Hierarchical syntactic structure in human-like language models

Michael Wolfman
University of Georgia
michael.wolfman@uga.edu

Donald Dunagan
University of Georgia
dgd45125@uga.edu

Jonathan Brennan
University of Michigan
jobrenn@umich.edu

John T. Hale
University of Georgia
jthale@uga.edu

Abstract

Language models (LMs) are a meeting point for cognitive modeling and computational linguistics. How should they be designed to serve as adequate cognitive models? To address this question, this study contrasts two Transformer-based LMs that share the same architecture. Only one of them analyzes sentences in terms of explicit hierarchical structure. Evaluating the two LMs against fMRI time series via the surprisal complexity metric, the results implicate the superior temporal gyrus. and This underlines the need for hierarchical sentence structure in word-by-word models of human language comprehension.

1 Introduction

Interest in language models (LMs) has exploded due to their recent success on language-related tasks (Min et al., 2021), with many commentators speculating about their implications as models of human language processing (see Millière, 2024, §IV.ii, for a review). The methodological utility of natural language processing tools for isolating language-processing functions in the brain is by now well-established (Brennan et al., 2012; Wehbe et al., 2014; Henderson et al., 2016; Shain et al., 2020; Stanojević et al., 2023); however, controversy persists regarding the role of hierarchical structure as useful or not in characterizing human language comprehension (e.g., Frank et al., 2012; Christiansen and Chater, 2015), yielding two related questions.

1. Is hierarchical structure part of the best description of human language comprehension?

2. If so, what brain regions subserve this aspect of processing?

This study investigates these questions by comparing two language models with the same underlying architecture. One is constrained via a special attention mask that captures hierarchical structure in the form of syntactic constituency, while the other lacks this attention mask, capturing only word-level information. The hierarchy-biased model is a Transformer Grammar (TG; Sartran et al., 2022), which differs only from the unconstrained model, Transformer-XL (TXL; Dai et al., 2019), in the presence of this attention mask.

We pair these language models with surprisal, a word-by-word information-theoretic complexity metric (see Hale, 2016, for a review) to derive predictions about neuroimaging data. Surprisal from the hierarchy-biased TG compares against surprisal from the unconstrained TXL in the task of predicting fMRI data (Li et al., 2022). This sets up a clean contrast between hierarchical and non-hierarchical conceptions of language comprehension.

The results, reported in section 6, support the role of hierarchical structure in language comprehension. Surprisal values derived from a Transformer Grammar predict fMRI timecourses in bilateral superior temporal gyrus (STG) better than those from TXL. This supports the view that the STG is sensitive to hierarchical sentence structure (Friederici and Gierhan, 2013; Friederici, 2017).

2 Phrase Structure

The Penn Treebank operationalizes one notion of hierarchical structure (Marcus et al., 1993). The present study uses these trees, exemplified in Figure 1. The syntactic analyses that they express date back to Chomsky's Standard Theory (1965) and can be motivated by considerations such as substitution, compositionality and structure-dependence of transformational rules which are reviewed in introductory linguistics textbooks (e.g. Akmajian et al., 2010). For a broad, comparative discussion of hierarchical structure in language, see Coopmans et al. (2023).

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 72–80
August 15, 2024

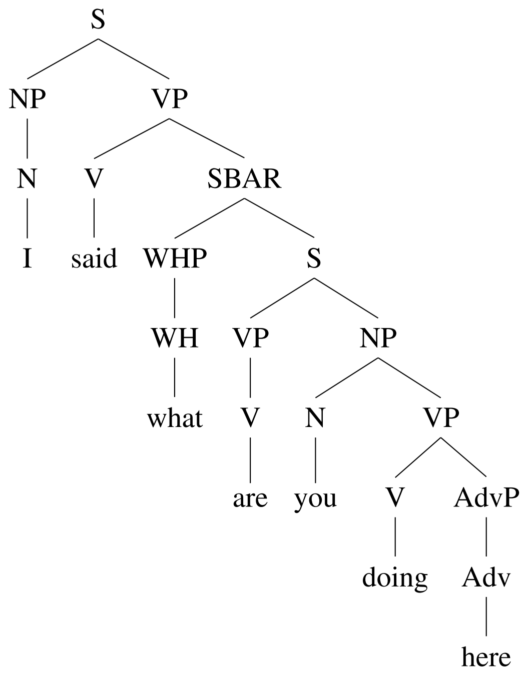

Figure 1: An example sentence attested in the stimulus text (*The Little Prince*) used in the fMRI study, see section 5.3.

