MNI) space. Preprocessing details are in Appendix B.

2.7 Predicting brain activity using various features

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 Ridge regularization parameter is chosen independently. We use Ridge regression because of its computational efficiency and because of the results of Wehbe et al. [21] showing that with this type of fMRI data, as long as proper regularization is used and 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 [20, 22–24].

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 interval 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 [20, 23, 24]. 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 (the evaluation and training folds are distinct for each model i.e. a 4-fold cross validation setting). 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 each validation fold are all 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 both 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 (coefficient of determination) is computed for every voxel using the predictions and the real signals.

We run a block-permutation test to test if R^2 scores are significantly higher than chance. Block-permutation refers to permuting a block of contiguous fMRI TRs, instead of individual TRs, to better estimate chance performance by accounting for the slowness of the underlying hemodynamic response. We choose a common value of 10TRs [25]. The predictions are block-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

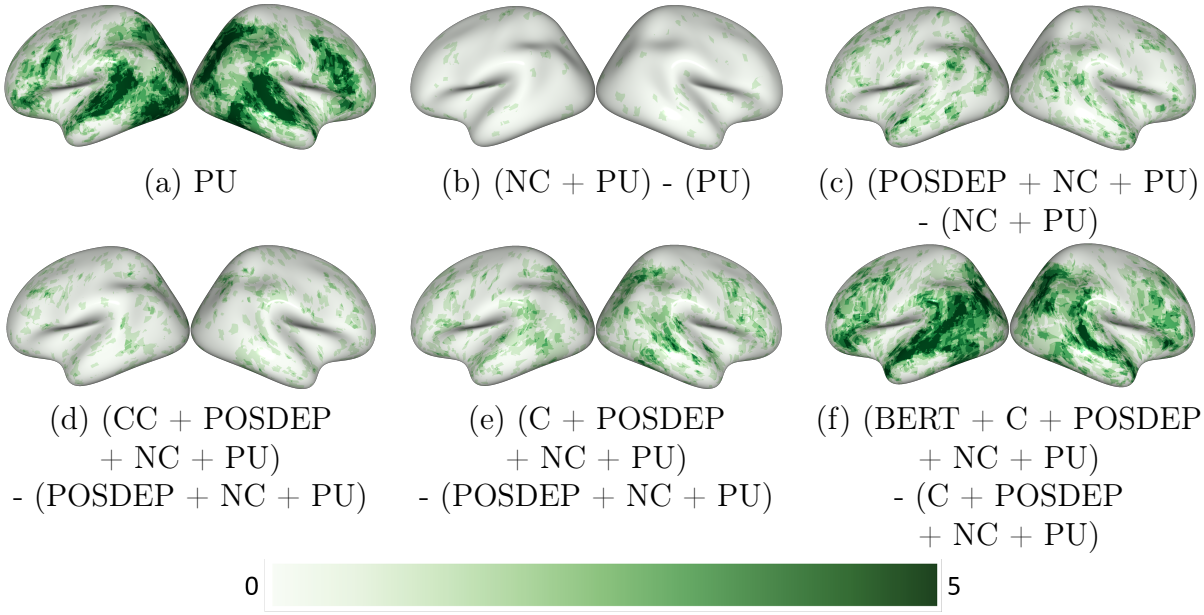
is estimated. We run a block-bootstrap test to test if a model has a higher R^2 score than another. We first bootstrap 10TR blocks of predictions for each model and for the real data (the blocks are matched for all three, for each bootstrap sample). We compute the difference between the R^2 scores of each model for that sample. We repeat this 5000 times and use this empirical distribution to compute a confidence interval, from which we can compute a p-value for the difference between the models being larger than 0. Finally, the Benjamni-Hochberg False Discovery Rate correction [26] is used for all tests (appropriate because fMRI data is considered to have positive dependence [27]). To compute Region of Interest (ROI) statistics, left-hemisphere ROI masks for the language system obtained from a "sentence vs. non-word" fMRI contrast [28] are obtained from [29] and mirrored to obtain the right-hemisphere ROIs.

