Emma Chabane

Final Project Report

Revealing surprisal neural representation in real-world language processing

Abstract

Understanding language processing and its neural representation within the human brain still poses a challenge to neuroscientists. Indeed, drawing significant results about language structure and its representation in the brain requires investigation methods with high spatiotemporal resolution and a large amount of data. We resolve these issues through our one-of-a-kind BrainTree-bank corpus, the only large scale treebank accompanied by neural data. We recorded more than 55 hours of intracranial recordings of children and adolescents watching Hollywood movies, using implanted electrodes. With BrainTree-Bank, we investigated the neural basis of surprisal within the brain and explored the relationship between surprisal and cognitive activity. Our study identified a robust link between surprisal and neural activity within the functional language regions of the brain. Our results suggest words with higher surprisal elicit stronger electrical responses, therefore strengthening the relationship between cognitive activity and surprisal.

Background

Humans process language incrementally (Rayner, 2008). With each word we parse when reading, talking to each other, or watching a movie, humans build a linguistic landscape using the information they have been processing. Using this linguistic and contextual information, our brains are able to create expectations of what words may come next.

Surprisal is a quantity therefore used to operationalize word expectations within a larger linguistic scope. Surprisal is a measure of how surprising or unexpected an event is. Surprisal plays a crucial role in language processing, helping humans parse and better understand language. As such, when linguistic expectations are violated, surprisal is higher and processing is more difficult (Levy, 2008). Because the brain uses previous linguistic information to predict what happens next in a sentence and conversation, breaking these built expectations heightens surprisal, and also helps the brain update its predictions. There is a reverse relationship between surprisal and word probability within the linguistic context. Furthermore, we also use surprisal to disambiguate words and sentences, using the context of the sentence and its expectations to determine the correct meaning. Sentences that are more syntactically complex yield to higher surprisal values (Levy, 2012).

Recent research has aimed to study the basis of surprisal in the brain. Using fMRI, the neural basis of surprisal was researched and certain brain regions activated during sentence and context processing (Bhattasali, 2021). Moreover, strong relations between surprisal and ERP components were found using EEGs (Brouwer, 2021). Nonetheless, our research pushes past these approaches. The two former papers fall short in the methodology used. Indeed, the use of fMRI and EEGs are bonafide impediments in analyzing the neural basis of surprisal. While EEGs have decent temporal resolution, they have poor spatial resolution. The reverse holds for fMRIs. Instead, our dataset is instead highly spatiotemporally accurate, and is the largest dataset of its kind.

Approach

BrainTree-Bank background

Subjects (n=10, 5 female, mean age= 11.9) aged 4 to 19 under treatment for epilepsy at the Boston Children's Hospital participated in the data collection. These subjects watched a total of twenty-eight Hollywood movies (ranging from Marvel to Disney movies, such as Cars 2 and Spider-Man). Each subject was implemented with an average of 169 electrodes through stereoelectroencephalography (SEEG), resulting in highly accurate intracranial recordings of the subjects brains. Moreover, the movie transcripts were annotated and parsed at the millisecond level. This resulted in the BrainTree-Bank corpus, the only large scale treebank accompanied by neural data. The dataset has more than fifty five hours of intracranial recordings, and high language and speaker variability. The BrainTree-Bank will henceforth be used for all our analysis of the neural basis of surprisal.