3 Transformer Grammar

Transformer Grammars (Sartran et al., 2022) model the joint-probability of a surface string x and its corresponding phrase structure tree y, $p(x, y)$. They incorporate an inductive bias toward hierarchical syntax via special attention masks. These attention masks mark the only difference between TG and a general Transformer-XL (Dai et al., 2019).

TGs apply the idea of parsing as language modeling (Vinyals et al., 2015; Dyer et al., 2016; Choe and Charniak, 2016) by assigning probability to labelled, bracketed strings. They innovate on that idea by restricting — via the additional attention mask – the information used in label assignment. This information is restricted to prior composed phrases and the direct subconstituents of the current phrase being composed. These restrictions result in stack representations that correspond to levels of a syntactic derivation (for more details on TG's recursive syntactic composition see Sartran et al., 2022, §2.1).

4 Previous Work Investigating Hierarchy using Computational Modeling and Neuroimaging Data

This work builds on research that compares word-by-word difficulty predictions against neuroimaging data. Previous work of this type has found support for hierarchical structure (Brennan et al., 2012, 2016; Henderson et al., 2016; Li and Hale, 2019; Shain et al., 2020; Reddy and Wehbe, 2021; Stanojević et al., 2023; Sugimoto et al., 2023; Oota et al., 2023). Hale et al. (2022) and Uddén et al. (2020, §2) review this interdisciplinary line of work from computational and neuroscientific perspectives, respectively.

Others, following in the tradition of Elman (e.g., 1990, see also Frank et al., 2012, Christiansen and Chater, 2015), have questioned the need for hierarchical structure. Proponents of this view point to the successes of LMs that rely just on overt word sequences in encoding (or decoding) human brain responses to language (e.g., Caucheteux et al., 2021a; Caucheteux and King, 2022; Toneva et al., 2022; see Karamolegkou et al., 2023 for a review). The most extreme form of this view holds that word-prediction alone suffices to explain human language processing (Schrimpf et al., 2021; Goldstein et al., 2022a).

The present study addresses this debate regarding the role of hierarchy in language comprehension by comparing two language models with the same underlying architecture, the only difference being that hierarchical structure is explicitly present (vis-a-vis the additional attention mask) in one (the TG) and not in the other (the TXL).

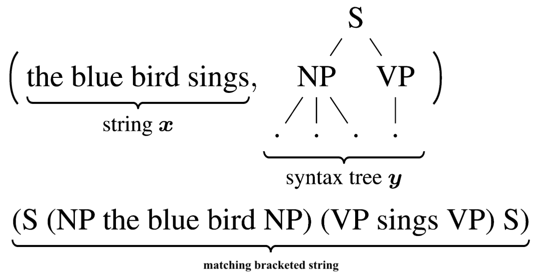

Figure 2: An example of a string x and tree y, which are modeled by a labelled bracketed sequence of (x,y) (Adapted from Sartran et al., 2022, Figure 1).

5 Methodology

5.1 Language Modeling

A 252M parameter, 16-layer, 8-attention-head TG was used as the hierarchy-biased model.[1] A 252M parameter, 16-layer, 8-attention-head TXL (Dai et al., 2019) was used as the unconstrained lan-

