Neural Discrimination of Syntactic Structures: Connecting Theory to Brainwaves

by

Isobel Moure

A Thesis

Submitted to the Columbia University Cognitive Science Department

In Partial Fulfillment of the Requirements

For the Degree of Bachelor of Arts

May 2023

Advisors: Drs. Christos Papadimitriou, Tony Ro, Tatiana Aloi Emmanouil

Abstract

Do our theories of syntax match what we can measure in the brain? This project sought evidence that the mind discriminates between different syntactic structures, replicating and extending the findings of the Ding et al. 2016 study. Participants, monitored with an EEG cap, were presented with synthetic speech sentences of different classes that would hierarchically merge at different rates, according to English phrase structure. A story class was added to test if sentences that were part of a larger narrative were processed differently. A Support Vector Machine was trained on the data and used to attempt to decode the sentence class from the EEG data. It was found that on the majority of conditions, the classifier obtained an accuracy higher than chance. These data suggest hierarchical processing of constituent structure and provide further evidence for the Assemblies hypothesis as well as support the previous findings from Ding et al.

Neural Discrimination of Syntactic Structures: Connecting Theory to Brainwaves

How can we connect theories of syntax to the actual functioning of the brain? If syntacticians seek to describe the way language is structured (partly in order to explain how the brain processes language), then it is of interest if the brain recognizes the proposed structures. Additionally, understanding what models the brain utilizes can help us understand how neurons implement language processing. Language comprehension has long been a fascinating and frustrating challenge to all those who study the brain. There is still no agreed upon model of language processing that is defined down to the neural level, however. Research often focuses either on the neuron-to-neuron level or the theoretical level of language structure (Papadimitriou et al., 2020). This project seeks to further the task of connecting measurable neural outputs to our theories of language structure and processing.

Specifically, this work seeks to determine if the brain hierarchically merges constituents to parse natural language. Constituent structure dominates much syntax literature as the basis for how language is organized. Constituents are simply groups of words bound together to form a semantic unit. They are typically analyzed using phrase structure rules that define how constituents hierarchically merge into an entire phrase and sentence (Chomsky, 2002). Figure 1 shows two example sentences diagrammed in simple syntax trees. The sentence “Girls push large carts,” is broken down into the noun phrase (NP), “girls” and the verb phrase (VP), “push large carts”. The VP is further parsed into a verb, “push”, and another NP, “large carts.” The same treatment is given to the second sentence, but the subject NP is comprised of two words. The words themselves, as well as the phrases, make up constituents. The tree diagram makes clear the hierarchical structure of these

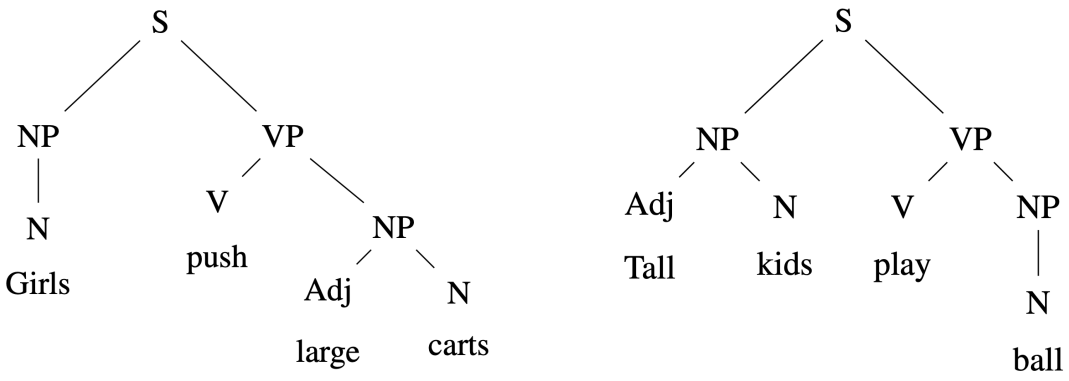

Figure 1. Simple syntax trees for the sentences "Girls push large carts" and "Tall kids play ball"

sentences. The constituents merge to form increasingly larger phrases with progressivley more complicated semantic meaning.