Procedure

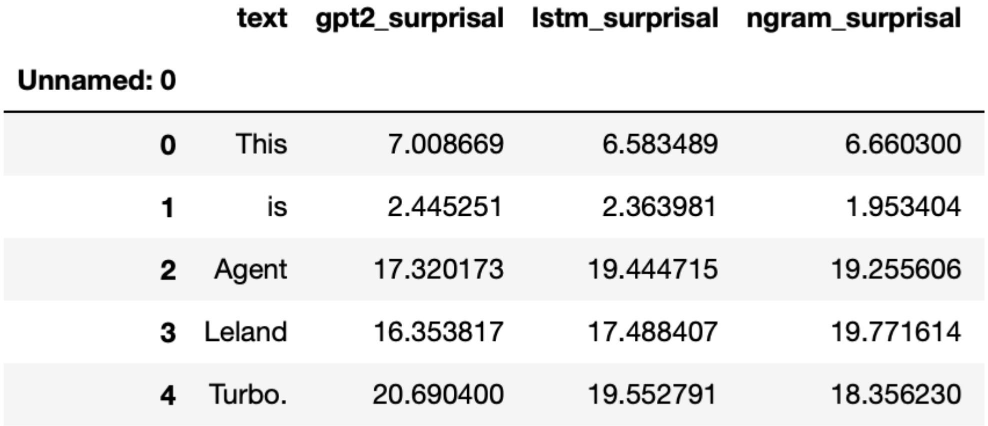
Initially, we will start by choosing a relevant word probability model to calculate suprisal on the words of our dataset. There are multiple models we can choose from, such as GPT-2, LSTM language models, or even n-grams models. We will investigate which model provides the best approach using statistical, qualitative, and quantitative analysis. Then, we will extract the movie transcripts we will choose to run our research on: we currently have annotated datasets for all of our movies, with many linguistic annotations at the word level (e.g. part-of-speech). After having computed the surprisal scores for all words in our chosen transcript, we will use those computations and match them with our intracranial recordings, to begin to look into what specific regions are active when surprisal is high. We will plot neural activity by distinguishing between low and high surprisal utterances, and compare the voltage activity within the brain.

Surprisal calculations

Surprisal was quantified as the log of the inverse of the probability of a word given the context. Word probabilities can be calculated through a variety of natural language processing methods. In this paper, we focused on three main language models: GPT-2 probabilities, LSTM probabilities, and n-gram probabilities. GPT-2 probabilities were computed using the Hugging Face Transformers library with GPT-2 large. Our LSTM model had 2 layers, a dropout of 0.2, and 200 input/output dimensions. Dropout helps us regularize our model by reducing overfitting and generally improving model performance. The LSTM probabilities were pre-trained on

BLLIP probabilities. The Brown Laboratory for Linguistic Information Processing (BLLIP) dataset is a language corpus containing a tagged and parsed treebank corpus, for an overall thirty million words. Finally, we calculated n-gram probabilities using a 5-gram model. The Python KenLM model was used, and was pre-trained over BLLIP much like the LSTM probabilities. Word particles, words that fall outside of the main parts of speech, were combined by summation.

Every word in the chosen movies (Cars 2, Lord of the Rings 1 & 2) had its surprisal score calculated using the GPT-2, LSTM, and n-gram probabilities. The surprisal scores for the first few words of Cars 2 are in Table 1.



	text	gpt2_surprisal	lstm_surprisal	ngram_surprisal
Unnamed: 0				
0	This	7.008669	6.583489	6.660300
1	is	2.445251	2.363981	1.953404
2	Agent	17.320173	19.444715	19.255606
3	Leland	16.353817	17.488407	19.771614
4	Turbo.	20.690400	19.552791	18.356230

Table 1: Surprisal scores computed on the Cars 2 movie transcript with our three word probability language models (a) GPT-2 (b) LSTM and (c) n-gram.

The difference in surprisal scores between the three different word probability models were not significant. In general, GPT-2 probabilities were lower than LSTM probabilities, which were themselves lower than n-gram probabilities. Ultimately, the minute differences in surprisal scores between the three did not affect the final results.

Surprisal plotting

We aligned the intracranial neural recordings with the surprisal scores and plotted the results using the matplotlib package in Python. On average, the mean of surprisal scores were 9.53, while their median was 8.28. Generally, simple and expected words tended to fall under a surprisal score of 10, while surprising and highly surprising words had scores in the late tens and up. The highest score we found within the movie transcripts we analyzed was 53. After a qualitative and quantitative analysis, we first chose to distinguish between high and low surprisal by setting the baseline at 8. This provided us with good results within the plots.