3 Results

Figures 3 and 4 summarize our results and the raw prediction results are in Appendix A. Many of our features contain overlapping information. POS tags include punctuation features, BERT vectors have been shown to encode syntactic information [18] and ConTreGE vectors were built using constituency trees and encode POS tags to a certain extent. To detect brain regions that are sensitive to the distinct information given by each feature space, we build hierarchical feature groups in increasing order of syntactic information and test for significant differences in performance between two consecutive groups. We start with the simplest feature: punctuation, then add more complex features in order: node count, POS and DEP tags, one of the ConTreGE vectors and the semantic vectors derived from BERT (which can be thought of as a super-set of semantic and syntax). At each step, we test if the introduction of the new feature space lead to a significantly larger than chance improvement in R^2.

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

Figure 3(b) indicates that the syntactic information provided by Node Count is not very predictive of brain activity when we control for punctuation. Figures 3(c), (d) and (e) show that the features which explicitly encode syntactic structure are much more predictive of brain activity than Node Count with the distinct information provided by ConTreGE being the most predictive. These results are made even clearer by Figure 4. We see that there are very few voxels for which there is a significant increase in the R^2 scores obtained after including Node Count in a feature group. On the other hand, we see increases in predictive performance when we add POS and DEP tags and ConTreGE to the feature groups across almost all of the ROIs. We also notice that ConTreGE Comp is not as predictive as ConTreGE, hinting that future syntactic information helps in predicting current brain activity.



(a) PU

(b) (NC + PU) - (PU)

(c) (POSDEP + NC + PU) - (NC + PU)

(d) (CC + POSDEP + NC + PU) - (POSDEP + NC + PU)

(e) (C + POSDEP + NC + PU) - (POSDEP + NC + PU)

(f) (BERT + C + POSDEP + NC + PU) - (C + POSDEP + NC + PU)

Figure 3: The first figure 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, POSDEP = POS and DEP Tags, C = ConTreGE, CC = ConTreGE Comp, BERT = BERT embeddings and the '+' symbol indicates that these features were concatenated in order to make the predictions. The '-' symbol indicates that we tested the difference in R^2 scores between the two feature groups (bracketed) that it is between. The distinct information given by syntactic structure-based features is predictive of brain activity but that given by Node Count is not. The semantic vectors are also very predictive and many well-predicted regions overlap with those that are predicted by syntax.

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

In this section we analyze the information contained in ConTreGe to be able to interpret its brain prediction performance. We estimate how much of the constituency tree's information is captured by each feature of a word by trying to predict the level N ancestor of the word in the tree using that feature. We vary N from 2 to 9 and train a logistic regression model for each level. POS tags are the level 1 ancestors of words, we thus start this analysis with N=2. Because there are a lot of phrase and clause labels, we group them into 7 larger buckets - noun phrases (NP, WHNP), verb phrases (VP), adverb phrases (ADVP, WHAVP), adjective phrases (ADJP, WHADJP), prepositional phrases (PP, WHPP), clauses (S, SBAR, SBARQ, SINV, SQ) and other miscellaneous labels. Also, if a word's depth in its tree is less than n, the root of the tree is considered its level n ancestor.

Table 1 shows the results of this analysis. Given the skewed label distribution, the optimal strategy for a predictor that takes random noise as input is to always output the most popular ancestor at that level. We include the frequency of the most popular label in Table 1 as chance accuracy. Node Count is not predictive of the labels at any level, with the trained

