Surprisal is Influenced by Syntax and Semantics, but not Equally across Language Models

Alessandro Lopopolo[1], Milena Rabovsky[1]
[1] Department of Psychology, University of Potsdam
lopopolo@uni-potsdam.de

Lexical surprisal, widely used to explain neural responses like the N400 and BOLD signal [1, 2], is often viewed as a measure of lexical prediction. However, it remains unclear how much it reflects syntactic and semantic structure beyond surface-level cues [3]. This study quantifies how surprisal is shaped by features such as semantic distance, and constituency and dependency structure (based on Figure 1A). We estimate word-wise surprisal from nine language models (GPT-2 [4], Falcon [5], BERT [6], RoBERTa [7], BART [8], T5 [9], and N-gram models) widely used in pyscholinguistics studies. Analyses were conducted on 60,000 English sentences.

Analyses: We fitted independent linear regression models predicting each of word-wise surprisal estimates from each of the nine LMs, using either (i) word position and lexical frequency alone (which yielded an average R^2 0.446 $\pm$ 0.082 across models) or (ii) structural (constituency, dependency) and semantic predictors alongside the baseline. Figure 1B shows the gains in explained variance (ΔR^2) from adding structural and semantic predictors. Constituency predictors included syntactic depth and the number of closed phrases per word; dependency predictors captured the number of dependencies per word and the distance to each word's dependency head. Dependency features yielded overall larger ΔR^2, constituency features produced moderate gains, and semantic distance – measured via FastText-based contextual dissimilarity –showed variable effects. GPT-2 and Falcon integrated both dependency and semantic information most strongly. BART showed weaker effects, particularly for constituency. BERT and RoBERTa benefited from syntax, though RoBERTa was less sensitive to semantics. N-gram models relied inconsistently on structure, with modest constituency and semantic effects but relatively strong sensitivity to dependency features, suggesting a shallow structural encoding. Additionally, we conducted SHAP (SHapley Additive exPlanations) analyses to assess each predictor's contribution to the nine surprisal estimates. SHAP quantifies feature impact by computing marginal contributions across feature subsets. Results (Figure 1C) show that semantic distance was the most influential predictor for transformer LMs, particularly GPT-2, Falcon, and BART. The number of left dependencies was key for T5 and N-gram LMs, while constituent depth notably impacted RoBERTa. Overall, transformers relied more heavily on semantic predictors, whereas N-gram models showed a more balanced or weaker feature profile. Hierarchical clustering confirmed distinct grouping patterns between transformer and N-gram architectures based on predictor sensitivities.

Conclusions: Overall, despite the conspicuous explanatory effects of the baseline predictors, structural predictors enhance surprisal estimates, confirming that surprisal reflects not just surface-level properties. However, the nature and extent of this sensitivity is not universal but varies sharply across models, reflecting their architectural and training distinctions. These findings highlight the importance of evaluating surprisal's informational content in model-specific terms when used to interpret cognitive or neural data, and they encourage an empirically driven definition of surprisal—one that reflects how different models relate to both sequential and structural information. In ongoing work, we test whether model-specific sensitivities to structure and semantics shape how surprisal maps onto neural correlates of predictive processing.

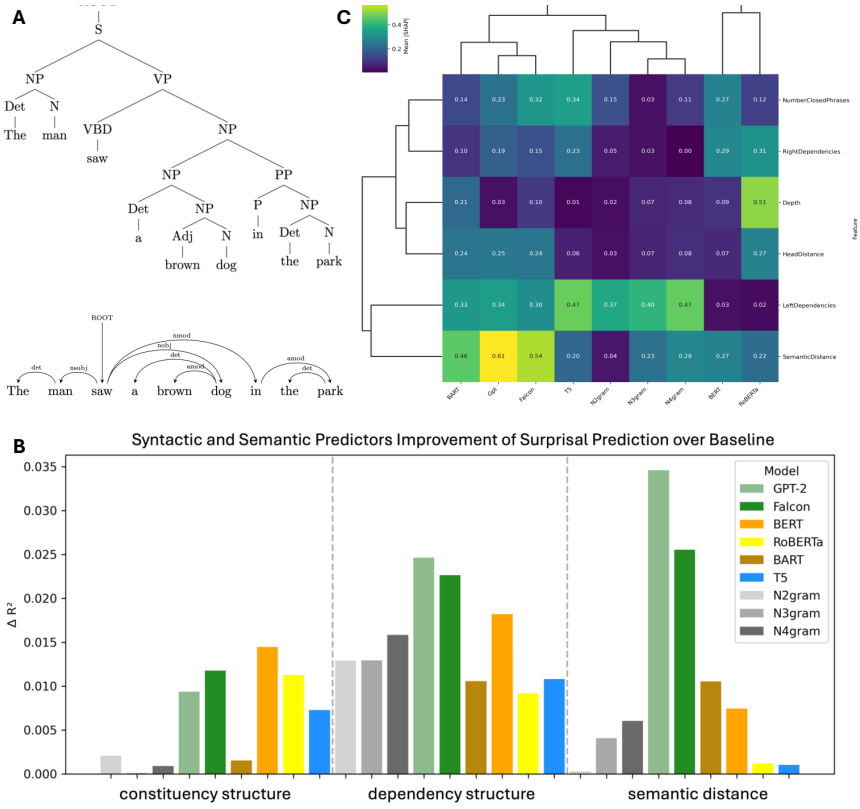

Figure 1: (A) Constituency and Dependency structure representations of a sentence; (B) ΔR^2 from linear models predicting surprisal across 9 language models. Each bar shows the increase in R^2 over a baseline using position and lexical frequency, after adding constituency, dependency, or semantic predictors; (C) SHAP scores and clustering per predictors for each of the 9 surprisal estimates.

References

[1] Willems, R. M., Frank, S., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral cortex*, *26 6*, 2506–2516.

[2] Michaelov, J. A., Bardolph, M. D., Petten, C. K. V., Bergen, B. K., & Coulson, S. (2024). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, *5*(1), 107–135.

[3] Slaats, S., & Martin, A. E. (2025). What's Surprising About Surprisal. *Computational Brain & Behavior*.

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners [Technical Report].

[5] Penedo, G., & et al. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and the Importance of Benchmarking.

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

[7] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*.

[8] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv*.

[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.