Results & Discussion

Our result suggested the existence of a strong correlation between surprisal scores and neural activity. We plotted neural voltage using our three different surprisal models, as well as three different movies. We plotted neural activity across different regions of the brain in order to find the functional areas for surprisal processing.

Figure 2 shows plots of the neural activity in the superior temporal gyrus region. Neural activity is divided between low and high surprisal in order to investigate the differences in cognitive effort and activity when words are expected and unexpected.

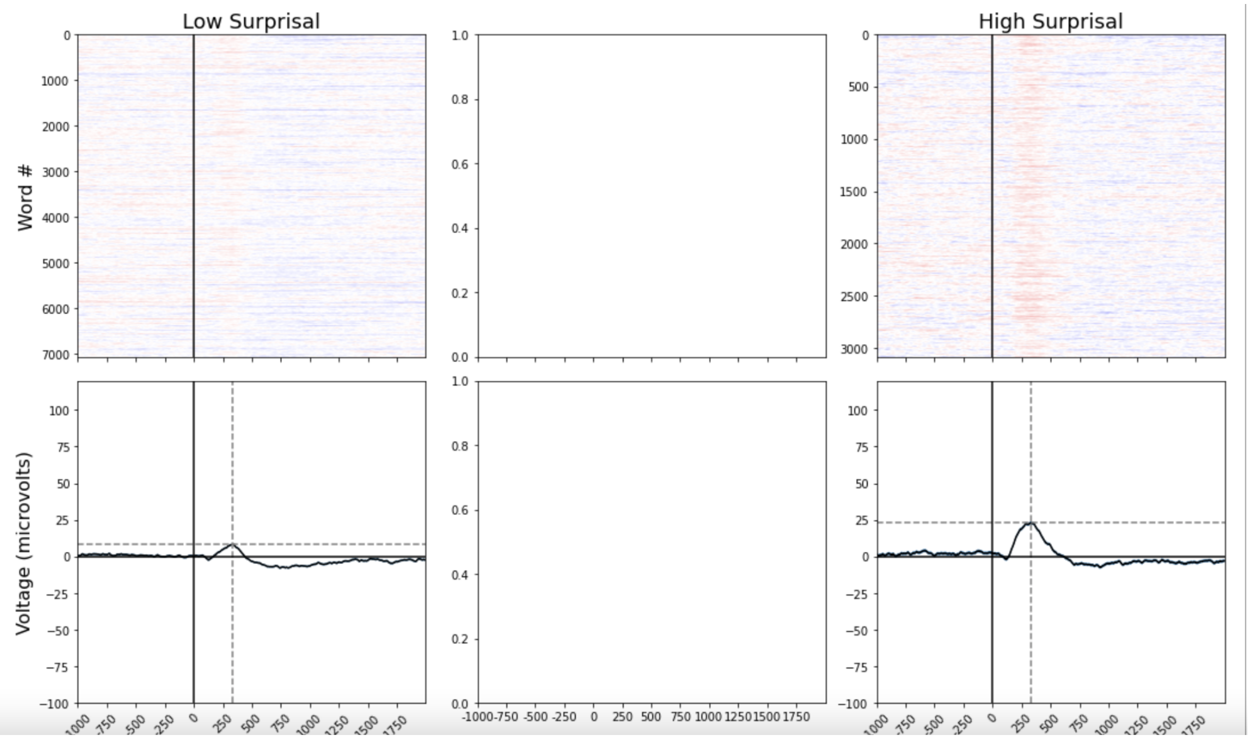


Figure 2: Raster plots and graphs of the neural voltage activity in the superior temporal gyrus region of the brain for Cars 2. Surprisal was computed as the negative log of word probability, using word probabilities from the GPT-2 language model.

