CCG parsing effort and surprisal jointly predict RT but underpredict garden-path effects

Satoru Ozaki
University of Massachusetts Amherst
sozaki@umass.edu

Aniello De Santo
University of Utah
aniello.desanto@utah.edu

Tal Linzen
New York University
linzen@nyu.edu

Brian Dillon
University of Massachusetts Amherst
bwdillon@umass.edu

1 Introduction

A prominent approach to explaining sentence processing difficulty is surprisal theory (Hale, 2001). In recent years, surprisal has often been estimated using large language models that do not have explicit representations of syntactic structures, let alone structure-building operations; even so, it predicts word-level difficulty in incremental processing (Oh et al., 2022). While surprisal theory has been prominent in the study of garden path effects as a model of ambiguity resolution, it has been shown to under-predict the magnitude of such effects in self-paced reading (van Schijndel and Linzen, 2021; Arehalli et al., 2022). On the other hand, models incorporating complexity metrics that take incremental structure-building operations explicitly into account have been shown to improve fit to eye-tracking (Demberg and Keller, 2008; Demberg et al., 2013) and neuroimaging data (Brennan et al., 2016; Stanojević et al., 2023).

Building on these lines of work, we first ask **(Q1)** whether a structure building-based complexity metric derived from a CCG (Combinatory Categorial Grammar) parser improves model fit to reading time data beyond surprisal estimates. We then explore **(Q2)** the extent to which this complexity metric can predict processing effort related to the recovery of temporally ambiguous sentences. While our metrics do not straightforwardly predict garden path effects, they predict processing effort in unambiguous sentences.

2 Q1: Fit to reading time beyond surprisal estimates

We first test whether explicitly considering structure-building operations from an incremental parser improves our ability to account for human behavioral data during sentence processing. Thus, we adopt an incremental parser for CCGs coupled with a metric that measures effort at each word as the number of nodes added to the parse tree upon processing that word (*node count*, Stanojević et al. 2023); see Figure 1 for an example.

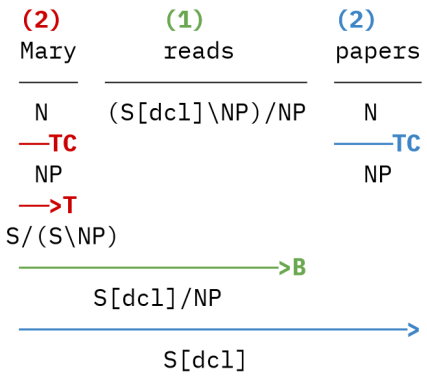

Figure 1: Node count example. Node count is indicated above each word. TC = type-changing, >T = forward type-raising, >B = forward composition.

CCGs offer a wide coverage of natural language, and node count derived from this particular incremental model has been shown to improve fit to neuroimaging data (Stanojević et al., 2023). We fit a linear mixed-effects model to the self-paced reading data available for the English filler sentences in the Syntactic Ambiguity Processing Benchmark (Huang et al., 2024). The predictors include node count, surprisal, as well as relevant lexical and orthographic control predictors. The surprisal values, taken from the Syntactic Ambiguity Processing Benchmark, are estimated from an LSTM model trained on 80 million tokens of English Wikipedia text (Gulordava et al., 2018). Node count of both the current word and the preceding word is associated with significantly slower RTs (Table 1).

3 Q2: Predictions for garden path constructions

Building on the results for Q1, we adopt the same CCG parser and implement a naive reprocessing

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 362-364.
Irvine, California, June 27-29, 2024

Effect & word pos.		Estimate	SE	*t*
Node	*i* th	0.66	0.27	2.52
count	*i-1* th	3.98	0.36	11.09
	i-2 th	-0.10	0.28	-0.35
Surprisal	*i* th	2.52	0.49	5.14
	i-1 th	3.61	0.37	9.82
	i-2 th	1.39	0.33	4.24

Table 1: Fixed effects from the best-fit linear mixed-effects model fitted to the self-paced reading time data of the English filler sentences in the SAP Benchmark. *i* th, *i-1* th and *i-2* th refer to the current, the previous and the previous previous word, respectively. Omitted lexical and orthographic control predictors: punctuation, word position, log frequency, length.

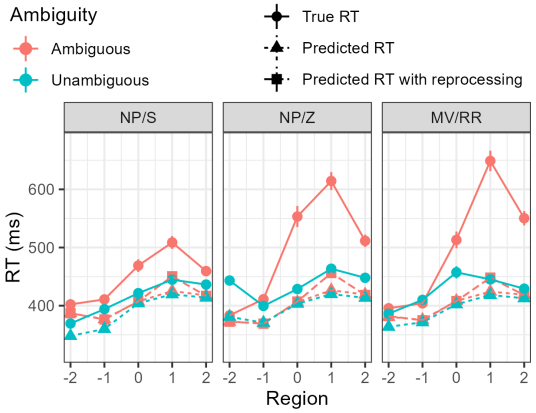

Figure 2: True RT for the three types of English garden-path sentences in the SAP Benchmark as well as RTs estimated by the model in Table 1. The x-axis indicates word position relative to the disambiguating word (Region 0). The solid line indicates true RT. The dotted line indicates RT estimated using CCG node count estimates without reprocessing, while the dashed line indicates RT estimated using effort values including reprocessing.

account of garden path recovery (Grodner et al., 2003): garden path difficulty is modeled as the total effort associated with reprocessing the entire string with the correct parse up to and including the word at which parsing breaks down. The disambiguating word is associated with this reprocessing effort in garden path constructions. We evaluate the model fit in Section 2 on three garden path constructions out of the set of critical contrasts in the Syntactic Ambiguity Processing Benchmark (NP/S; NP/Z; MV/RR). We find that our implementation of reprocessing consistently underestimates the magnitude of the garden path effects and, similarly to what is reported for surprisal estimates, it fails to predict differences across constructions (Figure 2).

4 Conclusion

Our results show that predictors relying on explicit structure-building operations improve our ability to model word-by-word reading times, independently of the contribution of surprisal measures — strengthening the evidence for structure building operations in a comprehensive model of human sentence processing. While our naive account of a dual-stage approach to ambiguity resolution underpredicts human effort in GP constructions, these results also showcase how this type of model can be used to implement various theories of sentence comprehension. In future work, this will allow for the evaluation of more fine-grained models of reanalysis processes as applied to GP effects.

5 Acknowledgments

This work was supported by the National Science Foundation, grant no. BCS-2020914 to BD.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *CoNLL 26*, pages 301–313.

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.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, Predictive Parsing with Psycholinguistically Motivated Tree-Adjoining Grammar. *Computational Linguistics*, 39(4):1025–1066.

Daniel Grodner, Edward Gibson, Vered Argaman, and Maria Babyonyshev. 2003. Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research*, 32(2):141—-166.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *NAACL 2*.

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.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5.

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.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.