[1] https://github.com/google-deepmind/transformer_grammars

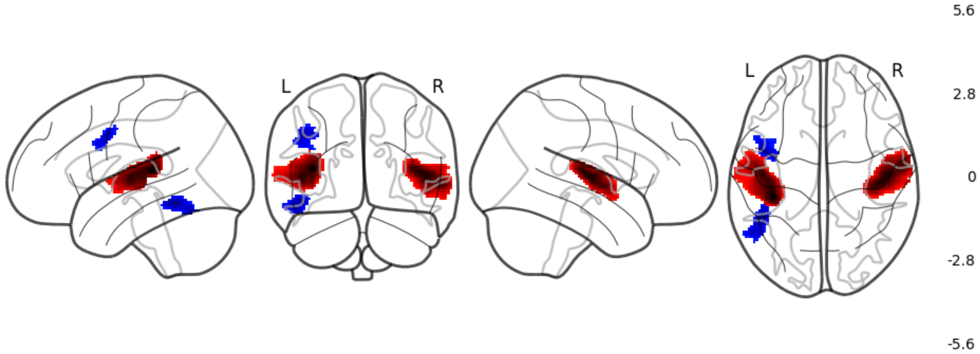

Figure 3: Glass brain z-map showing significant clusters of r^2 increase for hierarchy-biased TG surprisal (red) or unconstrained TXL surprisal (blue), thresholded with an expected false discovery rate (FDR) < 0.05 and a cluster threshold of 50 voxels.

guage model.[2] Both models were trained on the BLLIP-LG dataset (Charniak et al., 2000), as split by Hu et al. (2020). The training set is comprised of 1.8M sentences (≈40M words). Tokenization was performed with SentencePiece (Kudo and Richardson, 2018) using a subword algorithm (Kudo, 2018) with a 32K word-piece vocabulary.

The only difference between the TXL and TG in this study is the additional attention mask. Their number of parameters, layers, attention heads, and training/evaluation data (excluding the annotations used for TG) are identical. Indeed, as reported in Table A.3, the trained models arrive at highly similar test set perplexities.

5.2 Linking assumptions

To link brain data to language models, we use the surprisal complexity metric (Hale, 2001; Levy, 2008). Surprisal is the negative logarithm of the conditional probability of the next token, given previous tokens, on a particular LM (for a review, see Hale, 2016). These per-token numerical values serve as theoretical predictions that may explain time-dependent neural signals from people hearing those words. In this case, the neural signal is the blood oxygen level dependent (BOLD) signal measured with fMRI at each voxel in the brain (see §5.3 below).

Whereas surprisal values from the string-oriented TXL are exact, surprisals from the tree-oriented TG are approximated using the top 300 trees sampled from a Recurrent Neural Network Grammar (Noji and Oseki, 2021).

5.3 fMRI

5.3.1 Data

The fMRI data analyzed was the the English section of the Little Prince Datasets (Li et al., 2022, N = 49). Participants were scanned while they engaged in the naturalistic task of listening to an audiobook recording of David Wilkinson's English translation of *Le Petit Prince* (*The Little Prince*), read by Karen Savage. Data collection protocols and preprocessing steps are reported in the cited paper.

5.3.2 Statistical Analysis

To assess both LMs with respect to human neuroimaging data, we pursue an r^2 analysis, following Crabbé et al. (2019, §5).

Single-Subject Statistics For each subject, we calculate how much the inclusion of the variables of interest—TG surprisal and TXL surprisal—increases cross-validated BOLD r^2 with respect to a base model with only predictors of non-interest. Here, r^2 values indicate the voxel-wise variance explained. Thus, at the first level, two brain maps are calculated for each participant: one indicating the increase in cross-validated brain activity r^2 associated with adding TG surprisal to a baseline model; and one indicating the increase in cross-validated brain activity r^2 associated with adding TXL surprisal to a baseline model. The baseline model included: spoken word rate, word frequency, 5 principal components derived from fastText word vectors (Bojanowski et al., 2016), and the pitch and acoustic intensity of the narrator's voice.

BOLD signal is modeled, at each voxel, for each participant, via generalized linear model.