This readily shows us that there is a strong neural basis for surprisal. On the raster plots, red denotes high voltage. Higher neural activity is more prevalent for high surprisal. This makes sense, and previous research on surprisal and language processing confirm this. Levy showed that violated linguistic expectations lead to a higher surprisal in language processing (2007). The peak voltage differences between low and high surprisal is of about fifteen microvolts, a significant difference between the language processing ability of our subjects.

These measurements were taken from the superior temporal gyrus, therefore showing that there is a basis for surprisal processing within that area. Previous research has shown the STG is involved in language and auditory processing (Bigler, 2007). Our results corroborate these findings, and suggest surprisal processing takes place within these same highly specific functional language areas. Indeed, linguistic functions have been shown to only happen in specific regions (Fedorenko, 2011). Plotting different electrodes throughout the brain showed that surprisal neural activity is only relevant in the already established language regions of the brain. Figure 3 shows an example of surprisal neural activity in a non-language specific region of the brain, the lateral orbitofrontal cortex. The latter is a region involved in decision-making, therefore validating Fedorenko's earlier findings.

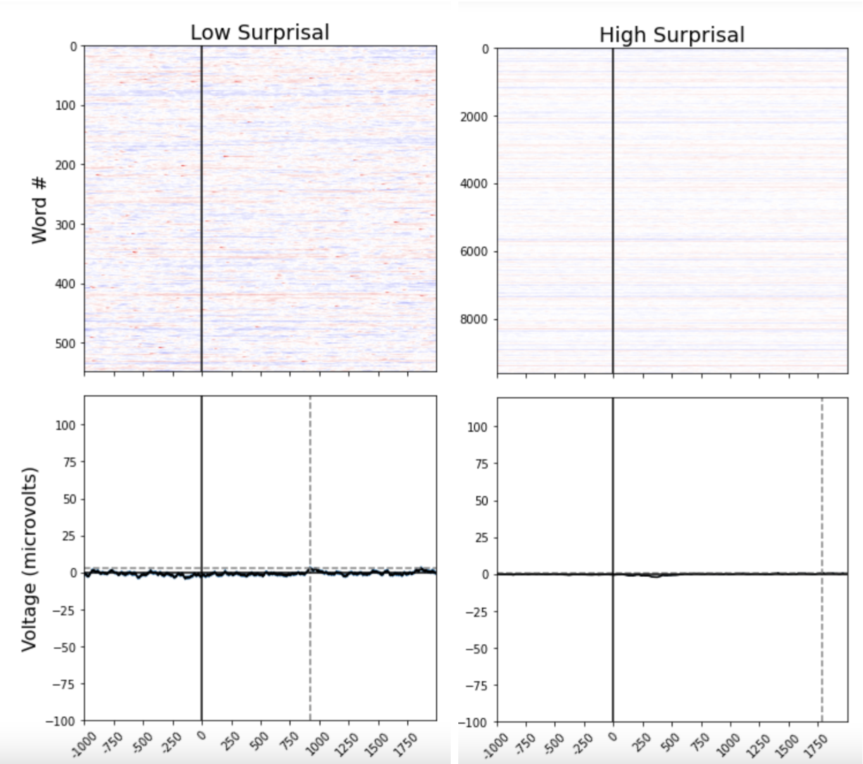


Figure 3: Raster plots and graphs of the neural voltage activity in the lateral orbitofrontal cortex region of the brain. Surprisal was computed as the negative log of word probability, using word probabilities from the GPT-2 language model.

It is quite clear from these results that high surprisal indeed correlates to higher neural activity and longer cognitive processing. Testing these results with other movies and word probabilities yield similar results, as seen in *Figure 4*. The latter shows raster plots and graphs from plotting differences in surprisal in Lord of the Rings, still using GPT-2 probabilities. In this plot, moments of high neural activity are once again prevalent at the onset of high surprisal words. In the low surprisal plot, neural voltage is less important and more spread-out.