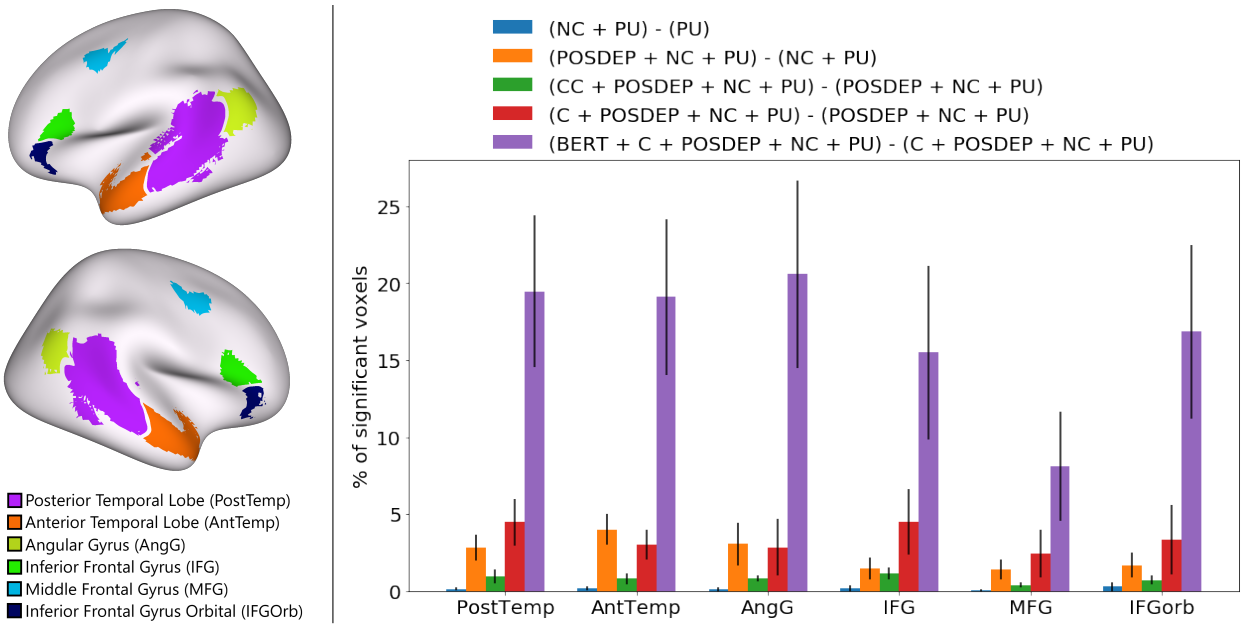


Figure 4: Region of Interest (ROI) level analysis of the prediction performance of our feature groups. The figure on the left shows the ROIs that are typically associated with language processing. The figure on the right shows the percentage of the number of significantly predicted voxels in each ROI. Each bar represents the average percentage across subjects and the error bars show the standard error across subjects. Again, we use the same abbreviations for the feature names as in Figure 3. We see the same trends as in Figure 3 across all ROIs.

model almost always outputting the most popular label for all words. Hence, we can conclude that Node Count is indeed a bad representation of the information in constituency trees. We also notice that POS and DEP tags are very predictive of labels at all levels and even produce the highest accuracies for many levels (mostly the lower ones).

ConTreGE is the most predictive of higher level ancestors but perform relatively poorly at predicting lower level ones. ConTreGE Comp is better than ConTreGE at predicting lower level ancestors but are not as good at predicting higher level ones. A possible explanation is that the graph embeddings of a tree tend to capture more of the information near the tree's root. This is because a random walk through the tree is likely to contain more occurrences of nodes near the root, leading to embeddings containing more information about them. Thus, the relatively shallow complete trees used to create ConTreGE Comp are likely to produce vectors that encode lower level ancestors and the relatively deeper trees used to create ConTreGE are likely to to produce vectors that encode higher level ancestors. Given that ConTreGE is predictive of brain activity and that they contain information about the higher level ancestors of a word, this analysis suggests that the brain represents complex hierarchical syntactic information such as phrase and clause structure.