Ding et al. (2016) found evidence of brain activity tracking these constituent structures when parsing speech. In the study, researchers recorded brain activity with magnetoencephalography (MEG) while participants listened to synthetic speech. Stimuli were sentences comprised of four monosyllabic words. Words were presented at a rate of 4 Hz, which regularly combined into phrases at a rate of 2 Hz, which in turn combined into sentences at a rate of 1 Hz. Recording data were then transformed into the frequency domain using a Fast Fourier Transform (FFT) and plotted. As hypothesized, they found increased power at 1, 2, and 4 Hz, exactly the rates that the constituents combined; the neural cycles matched up with the constituent cycles. These results strongly indicate that the brain is somehow tracking constituent structure to parse language.

The question still remains of exactly how neurons track the structure of language. Papadimitriou et al. (2020) sought to define exactly how neurons were implementing algorithms by proposing the Assemblies Hypothesis. This hypothesis seeks to fill in the gap between our understanding of neuron-to-neuron connections and whole brain models of

cognition. Fully describing the Assemblies Hypothesis is beyond the scope of this paper, but the essential idea behind the hypothesis is that neurons work together in groups called assemblies. Assemblies are based on the Hebbian model of neural encoding and present Hebbian plasticity, i.e., neurons that fire together once are more likely to fire together again (Morris, 1999). Repeated firing together eventually creates a stable set of neurons which form an assembly. Assemblies have a small, discrete set of operations which the authors claim can carry out arbitrarily complex processes under a small set of assumptions. Assemblies and their operations constitute a model of the brain called the Assembly Calculus (AC).

In 2021, the same group proposed a "plausible parser" which took the initial set of assembly operations and showed how they could be used to parse language. A computational model demonstrated that the parser is able to sufficiently parse relatively complicated sentences. The parser works by creating a hierarchical dependency-based structure, incrementally merging constituents to create increasingly complex meaning. Since the parser is based on a hierarchical parsing of language, the AC offers a plausible explanation for the previously discussed Ding et al. results. In turn, more evidence for hierarchical processing would support this implementation of parsing by the AC.

This project seeks to provide more evidence for cortical tracking of hierarchical structures in natural language by extending the Ding et al. study and considering the implications for AC. Our stimuli are consistent with the Ding et al. study, with the addition of two classes of sentences: imperative and story. In addition to testing intra-sentential parsing, it was hypothesized that sentences that were part of a discourse were merged in similar cycles. Participants were presented the stimuli in the same, regulated frequency of constituent merging as in the Ding et al. study. Instead of plotting the frequency domain data, results were

analyzed using a machine learning algorithm. As opposed to examining the data to see if the distinct hierarchical cycles were present, we measured how effective a model was at discriminating between different syntactically formed sentences. If the brain is tracking constituent structures hierarchically, then different structures would produce different neural cycles. It was hypothesized that a machine learning model should be able to effectively classify what syntactic structure was attended to when trained on neural data in the frequency domain. These results would replicate and extend the Ding et al. findings in a novel way as well as give further support to the Assemblies Hypothesis.

Method

Participants

Two participants' data were tested. The initial participant was 21 years old and left-handed and the second was 25 years old and right-handed. Both participants were native English listeners, meaning they spoke English fluently before the age of 6, and did not use any devices to aid in hearing.

Stimuli

Test sentences were split into five categories: balanced, unbalanced, imperative, story, and "blahs". All sentences contained four monosyllabic words that are common in the English language. Balanced sentences had a noun phrase (NP) and verb phrase (VP) of equal length--two words each. The sentences took the form [ADJ] [N] [V] [N], for example, "tall kids play ball." Unbalanced sentences comprised of a single word NP and a three-word VP, taking the form [N] [V] [ADJ] [N]. For example, "mice smell old cheese". The imperative class

comprised of four-word imperative sentences of the form [V] [PRO] [DET] [N]. For example, “send them the gifts.” The story class contained four twelve-sentence stories. A story was defined as a series of sentences that were connected narratively; each story had a small narrative arc with a resolution at the end.

All classes avoided use of the English copula, the verb “to be”. Each class (other than the “blah” class) had 12 unique sentences although a few nouns were repeated throughout different sentences. In addition, filler “sentences” were constructed with four blahs to take the form “blah blah blah blah.” The stimuli were recorded using the Amazon speech synthesizer with the voice Polly in order to avoid prosodic cues and to keep the phrases of equal time length and at a regular pace. The synthesized syllables were padded with silence or truncated to maintain an equal length of 320 milliseconds. Syllables/words were at the frequency of 3.125 Hz, phrases were at 1.5625 Hz, and sentences were at 0.78125 Hz. This frequency is similar to the speed of natural speech by native English speakers, so participants were not challenged by understanding the stimuli. Each sentence was 1281 milliseconds long.