Region	Cluster size (mm^3)	MNI Coordinates X	Y	Z	Peak Stat (z)
Left Superior Temporal Gyrus (STG)	11208	-38.0	-32.0	10.0	5.57
		-46.0	-14.0	4.0	5.26
Right STG	10680	48.0	-18.0	6.0	5.36
		60.0	-10.0	0.0	5.09
Left Fusiform Gyrus	2224	-46.0	-48.0	-14.0	-4.48
		-52.0	-58.0	-18.0	-4.15
Left Pre-Motor Cortex	1552	-44.0	4.0	38.0	-4.16

Table 1: Results of paired T-test between hierarchy-biased and unconstrained cross-validated r^2 increase, thresholded with an expected false discovery rate < 0.05 and a cluster threshold of 50 voxels.

The word-level metrics are temporally annotated at the offset of each word in the audiobook, while the speech-related metrics are annotated every 10ms. All regressors, described in Table A.1, were convolved with the SPM canonical hemodynamic response function (Poldrack et al., 2011). Regressors of non-interest are included to ensure that any effects found are not due to other facets of linguistic processing (Lund et al., 2006).

Group-Level Statistics The single-subject r^2 increase brain maps (one TG map, one TXL map, per subject) were entered into a paired t-test to compare the impact of the additions of TG surprisal and TXL surprisal to base model of the BOLD signal. The results indicate where the addition of one variable to the base model (either TG surprisal or TXL surprisal) contributes to explaining the BOLD signal significantly better than the other.

6 Results

The addition of surprisal derived from the hierarchy-biased TG model performed above-and-beyond the addition of surprisal derived from the unconstrained TXL model in goodness-of-fit (r^2) to the measured BOLD signal in bilateral STG (Fig. 3; Table 1). The unconstrained model performed above-and-beyond the hierarchy-biased model in the left fusiform gyrus and pre-motor cortex. The significant clusters found were thresholded using an expected false discovery rate < 0.05 and a cluster threshold of 50 voxels.

7 Discussion

The findings support the role of STG in hierarchically-sensitive sentence processing (Friederici and Gierhan, 2013; Friederici, 2017). Notably, the results for surprisal in STG are largely localized to auditory cortex (see also Willems et al., 2016). These results suggest, in line with the sensory hypothesis (Dikker et al., 2009), that hierarchical structure from earlier in the sentence can impact low-level sensory processing. Prior investigation into early (< 150 ms) processing using MEG has found that auditory cortex is sensitive to phrase structure (Herrmann et al., 2009). This early sensitivity to hierarchical structure indicates that previously encountered structure may modulate sensory processing of subsequent words in a top-down manner. Employing a precise regression analysis and holding the architecture of LMs constant, the current study offers novel evidence in support of the sensory hypothesis and the early influence of hierarchical structure in language comprehension.

One region that has been largely implicated in predictive processing such as the type modeled here (e.g., Henderson et al., 2016; Brennan et al., 2020; Shain et al., 2020) is the left inferior frontal gyrus (LIFG). The present study does not implicate LIFG. It is possible that this null result could be due to the fact that the level of prediction and prediction violation here is too modest to invoke the LIFG, which seems more associated with processing particularly complex stimuli.

The success of modern LMs in natural language processing tasks has revived hope (see §4) that hierarchical structure could be left out of an adequate cognitive model. The results reported here suggest contrariwise. This echos Huang et al. (2024), who find that LMs strongly under-predict human reading time on syntactically challenging constructions, Antonello and Huth (2024) who differentiate LM layers that better-predict successor words from layers that better-predict fMRI data, and Yedetore et al. (2023), who find that unbiased LMs fail to generalize structurally-dependent constructions in a human-like way. With Antonello and Huth, we acknowledge that unconstrained LMs learn something about syntax. But it is not enough; in the context of cognitive modeling, additional bias towards hierarchical structure seems to be needed (Coopmans et al., 2022).

8 Conclusion

Hierarchical structure remains a key part of the best characterization of human language comprehension. This conclusion rests upon the increase in BOLD r^2 from the addition of TG-derived surprisal compared to the addition of TXL-derived surprisal. This obtains in a well-known temporal node of the language network and shores up the view of the language-processing brain as a system that performs hierarchical combinatorics. The results here also support recent arguments against unbiased LMs as cognitive models of human language.

