Modeling Hierarchical Structure during Comprehension in English with Transformer Grammars

Introduction. Interest in large language models (LLM) has exploded lately and many commentators have speculated about their implications as models of human language processing (see [1] for a recent review) The methodological utility of language models for isolating language-processing functions in the brain is by now well-established [3,4]; however, controversy persists regarding the utility of hierarchical syntax in characterizing these language-processing functions [5,6]. Is phrase structure part of the best description of incremental language processing? This study investigates this question by comparing two language models with the same underlying Transformer-XL (TXL; [7]) architecture: one informed by hierarchical syntax and one not. Transformer Grammars (TGs; [2]), offer a novel Transformer language model architecture that captures phrase-structure syntax through recursive syntactic composition implemented via an attention mask. A syntax-informed TG is compared against a syntax-uninformed TXL in the context of fMRI data recorded while participants listen to a storybook via *surprisal*. Surprisal [8, 9] is a word-by-word complexity metric which can be calculated by generative models—like Transformers—and which has previously been shown to correlate with human language processing difficulty. We find that, for several language network regions, the addition of TG-derived surprisal to a baseline model results in a better fit to observed fMRI BOLD than the addition of TXL-derived surprisal, thus reaffirming the importance of hierarchical processing in human sentence processing. **Materials.** The fMRI data analyzed was the English component of the *Little Prince Datasets* [10]. Participants (N=49) were scanned while they engaged in the naturalistic task of listening to an audiobook recording of a children's story, *The Little Prince*. Word-by-word syntax-informed surprisal values were calculated using a TG model trained on the BLLIP-LG phrase structure treebank as prepared by [11]. Word-by-word syntax-uninformed surprisal values were calculated using an equivalently-sized TXL model, trained on the BLLIP-LG corpus. **Methods.** The analysis compared ability between TG and TXL surprisal to account for variance in the observed BOLD signal during naturalistic listening. For each subject, a base-level general linear model (GLM) containing linguistic regressors of non-interest was fit to each voxel. Two additional subject-level GLM models were estimated, one each adding TG or TXL surprisal to the predictors in the base-level model. For each subject, we computed how much the addition of the surprisal regressors increased cross-validated r^2 with respect to the base-level model. This results, for each subject, in two statistical brain maps: one indicating the increase in variance accounted for by TG surprisal and the other, the increase in variance accounted for by TXL surprisal. At the second level of the analysis, the subject-level TG and TXL surprisal r^2 increase maps were compared using a paired t-test. The resulting z-map is given in Figure 1 and was thresholded with a false discovery rate < 0.05 and a cluster threshold of 50 voxels. **Results.** The results of the paired t-test (Figure 1) indicate greater r^2 increase for TG surprisal compared to TXL surprisal in left pars opercularis (Broca's area; BA 44), right middle temporal gyrus (rMTG; BA21), and right pre-frontal cortex (rPFC; BA10). There were no regions in which TXL surprisal had greater r^2 increase than TG surprisal. **Discussion.** These results reinforce the role of Broca's area in syntactic structure processing, but also implicate right hemisphere homologues of left hemisphere language network regions. The rPFC and rMTG results are interpreted as supplementally supporting the traditional left hemisphere language network in two ways: 1) for processing syntactically difficult constructions; and 2) processing *quotation*, a complex metalinguistic phenomenon consisting of indirect speech [12, 13]. Indeed, 44% of the text of *The Little Prince* is in quotation. Further, these results reinforce the notion that human sentence processing in the brain involves more than just surface-level statistical patterns: it involves syntax.

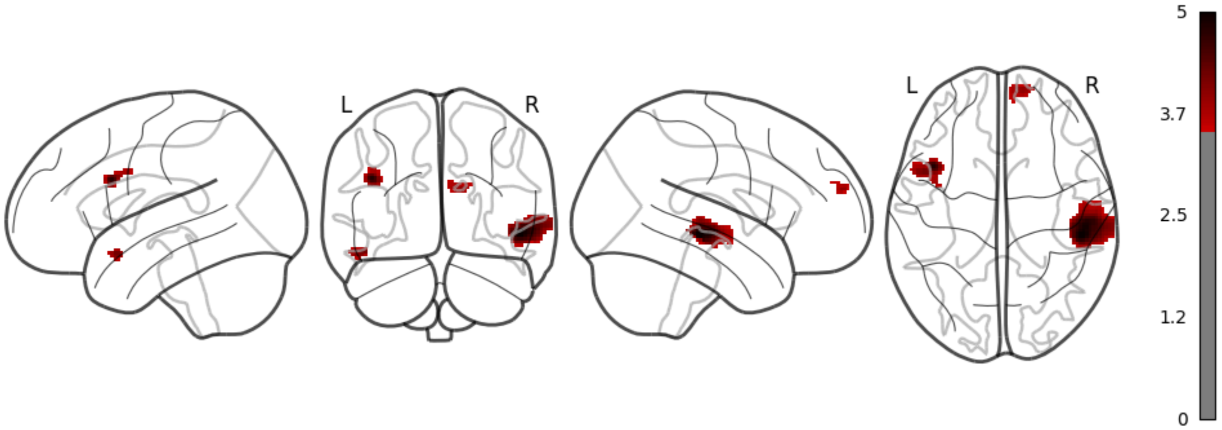

Figure 1: Results (z-valued) of the paired t-test between r^2 increase for syntax-informed Transformer Grammars surprisal and syntax-uninformed Transformer:XL surprisal, thresholded with an expected false discovery rate < 0.05 and a cluster threshold of 50 voxels. Red indicates significantly greater r^2 increase for Transformer Grammar surprisal > Transformer:XL surprisal.

References

[1] Millière, Raphaël. (2024). *Language Models as Models of Language*. To appear in *Oxford Handbook of the Philosophy of Linguistics.*

[2] Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, Chris Dyer. (2022). Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *TACL.*

[3] Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia 138*.

[4] Brennan, J. R., Dyer, C., Kuncoro, A., & Hale, J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, *146*.

[5] Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. Cognitive Science 47.

[6] Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition, 203*.

[7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. 57th *ACL Proceedings.*

[8] Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *2001 NAACL*.

[9] Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3).

[10] Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., Brennan, J. R., Yang, Y., Pallier, C., & Hale, J. (2022). *Le Petit Prince* multilingual naturalistic fMRI corpus. *Scientific Data*, 9(1).

[11] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. *58th ACL Proceedings.*

[12] Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex*, *24*(10).

[13] Brendel, E., Meibauer, J., & Steinbach, M. (2011). *Understanding quotation*. De Gruyter Mouton.