Feature	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9
Most Popular Label %	51.0046	38.7558	54.4243	64.0456	73.4351	78.2457	82.3802	85.8192
Node Count	51.0046	42.6584	55.4675	64.0263	73.4158	78.1298	82.3802	85.8192
POS and DEP tags	**92.1368**	**71.1360**	**67.3686**	**69.9189**	**77.4923**	82.0711	86.1862	89.6638
ConTreGE Comp	66.5495	51.1399	59.6832	67.9212	77.2720	82.0827	86.2674	89.6832
ConTreGE	52.8207	45.8192	57.8941	67.4420	77.0788	**82.4691**	**86.5881**	**90.3748**
BERT Embeddings	80.7573	63.4853	66.1515	69.2233	76.8354	81.7233	85.9737	89.6252

Table 1: 4-fold cross validation accuracies in predicting the level N ancestors of a given word. The first row of the table shows the percentage of the most popular label at a given level. POS and DEP tags best predict lower level ancestors while ConTreGE vectors best predict higher level ones.

3.3 Syntax and semantics are processed in a distributed way in overlapping areas across the language system

Our results indicate that syntactic and semantic information are processed in a distributed fashion across the language network. While many regions are better predicted by semantic embeddings, there is a big overlap with those that are also predicted by syntactic features.

4 Discussion and Related Work

4.1 Syntactic representations

Apart from Brennan et al. [7] and Hale et al. [8], many others including [9, 30–33] use effort-based metrics to study the syntactic processing which occurs during natural reading or listening tasks. However, some studies have used features that encode syntactic structure. For example, Wehbe et al. [20] use POS and DEP tags and find that they are the most predictive features out of a set of other spaces describing semantic and discourse level features.

Moving away from popular approaches that are dependent on effort-based metrics, we extended the work of Wehbe et al. [20] by developing a novel graph embeddings-based approach to explicitly capture the syntactic information provided by constituency trees. Our results showed that these explicit features are substantially more predictive of brain activity than Node Count, a popular effort-based metric. Given these results, we believe that future work in this area should move away from effort-based metrics towards features that explicitly encode syntactic structure.

4.2 Syntax processing in the human brain

Traditionally, studies have associated a small number of brain regions, usually in the left hemisphere, with syntactic processing. These include various parts of the inferior frontal gyrus (IFG), ATL and Posterior Temporal Lobe (PTL) [2, 3, 34, 35]. However, some literature points to syntactic processing being distributed across the language system. Blank et al. [4] support this theory by showing that significant differences in the activities of most of the regions of the system can be seen when phrases that are harder to parse are read compared

to when easier phrases are read. Wehbe et al. [20] 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 very short phrases that have subtle syntactic variations or violations. Regions whose activity is well correlated with the presentation of such variations/violations are thought to process syntax [3]. Observations from such studies have a 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 processed and represented 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 areas.

Some studies have also doubted the importance of syntactic composition for the brain. Pylkkänen [36] proposes that there is no conclusive evidence to indicate that the brain puts a lot of weight on syntactic composition, and that though studies (some with effort-based metrics) have associated certain regions like the left ATL with syntactic processing, there have been numerous studies which have later shown that the left ATL might instead be involved in a more conceptually driven process. Gauthier and Levy [37] 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 areas of the language system. We attribute these differences in conclusions to our naturalistic stimuli and word-by-word evolving representations of syntax. Pylkkänen's conclusions are mostly based on studies that present a phrase with just two words (like "red boat"). Gauthier and Levy 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 appropriate for capturing such information.

Furthermore, our analysis of the syntactic information contained in various features highlighted that our ConTreGE 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. We observed that several regions of the brain's language system are predicted by ConTreGE, hinting that the brain does indeed encode complex syntactic information. Another potentially interesting observation is that including ConTreGE increases prediction performance in the PTL and IFG by more than when we include POS and DEP tags (Figure 4) but not for the ATL and the Angular Gyrus (AG). These observations very loosely support the theory by Matchin and Hickok [35] - 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 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 is more predictive than ConTreGE Comp with the latter being very weakly predictive. Thus, future syntactic information appears to be very useful while predicting BOLD signals, indicating that that the brain anticipates sentence structure while reading.