Procedure

Participants were monitored using a 32 channel Brain Vision EEG cap at a sampling rate of 1000 Hz. In a sound booth, participants were instructed to watch a fixation cross on a computer to prevent visual confounding variables. The screen remained the same throughout the experiment and no stimuli were shown on the screen as only audio speech perception was being tested. In order to ensure participants were actively listening to the stimuli, they were required to count the number of times they heard the phrase “blah blah blah blah” and report

to the administrator the count. This was done to gauge if the participant was paying sufficient attention.

Each 12-sentence set (other than the story class) was played twelve times for a total of 144 trials per class. Every time the set was played, it was arranged in a different order. The story class was played at the end of the session. Each story was presented three times for a total of 12 stories played. Between every story a "blah" filler sentence was played to indicate the barrier between narratives. In every block other than the story condition, one to three blah sentences were randomly inserted. There were no gaps between sentences for all of participant A's trials. Participant B did one trial with no gaps (identical to participant A) as well as a second trial with a gap of 1274 milliseconds between every sentence.

Analysis

The data were analyzed using a Support Vector Machine (SVM). An SVM is a supervised machine learning algorithm which takes labeled data and learns to separate it into classes and predict what class new data points are in. The fundamental idea of an SVM is to plot the data into an n-dimensional space (where n is the number of features) and find the hyperplane which best separates the data. This is most easily imagined in 2-dimensional space, but this analysis can be done in datasets with arbitrarily large feature sizes. The algorithm evaluates the ideal separating plane by measuring the "support vectors." Support vectors are the data points closest to the boundary between classes. Removal of these boundaries would affect where the hyperplane is drawn. The algorithm chooses a boundary that maximizes the margin of space between all of the support vectors and the boundary itself.

By maximizing the margins, the algorithm approximates the actual boundary of change in the data.

If the data are not linearly separable in the dimensionality of their features, they are mapped onto a higher dimension using a kernel function. This is easiest to imagine in 2-dimensional space where a class is encapsulated within another in the space. The hope is when the data are mapped onto 3D space, the boundary can be a linear plane that goes between the "levels" of data. Sometimes, however, the data need to be transformed into more than one higher dimension to be separable.

Scikit-learn's implementation of an SVM was used to analyze these data after some minor preprocessing. The data were put through a bandpass filter with a low of 0.1 Hz and high of 30 Hz to reduce signal noise. After epoching, the data were transformed into the frequency domain using MNE's implementation of the Fast Fourier Transform. The data were fit using crossfold validation with five folds. Scikit-learn's implementation of grid search was used to optimize for ideal hyper-parameters and it was found that C of 10 and gamma of 1 maximized performance. The radial basis function was used as the kernel for the model.

The SVM was trained on different combinations of channels and were used to draw out spatial components of the data. All electrodes, just the left temporal electrodes (T7 and TP9), the entire left hemisphere, or the entire right hemisphere were inputted. Code is available upon request.

Results

Consistent with our hypothesis, the classifier performed above chance for the majority conditions and classes. Since there were five classes, chance classification was 20%.

Anything above 20% is therefore considered important. For participant A, classification of all classes was well above chance with input from all electrodes with a total accuracy of 41%. Participant B's data showed lower results across the board, with an accuracy of 34% for the identical, ungapped trial. The main difference in accuracy appears to be in the blah class; participant A's blah class has an F1 score of 0.62 while participant B's ungapped blah class has an F1 score of 0.09.

Participant A (ungapped)				
	All electrodes	Only TP9 and T7	Right hemisphere	Left hemisphere
Overall accuracy	0.41	0.24	0.35	0.37
Balanced class F1	0.36	0.0	0.28	0.31
Unbalanced class F1	0.33	0.0	0.28	0.26
Imperative class F1	0.41	0.39	0.39	0.38
Blah class F1	0.62	0.0	0.64	0.59
Story class F1	0.43	0.0	0.33	0.44

Participant B (ungapped)				
	All electrodes	Only TP9 and T7	Right hemisphere	Left hemisphere
Overall accuracy	0.34	0.24	0.29	0.29
Balanced class F1	0.33	0.0	0.24	0.28
Unbalanced class F1	0.34	0.0	0.24	0.25
Imperative class F1	0.23	0.39	0.27	0.29
Blah class F1	0.09	0.0	0.03	0.0
Story class F1	0.56	0.0	0.44	0.38

Figure 1. Data table containing a summary of results for both participants and all conditions. Here, F1 score is defined as $\frac{2*p*r}{p+r}$ *where p is precision and r is recall. This is used because accuracy for a single class is undefined for a non-binary classifier. F1 score is standard machine learning metric that captures the essence of accuracy by taking the harmonic mean between precision and recall.*

For all trials, the two left temporal lobe electrodes proved insufficient to produce meaningful classification. Interestingly, the imperative class achieved a relatively high F1 score for this condition (0.39 for both A and B) while all other classes had a zero F1 score.

