From He to Holmes: Analyzing the neural link between pronouns and their referent using MEG

Ricky Rojas (rickyro@stanford.edu)
Bachelors of Science Candidate, Data Science
Stanford University

Abstract

Understanding how the human brain resolves pronouns with their referents during natural speech processing is crucial for deciphering complex language comprehension mechanisms. This study explores the neural dynamics of pronominalization using magnetoencephalography (MEG) data from participants listening to stories from ”The Adventures of Sherlock Holmes.” By employing naturalistic story stimuli, we aim to investigate the neural correlates of coreference resolution, a relatively under-explored area in contrast to traditional, controlled experimental setups.

We found that pronouns and common nouns are decipherable pre-onset while proper nouns are decodable later, potentially aligning with previous research on POS surprisal. We failed to find a neural trace linking pronouns to specific characters nor generalizability of neural signals between characters and their pronouns, reinforcing previous research that referents are processed within different regions of the brain from referential nouns.

Keywords: MEG; part of speech; coreference; decoding; neural processing

Introduction

The process of decoding continuous speech is complicated by ambiguity. From overlapping phonemes to homophones, the process of parsing and interpreting language is a complicated and important area of neuroscience. One source of auditory ambiguity comes from the use of pronouns to replace nouns in speech, a process called pronominalization. The meaning of pronouns like “I”, “he” and “it” are highly dependent on the context in which they are used, and the referent can shift over the course of a sentence. For example, consider the phrase “Sarah yelled at Sally. She is angry at her.” Despite the ambiguity of this sentence, humans are very good at connecting pronouns with their referents, a process known as coreference resolution (Brodbeck & Pylkkänen, 2017).

Previous research on coreference has focused on the regions of the brain responsible for processing referential language (Brodbeck & Pylkkänen, 2017; Brodbeck, Gwilliams, & Pylkkänen, 2016) or linguistic models of coreference resolution that align best with neural decoding (Jixing Li et al, 2020). These studies combine auditory and visual stimuli in structured experimental settings to isolate the process of referential processing. However, there isn’t much research on coreference using naturalistic story stimuli. Of the research that exists, it has been shown that there is a difference in neural response based on the type of coreferential structure within longer form language processing, which isn’t easily replicated in tightly controlled experimental settings (Brilmayer & Schumacher, 2021). This means that there is a need to build upon the research performed in constrained environments.

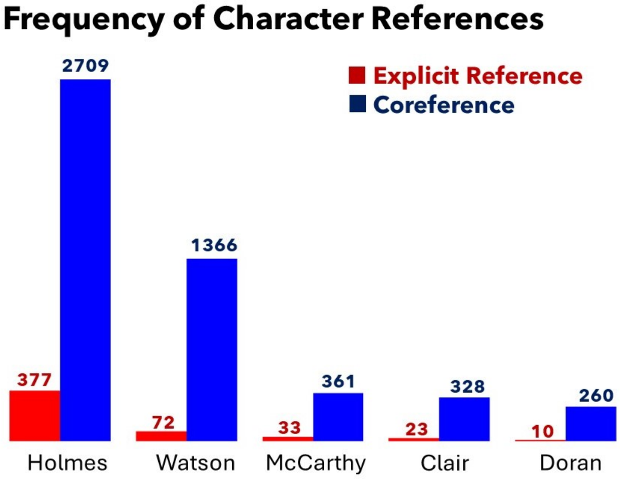

Figure 1: Difference between the frequency in explicit character references and references using referential nouns like pronouns for the 5 most common characters in 10 chosen stories of *The Adventures of Sherlock Holmes*

As such, the goal of this study is to investigate the link between pronouns and their referent during natural language processing. The process of pronominalization occurs frequently in narrative texts such as Sherlock Holmes (Figure 1). This complicates the analysis of the neural response to characters within narrative texts since the audio stimuli of the character’s name occurs very infrequently. By better understanding the relationship between the neural response and coreference resolution, we can better structure future analyses of coreference on naturalistic story stimuli. To do this we’ve devised three analyses to evaluate the decipherability between characters and their pronouns at timepoints before, during, and after word onset.