Limitations

The TG (Sartran et al., 2022) and TXL (Dai et al., 2019) models used in this study are 16-layer models. A recent study from Mueller and Linzen (2023) found that depth (number of layers) is a more important factor in a language model's generalization performance than width (embedding and hidden dimensions, feed-forward layer size). Applying these findings to the present study by increasing the depth of the TG and TXL models could yield interesting results. It is possible that adding more layers to both models could affect the magnitude and presence of correlations to brain regions by influencing the generalization patterns of both TG and TXL. Given that the procedure here is theoretically motivated and the results align with both these theoretical considerations and previous neuroimaging work (e.g., the large scale brain model of Friederici, 2017), we do not expect the pattern of results to change. Nonetheless, further investigation is warranted.

This study only considers English. Follow-up studies could be performed in additional languages to solidify and expand the conclusions drawn here.

Finally, as previously mentioned, it has been found (e.g., Toneva and Wehbe, 2019; Caucheteux et al., 2021a; Caucheteux and King, 2022) that intermediate layers of LMs are best at encoding neural data. An interesting follow-up to the current study could probe the representations learned by TXL in its earlier layers and compare how well they encode neural data against a TG.

Ethics Statement

Language models pose risks when used outside of their intended scope. The language models used here are available under a CC-BY 4.0 license, allowing free public use. The training data used here (Charniak et al., 2000) is semi-controlled in that it comes from the Wall Street Journal; however, it is generally important to investigate training data for harmful human bias, which could find its way into language models.

References

Adrian Akmajian, Richard A. Demers, Ann K. Farmer, and Robert M. Harnish. 2010. *Linguistics: an introduction to language and communication*. MIT Press.

Richard Antonello and Alexander Huth. 2024. Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language (Cambridge, Mass.)*, 5(1):64–79.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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

Jonathan R. Brennan, Chris Dyer, Adhiguna Kuncoro, and John T. Hale. 2020. Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146:107479.

Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157-158:81–94.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021a. Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1336–1348. PMLR.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10. Publisher: Nature Publishing Group.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. BLLIP 1987-89 WSJ Corpus Release 1. Artwork Size: 1048576 KB Pages: 1048576 KB.

Do Kook Choe and Eugene Charniak. 2016. Parsing as Language Modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2331–2336. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.

Morten H. Christiansen and Nick Chater. 2015. The language faculty that wasn't: a usage-based account of natural language recursion. *Frontiers in Psychology*, 6. Publisher: Frontiers.

Cas W. Coopmans, Helen de Hoop, Karthikeya Kaushik, Peter Hagoort, and Andrea E. Martin. 2022. Hierarchy in language interpretation: evidence from behavioural experiments and computational modelling. *Language, Cognition and Neuroscience*, 37(4):420–439. Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2021.1980595.

Cas W. Coopmans, Karthikeya Kaushik, and Andrea E. Martin. 2023. Hierarchical structure in language and action: A formal comparison. *Psychological Review*, 130(4):935–952.

Benoit Crabbé, Murielle Fabre, and Christophe Pallier. 2019. Variable beam search for generative neural parsing and its relevance for the analysis of neuro-imaging signal. 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 1150–1160. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context.

Suzanne Dikker, Hugh Rabagliati, and Liina Pylkkänen. 2009. Sensitivity to syntax in visual cortex. *Cognition*, 110(3):293–321.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211. Publisher: John Wiley & Sons, Ltd.

Stefan L. Frank, Rens Bod, and Morten H. Christiansen. 2012. How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, 279(1747):4522–4531.

Angela D. Friederici. 2017. *Language in Our Brain: The Origins of a Uniquely Human Capacity*. The MIT Press.

Angela D. Friederici and Sarah M. E. Gierhan. 2013. The language network. *Current Opinion in Neurobiology*, 23(2):250–254.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022a. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380. Number: 3 Publisher: Nature Publishing Group.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*, pages 1–8. Association for Computational Linguistics.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

John T. Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R. Brennan. 2022. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8(1):427–446.