Inspecting the precision and recall scores shows that recall for the imperative class is 1.0, suggesting that the classifier found the highest accuracy by classifying all sentences as imperative. The right and left hemisphere conditions gave similar results for all trials, with the left hemisphere preforming slightly better. This is to be expected as the vast majority of speakers process speech in the left hemisphere.

Participant B (gapped)				
	All electrodes	Only TP9 and T7	Right hemisphere	Left hemisphere
Overall accuracy	0.26	0.24	0.24	0.27
Balanced class F1	0.29	0.04	0.30	0.30
Unbalanced class F1	0.30	0.02	0.20	0.25
Imperative class F1	0.24	0.38	0.21	0.23
Blah class F1	0.07	0.0	0.11	0.12
Story class F1	0.28	0.02	0.30	0.36

Figure 2. Data table containing a summary of results for participant B's gapped trial. Blocks were played with a 1274 millisecond gap between every sentence.

The ungapped trials for participant B classified more accurately than that of the gapped trials. The overall accuracy for all electrodes in the gapped trial was only 26% while the ungapped accuracy was 34%. The overall patterns in different conditions remained constant between the gapped and ungapped trials. The gapped trial was very consistently a few percentage points below the ungapped trial.

Discussion

The data suggest that parsing different syntactic structures does generate different cyclical neural signatures, consistent with our hypothesis. The SVM was able to consistently classify the EEG data higher than chance, as high as double than chance. The model successfully identified patterns in the frequency domain data that were unique to each class in

order to find the boundary plane. This suggests that parsing different sentences produces different cyclical signals based on different syntactic structure, indicating that when the brain parses speech, it does so by binding together smaller units cyclically into one large unit, hierarchically parsing the individual words to combine the entire meaning.

This effect was shown most strongly in the all electrode, ungapped trials. The overall lower scores on the spatialized trials indicates that the EEG data was not sufficient to draw out spatial qualities in processing. This is unsurprising, as EEG has fairly weak sensitivity to spatial features (but makes up for this shortcoming in high temporal sensitivity). Alternatively, this result could mean that the parsing of speech is not limited to one region of the brain. While many previous results indicate that language is mainly processed in the left hemisphere of the brain, it is conceivable that the brain recruits from both hemispheres in order to completely parse a sentence.

Regardless, it is clear that the SVM was not able to identify any discernable pattern with only input from the left temporal lobe. The model found highest accuracy in all cases by classifying every trial as imperative. This may not, in fact, hold deep meaning, but rather reflect the optimization choice made by the model based on limited information which does not reflect an underlying feature. This is a common failure of machine learning models, emphasizing the need to examine all class accuracy, not just overall accuracy. Accuracy is fairly well distributed throughout classes in all other conditions, reinforcing the validity of these results. Further research should be done with recording equipment with more precise spatial abilities to examine the spatial qualities of language parsing, particularly in relation to assemblies.

The story class had an F1 score on par with the other classes. The sentences in this class did not differ substantially from the unbalanced sentences intra-sententially, but the SVM was still able to successfully categorize them. This result suggests that language that is part of a larger discourse is processed in a cyclically unique manner, possibly the same fashion in which words are hierarchically combined. More work should be done with more precise recording equipment to examine how discoursal speech is parsed.

The ungapped trial for Participant B produced an 8% higher overall, all electrode accuracy than the gapped trial. The gapped condition was introduced after concerns that the ungapped condition had low intelligibility as the sentences could sometimes blend together. This result supports the validity of testing with ungapped stimuli, however, as it is clear that the sentences remained intelligible. Additionally, since the cyclical nature of parsing is of particular interest in this study, the ungapped stimuli likely maintained a more regular cycle that the SVM was able to detect. This result should encourage future researchers using these stimuli to opt for an ungapped version.