Methods

Data

We used a magnetoencephalography (MEG) dataset from Armeni et al (2022). It consists of 3 right-handed, native English speakers (1 female) listening to ten stories of ”The Adventures of Sherlock Holmes” by Arthur Conan Doyle as read by David Clarke while in a MEG-compatible head cast to minimize motion artifacts. The dataset was pre-annotated with word onsets and offsets through automatic forced alignment of the audio recordings and the text. Due to issues with the dataset, patient 3, session 8, segment 7 as well as any words that end with an accented letter are excluded from analysis.

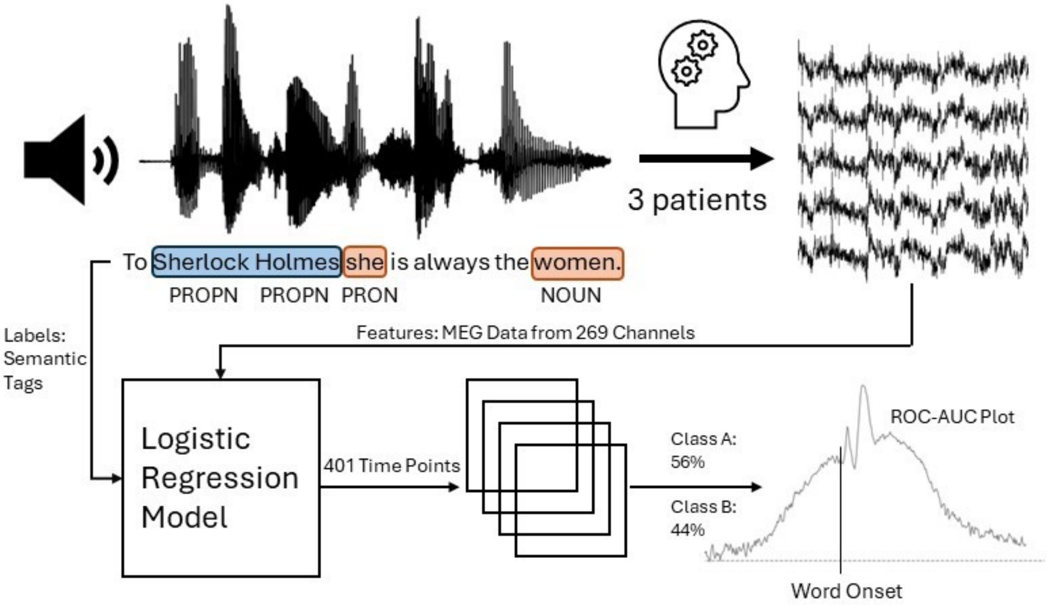

Figure 2: Process Diagram for Data Analysis. Part of speech and coreference labels are attached to the pre-annotated dataset. These labels are used to define classes for a logistic regression model which uses neural MEG data. Each logistic regression model is trained at all 401 timepoints and computes the probabilities that a given epoch belongs in either class A or B. The decodability (measured by ROC-AUC) is then graphed and evaluated for statistical significance ($p < 0.001$) using a permutation-cluster test

Preprocessing

Analysis was performed on data from all 269 MEG data channels after being down sampled to 100 Hz and bandpass filtered to 0.1 - 30 Hz using IIR forward-backward filtering. The preprocessed continuous MEG data was then epoched from -2s to 2s relative to word onset. No baseline correction was applied.

Data Annotation

We labeled the part of speech for all words in the dataset using the python package "spaCy," on the raw text (including punctuation) and aligned it with the pre-annotated data. We used the python package "fastcoref" to resolve referential nouns into clusters with their referents. This coreference strategy has a F1 score of 78.5, meaning this annotation process isn't perfect (Otmazgin, Cattan, & Goldberg, 2022). In order to simplify analysis between single token pronouns and multi-token character names, we ran all analyses on the epoch assigned to the last name of each character as Doyle overwhelmingly refers to characters by their last name when referenced explicitly. Because referents can be noun phrases and we are only concerned with character pronouns, we found the "head" of the cluster (the referent) by looking at the strings in each cluster and choosing the most common, last proper noun. Finally, we labeled all the elements in a cluster with the head. If no head/proper noun was found or a word was not in a cluster, then it was omitted from analysis.