Moreover, our results suggested that the difference in surprisal scores found earlier did not significantly impact our plots. Indeed, neural activity results stay similar. We indeed would expect surprisal scores to stay within a similar range for all three

If surprisal varied heavily based on the language model we used, this would indicate a flagrant error within the way surprisal and word probabilities were calculated, and would have most likely resulted from an error on our end. Highly volatile surprisal scores between models would have otherwise shown that surprisal may change between individuals in a non-negligeable manner. This would have been quite an interesting finding, not in line with current research done on surprisal. Thus, this similarity between surprisal scores is essential to validate our findings, as well as confirm previous ideas in the field of language processing.

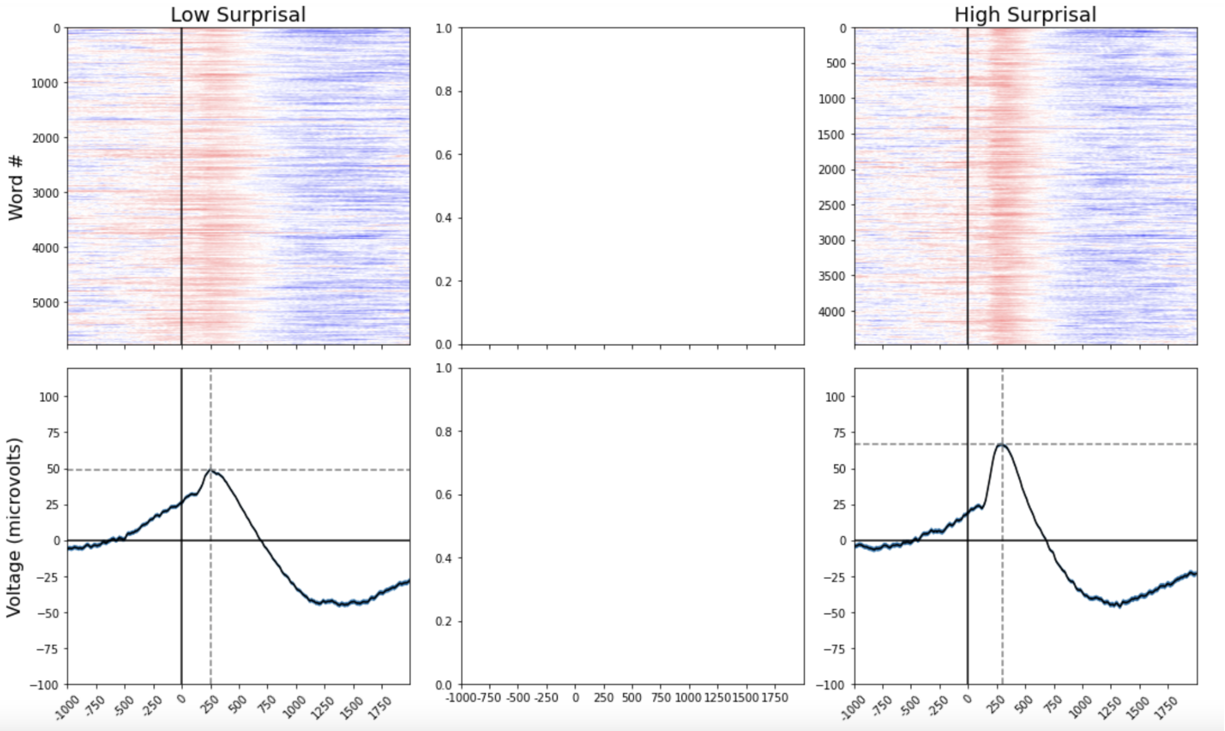


Figure 4: Raster plots and graphs of the neural voltage activity in the superior temporal gyrus region of the brain for LOTR 2. Surprisal was computed as the negative log of word probability, using word probabilities from the GPT-2 language model.

Conclusion

The present study aimed to investigate the neural basis of surprisal and explore the relationship between surprisal scores and cognitive activity. Our findings completely met these expectations. Indeed, our results confirm that language regions are highly specific, and language processing, specifically surprisal, has a neural basis within these regions.