The difference between the two participants' results underscores the need for further research in this area with more participants. It is difficult to pinpoint exactly why the accuracies differ. It is possible that Participant B was simply not attending as closely as Participant A. Alternatively, Participant B's data might have been noisier due to differences in electrode placement or any other range of factors. These preliminary results show promising evidence of the cyclical nature of speech parsing, but further studies should be done with a larger number of participants and, ideally, better spatialized data to support these findings.

Additionally, while using an SVM is an extremely effective tool to bring out patterns in the data not noticeable to humans, it is difficult to examine what exactly those patterns are.

Machine learning algorithms are notoriously a "black box," meaning it is often impossible to find meaning in the parameters. While the SVM proved effective in classifying the data, it is difficult to actually explain what features it was detecting. The Ding et al. study findings suggest that there is stronger energy in frequencies where constituents regularly merge, but it is impossible to say if the SVM is registering these peaks. The SVM is certainly identifying a difference in frequencies depending on the sentence structure, but it is not necessarily correlated to the constituents that we hypothesize to be real.

Another issue is that of non-phrasal languages or, more generally, languages that do not have a robust notion of phrase structure or even words. Phrase structure has been established as a productive theory for mostly Western languages such as English, German, or Spanish. There are many languages, however, in which phrase structure is not as useful or sensical. Many polysynthetic, agglutinative languages have freer word order and favor morphemes or other linguistic markers to indicate the syntactic role of words. Single "words" can carry the meaning of an entire sentence and words do not necessarily need to be temporally presented next to each other to be combined into a single semantic unit (Haspelmath & Sims, 2010). Cognitive science in general often makes the mistake of over-generalizing from a small, highly regular population, which tends to be WEIRD (Western, Educated, Industrialized, Rich, Democratic) (Henrich et al., 2010). Much linguistic literature ignores the pesky reality of wide language diversity in favor of analysis that works well only for languages of those in power and requires great stretches to fit onto smaller languages.

This does not necessarily mean that these results are useless, or only applicable to English-speakers. Rather, I propose that we (those who study the brain and language) should further examine how this hierarchical merging shows up in different languages. While

constituents may not be universal, the higher-level idea of merging discrete semantic concepts into increasingly complex ideas to parse a sentence seems to get at the core of what it is to process and produce language. The plausible parser, as proposed by Papadimitriou et al. is primarily based off of the "merge" assembly operation. The assemblies could conceivably be merging not constituents but morphemes or other small information units that can combine into a complete phrase or idea. More research is necessary to examine what neural signatures show when listeners parse non-constituent structure languages.

One can imagine a similar experiment using EEG and an SVM with a language like Lushootseed, spoken by the indigenous peoples of modern-day Seattle, Washington. Lushootseed is a polysynthetic language with a fairly free word order and verbless sentences as well as verb-only sentences (Hess, 1995). It would be enlightening to determine if listeners show these frequency peaks at regular merge points at the level of morpheme or sentential discoursal level. We should be ensuring that our theories are crosslinguistically effective and not overly focused on culturally dominant languages.

Overall, this study offers encouraging evidence for the hierarchical merging of constituents. These findings support the Ding et al. results and the Assemblies Hypothesis. They also further previous findings through the story condition which suggests sentences are merged together by a similar mechanism as are words. More work remains to be done to flesh out the Assemblies Hypothesis cross-linguistically and find more evidence to support these mechanisms.

References

Chomsky, N. (2002). *Syntactic structures* (2nd ed). Mouton de Gruyter.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), Article 1. https://doi.org/10.1038/nn.4186

Haspelmath, M., & Sims, A. (2010). *Understanding Morphology* (2nd ed.). Routledge.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hess, T. (1995). *Lushootseed Reader with Introductory Grammar*. Tulalip Tribes.

Mitropolsky, D., Collins, M. J., & Papadimitriou, C. H. (2021). A Biologically Plausible Parser. *Transactions of the Association for Computational Linguistics*, *9*, 1374–1388. https://doi.org/10.1162/tacl_a_00432

Morris, R. G. M. (1999). D.O. Hebb: The Organization of Behavior, Wiley: New York; 1949. *Brain Research Bulletin*, *50*(5), 437. https://doi.org/10.1016/S0361-9230(99)00182-3

Papadimitriou, C. H., Vempala, S. S., Mitropolsky, D., Collins, M., & Maass, W. (2020). Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences*, *117*(25), 14464–14472. https://doi.org/10.1073/pnas.2001893117