Analysis Overview

The analyses relied on the python package scikit-learn. This includes the functions LogisticRegressionCV, StandardScaler, and StratifiedKSplit. We used LogisticRegressionCV to optimize the regularization parameter at each time point, selecting the best model fit on ten log-spaced alpha parameters from 1e-4 to 1e+4. We found the best parameter to be 1e-3, which was then used for the rest of the analyses. We trained 1-v-all logistic regression models at all 401 time points for each of the analyses using 5-fold cross validation. For character-based analyses we limited the samples to the top 18 characters to maintain a reasonable sample size ($n > 80$).

Similarly for pronoun-based analyses we limited the samples to the top ten pronouns associated with the selected characters. To evaluate the performance of the models at each time point we used the receiver operating characteristic (ROC) area under the curve (AUC) which is a measure that balances the precision-recall of the classification. Finally, to find the time points in which the neural response is statistically decodable above chance ($p < 0.001$), we used permutation_cluster_1samp_test from the python library "mne" to perform a non-parametric permutation-cluster test with 10000 permutations and a paired t-test as the statistical test.

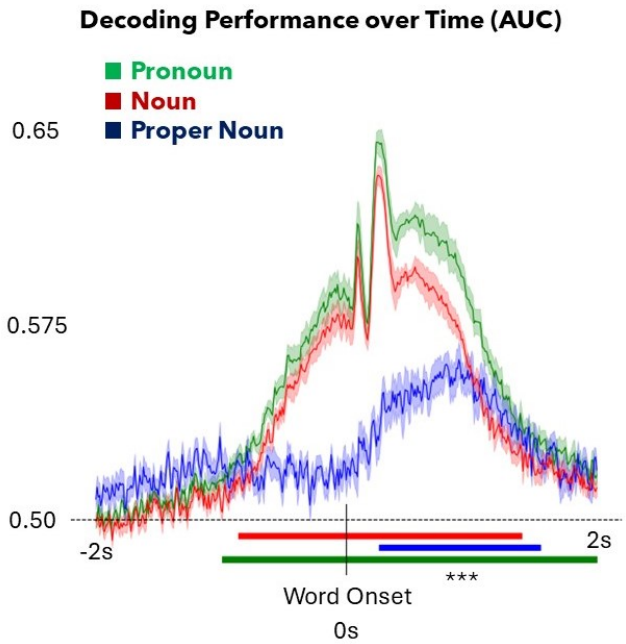

Figure 3: Timecourse of part of speech decoding. Nouns and pronouns are decodable around a second before word onset while proper nouns take 250 ms post onset. The bars at the bottom indicate times where decoding performance is statistically significant (p $<$ 0.001)

Results

Analysis 1

The first analysis we performed was to compare baseline decodability between different types of nouns (Pronouns, Common Nouns, Proper Nouns). We ran two analyses for this experiment. First was a 1-v-all logistic regression comparing the three classes (Figure 3), the second was a 1-v-1 logistic regression between pronouns and character last names. To have an adequate sample size for the permutation-cluster test, we randomly paired two sessions together for each patient for a total of 3 $\times$ 5 samples. We found statistically significant clusters (p $<$ 0.001) for all three types of nouns. Pronouns and Nouns were decodable at a rate statistically greater than chance beginning $\sim$800 ms pre-word onset and peaked $\sim$300 ms post-word onset. Proper nouns did not become decodable at a rate above chance until $\sim$250 ms post-word onset, peaking later $\sim$1 second post-word onset. The results of the 1-v-1 logistic regression between pronouns and characters matched the decoding window for proper nouns in the first analysis, with neural activity being significantly decodable at $\sim$250 ms post-word onset.

