Surprisal Outperforms Other Automatically Computable Predictive Metrics

Michael Vrazitulis (vrazitulis@uni-potsdam.de)[1], Kate McCurdy[2], Shravan Vasishth[1]
[1]Department of Linguistics, Potsdam University, [2]Department of Language Science and Technology, Saarland University

A longstanding goal in sentence processing research is to develop quantitative metrics that successfully predict processing difficulty. Over the past two decades, surprisal [1] has established itself as the dominant metric to account for non-trivial variation in behavioral and neurophysiological traces of human comprehension difficulty. With recent advances in language modeling technology, it is now easier than ever to compute surprisal values for arbitrary linguistic materials, e.g., from Transformer-based large language models like GPT-2 [2].

However, recent work introducing the large-scale SAP dataset [3] has found surprisal to dramatically underpredict the self-paced reading slowdown associated with syntactic disambiguation effects like garden paths. This raises the question whether other predictive metrics, beyond surprisal, would do a better job at accounting for syntactic disambiguation effects—and for processing difficulty more broadly. Towards this end, we examined the extent to which five different predictive metrics account for empirical effect sizes in the SAP dataset. Importantly, all examined metrics can be computed "hands-off" for arbitrary linguistic inputs, similarly to surprisal. These are the metrics we examined:

1. Resource-rational lossy context surprisal is surprisal but computed on an imperfect memory representation, as implemented by [4]. **2. Transformer attention entropy** [5] is computed from GPT-2 attention heads as a proxy for cue-based retrieval. **3. ACT-R interference** reimplements the cue-based retrieval model by [6], using categorical features from spaCy's [7] morphological parser. **4. DLT integration cost** is computed as defined by [8], based on Stanza [9] dependency parses. **5. BERT interference** is an interference metric based on BERT embedding [10] similarities.

We included each metric as a predictor in a separate log-normal mixed-effects model fitted to spillover-region reading times from the SAP dataset, alongside trivial predictors like word frequency, word length, and trial number. In the case of lossy context surprisal, this was done separately for 20, 50, and 80 % deletion rate variants. In the case of ACT-R interference, DLT integration cost, and BERT interference, we included **reanalysis cost** as an auxiliary predictor. We quantified reanalysis cost as the number of changes to the sentence's automatically computed [9] dependency parse tree before and after encountering the critical word. The BERT interference model was further enriched by a predictor summing up the **dependency distances** between the critical word and its previous codependents. Finally, a model with (plain) surprisal as a predictor was also fitted for comparison.

Figure 1 shows how the empirical effect sizes in the SAP dataset differ from model-predicted ones across seven experimental contrasts involving garden paths, attachment ambiguity, and agreement violation. None of the newly examined metrics is able to systematically outperform surprisal predictions. Instead, wherever surprisal already underpredicts empirical effect sizes (e.g., on MV/RR garden paths), other metrics tend to fare even worse. Further, DLT integration cost and ACT-R interference show a *negative* (i.e., opposite-to-expected) coefficient in their fit to human reading data.

These results indicate that surprisal still shows the highest predictive performance on key sentence processing phenomena, relative to a range of alternative, automatically computable metrics that are rooted in memory-based, reanalysis-based, or hybrid expectation–memory approaches.

References

[1] R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[3] K.-J. Huang, S. Arehalli, M. Kugemoto, C. Muxica, G. Prasad, B. Dillon, and T. Linzen. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510, 2024.

[4] M. Hahn, R. Futrell, R. Levy, and E. Gibson. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119, 2022.

[5] S. H. Ryu and R. L. Lewis. Memory for prediction: A Transformer-based theory of sentence processing. *Journal of Memory and Language*, 145:104670, 2025.

[6] R. L. Lewis and S. Vasishth. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29: 1–45, 2005.

[7] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, et al. spaCy: Industrial-strength natural language processing in Python, 2020.

[8] E. Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, 2000: 95–126, 2000.

[9] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 101–108, 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (long and short papers)*, pages 4171–4186, 2019.