Moreover, our research suggests that surprisal is highly correlated to neural activity within the language regions of the brain. Higher surprisal is caused by words our brains expect to be less likely to appear within the linguistic context. These less probable words lead to higher cognitive activity and effort within the brain, as shown by the higher levels of neural activity.

Through this research, we add onto the growing body of research on the neural representation of language and the linguistic, visual, and auditory components that work together to allow humans to use and understand language. Specifically, studying surprisal provides insights into the way humans constantly process language, and how past utterances can affect both future utterances and linguistic understanding.

Future plans

Future research might benefit from looking into the impact of age on surprisal processing. Our subjects encompass a range from early childhood to late adolescence, thus giving us plenty of data on how age could affect language processing and understanding. We could ask ourselves how surprisal neural activity changes through children and adolescence cognitive development. A priori, there could be differences with the way four years old process language compared to nineteen year olds, and neural activity analysis and modeling could address this.

Moreover, an intriguing research path could be to study the processing of ambiguous language and garden-path sentences within the brain. How does the neural activity of subjects change when language becomes harder to understand, either due to syntax, semantics, or pragmatics? This would require a whole new study and a new dataset, but could yield promising results in understanding linguistic ambiguity.

Finally, we hope to push this research forward by building computational models of surprisal processing within the brain. We will first start by building regression models capable of predicting surprisal scores based on neural recording. Building machines that process language like humans could change the way language models understand and parse language.

I am planning on MEnging with the InfoLab, and will be pursuing a UROP with them in the spring. Hopefully, I will get to work on similar work (or the same research). Thank you to the 9.58 staff for giving us the opportunity to meet and hear from so many insightful and talented people!

Acknowledgements

I would like to thank Andrei Barbu for mentoring this project and providing help and background information, as well as Christopher Wang and Vighnesh Subramaniam for their insights on the project and their support with getting set up. A huge thank you to Brian and Yudi for being such an integral part of 9.58, and Karen Pepper for being a great resource in learning how to better communicate!

References

1. Barbu, Andrei, et al. “The Compositional Nature of Event Representations in the Human Brain.” Center for Brains, Minds and Machines (CBMM), ArXiv, 2014, dspace.mit.edu/handle/1721.1/100175.
2. Bigler, E. D., Mortensen, S., Neeley, E. S., Ozonoff, S., Krasny, L., Johnson, M., Lu, J., Provencal, S. L., McMahon, W., & Lainhart, J. E. (2007). Superior temporal gyrus, language function, and autism. *Developmental neuropsychology*, *31*(2), 217–238. https://doi.org/10.1080/87565640701190841
3. Bhattasali, Shohini, and Philip Resnick. “Using Surprisal and FMRI to Map the Neural Bases of Broad and Local Contextual Prediction during Natural Language Comprehension.” *ACL Anthology*, Association for Computational Linguistics, 2021, https://aclanthology.org/2021.findings-acl.332.pdf.
4. Brouwer, Harm et al. “Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model.” Frontiers in Psychology, doi:10.3389/fpsyg.2021.615538.
5. Fedorenko, Evelina et al. “Functional specificity for high-level linguistic processing in the human brain.” Proceedings of the National Academy of Sciences of the United States of America vol. 108,39 (2011): 16428-33. doi:10.1073/pnas.1112937108
6. Kumar, S.*, Sumers, T. R.*, Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A.. “Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model.” 2022, https://www.biorxiv.org/content/10.1101/2022.06.08.495348v2.
7. Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006
8. Pylkkänen, Liina. “The neural basis of combinatory syntax and semantics.” *Science*, 2019, https://www.science.org/doi/10.1126/science.aax0050.
9. Rayner, K., & Clifton, C., Jr (2009). Language processing in reading and speech perception is fast and incremental: implications for event-related potential research. *Biological psychology*, *80*(1), 4–9. https://doi.org/10.1016/j.biopsycho.2008.05.002