Analysis 2

The second analysis we performed was to investigate the existence of a neural trace between character references and their pronouns. To do this, we trained a logistic regression model on the coreference labels of a specific character pronoun (like “he”). We then performed a 1-v-all logistic regression on the characters with at least 30 coreference labels associated with that pronoun. Finally, we performed an item-based permutation-cluster test for each sample across patients (each character, pronoun combination produces 1 sample). As seen in Figure 4, we did not find any time points that were statistically decodable above chance for this analysis (p $<$ 0.05). These results held when looking at the samples aggregated across pronouns and characters as well as separately.

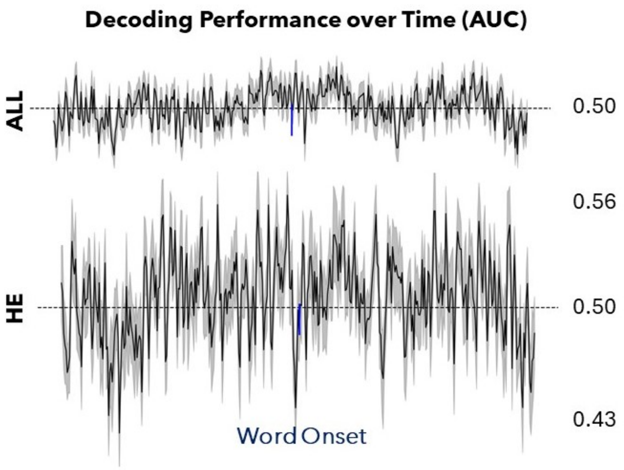

Figure 4: Timecourse of character decoding based on pronouns. The top graph shows the decodability over time of all characters and pronouns over time. The second graph just looks at the epcochs corresponding to the pronoun ”he”. In both analyses a permutation-cluster test indicates that there are no statisitcally signifcant points where the model preforms above chance. Both graphs are on the same scale.

Analysis 3

The final analysis was performed to test the generalizability between the neural signal between characters and their pronouns. To reduce the results of auditory differences, this analysis was only performed on the 13 characters with 2-syllable names. We trained the logistic regression model to classify characters based on epochs where the character name was explictly spoken and tested the model on the pronouns associated with that character. Based on the results of analysis 2, we’d expect null results, which is what we find as seen in Figure 5. This is likely due to a lack of observable neural trace.

Discussion

Analysis 1

From analysis 1 we found clear evidence that both proper nouns and the 18 significant characters previously identified are differentiable from other types of nouns. Additionally, we found a difference in when clusters become statistically de-

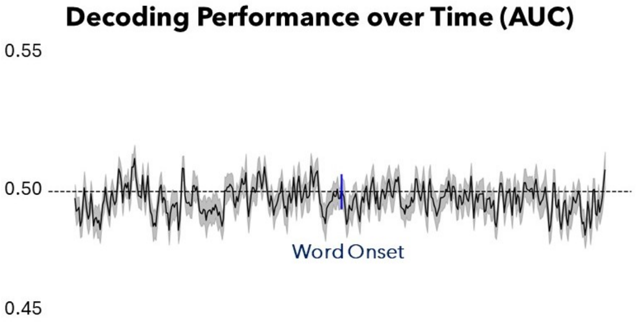

Figure 5: Timecourse of pronoun decodability based on character training,

codable, with proper nouns being decodable much later than common and pronouns. This result could align with the research from Heilbron et al (2022) which found that part of speech surpisal plays a word onset. Because pronominalization is so frequent in the text (as seen in Figure 1) the times in which a character name is used instead may lead to a lag due to the surprisal. However, this is just speculation and future work is needed to test this hypothesis rigorously.

Additionally this analysis is limited due to the class imbalance between proper nouns/characters and pronouns. Future analyses should subsample the classes so they have similar sizes.