Figures and Tables

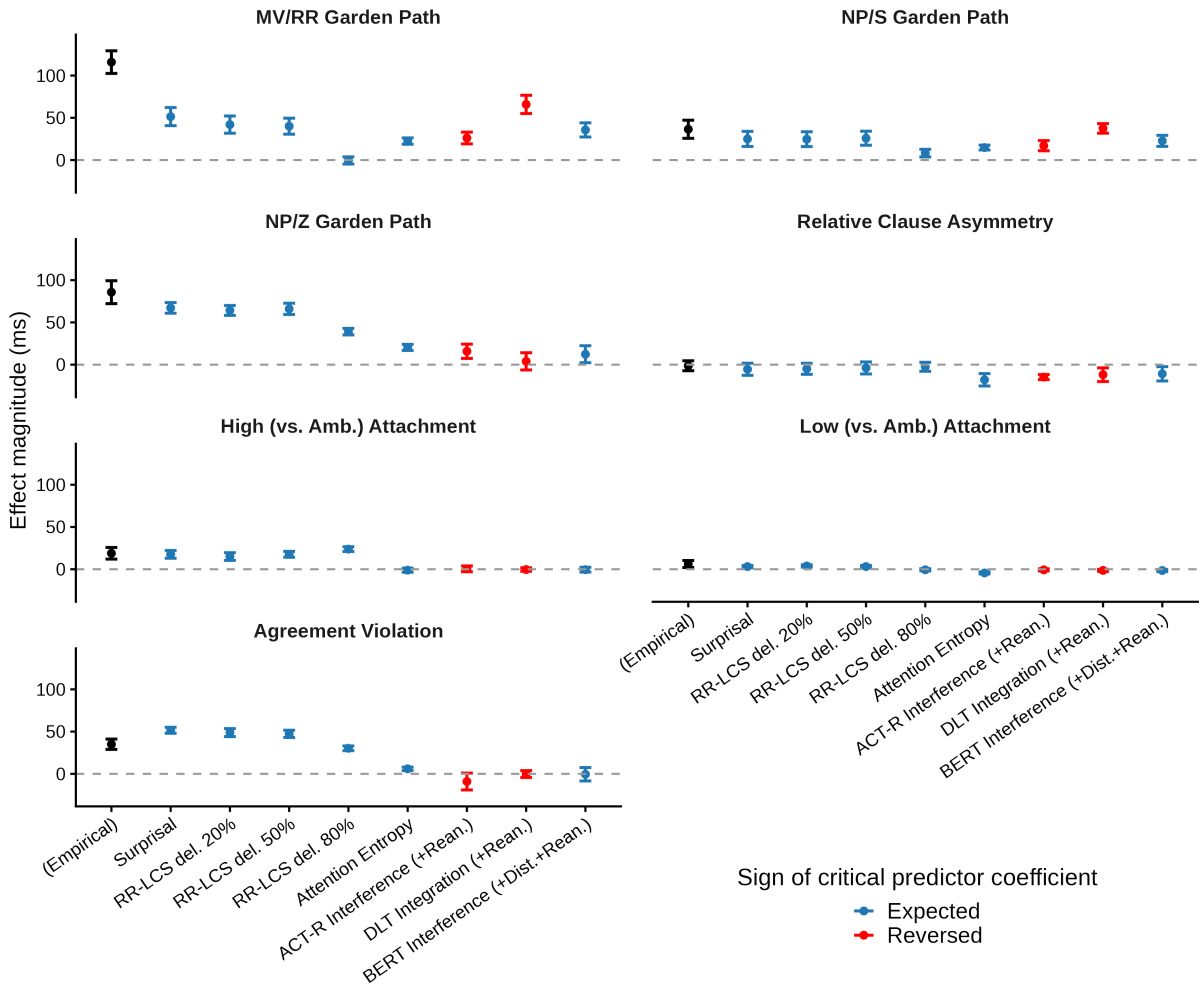

Figure 1: Empirical vs. model-predicted effects. Error bars indicate 95 % CIs.