4.3 Semantic processing vs. syntactic processing in the human brain

Finally, our results support the theory that syntax processing is distributed throughout the language network in areas that also process semantics. This theory is supported by many other studies [4, 5, 20]. On the other hand, Friederici et al. [34] among others argue that they are instead processed in specific and distinct regions by localizing the effects of semantic and syntactic violations. Again, these differences might be due to the specialized stimuli and high statistical thresholds that only reject the null hypotheses in the regions with the strongest effect size, thereby precisely identifying small areas. A less conservative threshold might have revealed a more distributed pattern (without leading to type I errors).

References

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

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

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

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

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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*, pages 170–182. Springer, 2018.

[11] 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.

[12] 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.

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

[14] *spaCy, en_core_web_sm model.* URL https://github.com/explosion/spacy-models/releases//tag/en_core_web_sm-2.2.5.

[15] *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.

[16] 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*, pages 14928–14938, 2019.

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

[18] 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)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://www.aclweb.org/anthology/N19-1419.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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)*, pages 289–300, 1995.

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

[28] 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.

[29] *Group-level functional parcels.* URL https://evlab.mit.edu/funcloc/download-parcels.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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)*, pages 529–539, 2019.

[38] 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.

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

[40] 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.

Appendix

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.

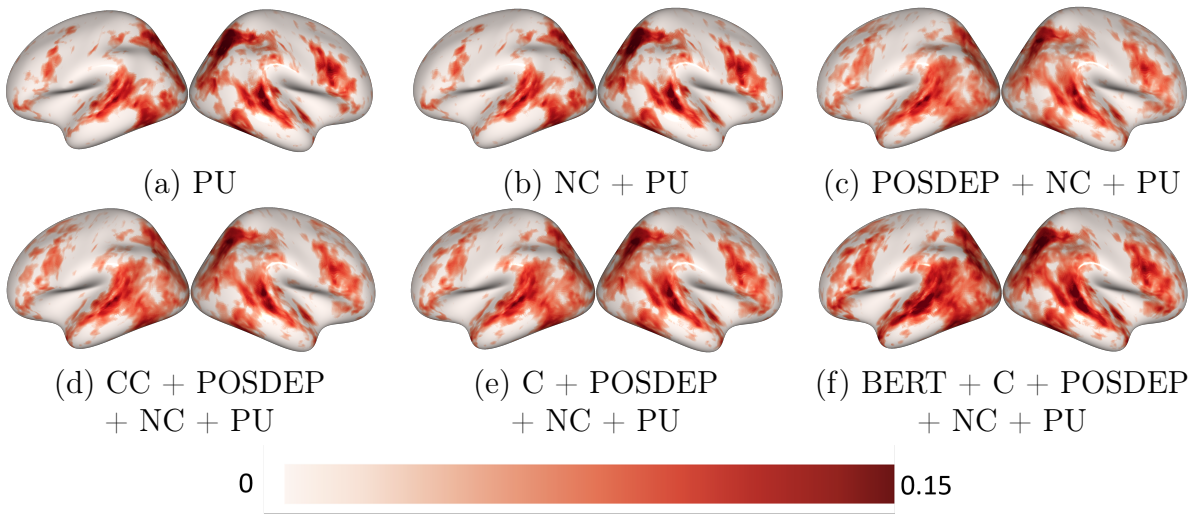


Figure 5: Cross-subject prediction performance of all syntactic feature groups. The figures show cross-subject average R^{2+} scores. Here, PU = Punctuation, NC = Node Count, POSDEP = POS and DEP Tags, C = ConTreGE, CC = ConTreGE Comp, BERT = BERT layer 16 vectors reduced to 15 dimensions using PCA and the '+' symbol 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[20]. 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 [38], 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 [39] and a grey matter mask is obtained. Pycortex [40] is used to handle and plot the data. All subject results are converted to MNI space using pycortex.