Analyses 2 and 3

Both analysis 2 and 3 failed to find a significant neural link between referential nouns and their referents. This is likely due to a few factors. Though we know that humans are able to connect referents and referential nouns, previous research has shown that referential nouns are processed outside of the language centers of the brain in the medial parietal lobe (Brodbeck & Pylkkänen, 2017). Combined with the high variance environment created by the naturalistic story stimuli, the task becomes ill-suited for the logistic-regression based analysis performed in this paper. This is because it relies on an assumption that there is some sort of correlation in specific MEG channels that can be used to decode the neural signal above chance.

Outside of the problems with higher-level experimental design, there were also smaller issues with these analyses. In analysis 2, treating each character/pronoun combination as a different sample only works if we expect for each combination to exhibit similar neural responses (which the response from the differing audio stimuli alone breaks this assumption). However, we can't just look at sub-tabs of the data (like specific pronouns or character/pronoun combinations) because the sample size drops off dramatically causing variance to rise. This can be seen in the second graph in Figure 4, and makes it hard to draw any meaningful results. Future analyses should consider bootstrapping the data in order to generate more balanced samples and reduce variance when exploring segments of the sample population.

In analysis 3, besides issues with differing audio stimuli between characters adding to noise, we were also severely limited by the number of epochs correlated to explicit mentions of characters. Bootstrapping could once again help, but the difference is likely marginal.

Finally, the added variance from the error in coreference tagging by fastcoref likely harmed the analysis. Not only did it add noise from misclassified words, but it also reduced the size of our already small sample set.

All of this is to say, these results aren't definitive. Though it's possible that a connection between referential nouns and their referents could be found using MEG, a more controlled environment specifically crafted to balance pronouns and character mentions would like be needed in order to control for a lot of these confounding factors. Additionally, since other research indicated that coreference is a memory-based process, fMRI may be a better modality for this task as spatial resolution is probably more impactful then temporal resolution.

Conclusion

This study examined the neural dynamics of pronoun resolution within natural language processing, focusing on the decoding of pronominalization in narrative contexts using MEG data. Our investigation revealed significant neural decodability for pronouns and common nouns prior to word onset. Proper nouns, however, exhibited delayed decodability, suggesting different cognitive mechanisms related to POS surprisal at play.

Despite these findings, we did not observe significant neural traces linking pronouns to specific characters nor did we find generalizable neural signals between characters and their pronouns. This aligns with previous research indicating that coreferential processing may occur in brain regions outside the traditional language centers.

The complexities and limitations encountered, such as class imbalance, annotation inaccuracies, and the high variability of naturalistic stimuli, suggest that future studies may benefit from more controlled environments or alternative imaging techniques like fMRI. These findings underscore the challenges in decoding natural language processing but also highlight pathways for refining our understanding of coreference resolution in the brain. Continued exploration with improved methodologies will be crucial for unraveling the intricate neural processes underlying language comprehension.

Acknowledgments

I'd like to thank Laura Gwilliams and the other students of Data Science 125 for their continued help and support. I greatly appreciate it and have learned a significant deal through the process.

References

Brodbeck, C., & Pylkkänen, L. (2017). Language in context: Characterizing the comprehension of referential expressions with MEG. *NeuroImage, 147*, 447-460. https://doi.org/10.1016/j.neuroimage.2016.12.006

Li, J., Wang, S., Luh, W.-M., Pylkkänen, L., Yang, Y., & Hale, J. (2021). Cortical processing of reference in language revealed by computational models. *bioRxiv*. https://doi.org/10.1101/2020.11.24.396598

Brodbeck, C., Gwilliams, L., & Pylkkänen, L. (2016). Language in context: MEG evidence for modality-general and -specific responses to reference resolution. *ENeuro*. https://doi.org/10.1523/ENEURO.0145-16.2016

Brilmayer, I., & Schumacher, P. B. (2021). Referential chains reveal predictive processes and form-to-function mapping: An electroencephalographic study using naturalistic story stimuli. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.623648

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *PNAS, 119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Armeni, K., Güçlü, U., van Gerven, M., & Schoffelen, J.-M. (2022). A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data, 9*, 278. https://doi.org/10.1038/s41597-022-01382-7