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

Björn Herrmann, Burkhard Maess, Anna S. Hasting, and Angela D. Friederici. 2009. Localization of the syntactic mismatch negativity in the temporal cortex: An MEG study. *NeuroImage*, 48(3):590–600.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. 2023. Mapping brains with language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9748–9762, Toronto, Canada. Association for Computational Linguistics.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. ArXiv:1808.06226 [cs].

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. Publisher: Elsevier.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022. Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data*, 9(1):530.

Jixing Li and John Hale. 2019. Grammatical predictors for fMRI timecourses. In Robert C. Berwick and Edward P. Stabler, editors, *Minimalist Parsing*. Oxford University Press.

Torben E. Lund, Kristoffer H. Madsen, Karam Sidaros, Wen-Lin Luo, and Thomas E. Nichols. 2006. Non-white noise in fMRI: Does modelling have an impact? *NeuroImage*, 29(1):54–66.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Raphael Millière. 2024. Language models as models of language. *Oxford Handbook of the Philosophy of Linguistics*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *CoRR*, abs/2111.01243.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada. Association for Computational Linguistics.

Hiroshi Noji and Yohei Oseki. 2021. Effective Batching for Recurrent Neural Network Grammars. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4340–4352. Association for Computational Linguistics.

Subba Reddy Oota, Mounika Marreddy, Manish Gupta, and Raju Bapi. 2023. How does the brain process syntactic structure while listening? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6624–6647, Toronto, Canada. Association for Computational Linguistics.

Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols. 2011. Statistical modeling: Single subject analysis. In Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols, editors, *Handbook of Functional MRI Data Analysis*, pages 70–99. Cambridge University Press.

Aniketh Janardhan Reddy and Leila Wehbe. 2021. Can fMRI reveal the representation of syntactic structure in the brain? In *Advances in Neural Information Processing Systems*, volume 34, pages 9843–9856. Curran Associates, Inc.

Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. 2022. Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118. Publisher: Proceedings of the National Academy of Sciences.

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.

Miloš Stanojević, Jonathan R. Brennan, Donald Dunagan, Mark Steedman, and John T. Hale. 2023. Modeling Structure-Building in the Brain With CCG Parsing and Large Language Models. *Cognitive Science*, 47(7):e13312.

Yushi Sugimoto, Ryo Yoshida, Hyeonjeong Jeong, Masatoshi Koizumi, Jonathan R. Brennan, and Yohei Oseki. 2023. Localizing Syntactic Composition with Left-Corner Recurrent Neural Network Grammars. *Neurobiology of Language*, pages 1–48.

Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. 2022. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757. Number: 11 Publisher: Nature Publishing Group.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32.

Julia Uddén, Mauricio de Jesus Dias Martins, Willem Zuidema, and W. Tecumseh Fitch. 2020. Hierarchical structure in sequence processing: How to measure it and determine its neural implementation. *Topics in Cognitive Science*, 12(3):910–924.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. *Advances in neural information processing systems*, 28.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE*, 9(11):e112575. Publisher: Public Library of Science.

Roel M. Willems, Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, and Antal van den Bosch. 2016. Prediction During Natural Language Comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

A Appendix

Predictor	Description	Model-Inclusion
TG Surprisal	Surprisal derived from TG at a word	hierarchy-biased
TXL Surprisal	Suprisal derived from TXL at a word	unconstrained
Word Rate	Annotation indicating the existence of a spoken word	base, hierarchy-biased, unconstrained
Word Frequency	Log lexical frequency of a word	base, hierarchy-biased, unconstrained
F_0	Pitch (fundamental frequency) of the voice of the narrator	base, hierarchy-biased, unconstrained
RMS Amplitude	Root Mean Square Amplitude of the voice of the narrator (reflecting intensity)	base, hierarchy-biased, unconstrained
Word Vector_5	5 regressors corresponding to values derived from a word's pretrained fastText vector	base, hierarchy-biased, unconstrained

Table A.1: Generalized linear model predictors

Language Model	Perplexity on Test Set
Transformer Grammar (Sartran et al., 2022)	32.82
Transformer-XL (Dai et al., 2019)	34.07

Table A.3: Perplexity values for the TG and TXL language models on the BLLIP-LG test set, as split by (Hu et al., 2020).