Benchmarking Sentence Processing Models:
Do Surprisal and Lossy-Context Surprisal Outperform Theory?

Michael Vrazitulis[1], Pia Schoknecht[1], Shravan Vasishth[1]
[1] University of Potsdam
vrazitulis@uni-potsdam.de

Background

GEPPU (**G**erman **E**valuation Benchmark for **P**sycholinguistics from **P**otsdam **U**niversity) is a benchmark dataset of German reading times from eye tracking and self-paced reading (SPR), based on multiple controlled experimental designs. It supports quantitative model evaluation across psycholinguistic phenomena. We use it to compare the predictive accuracy of surprisal, lossy-context surprisal, and theory-based qualitative predictions.

Method

We evaluated three different accounts of sentence processing difficulty:

1. **Qualitative predictions** based on psycholinguistic theory,[1] encoding each predicted main effect or interaction as a one-unit difference on the predictor variable;
2. **Surprisal** [1] from a German GPT-2 [2, 3] model;
3. **Lossy-context surprisal** [4] from German GPT-2, after probabilistically reconstructing distorted contexts with BERT [5] and Gibbs sampling.

Each account was tested as a predictor in a Bayesian log-normal hierarchical model of reading times in the critical sentence region. Random intercepts and slopes were included for subjects, items, and phenomena. Model comparison was carried out using Pareto smoothed importance sampling, which approximates leave-one-out cross-validation (PSIS-LOO).[6] This analysis was carried out separately on the SPR data and the eye-tracking data from GEPPU (on the subset of data collected by April 2025). For eye tracking, regression path durations served as the reading time measure of interest. Participants whose comprehension question accuracy was at chance level were excluded from the analysis.

Results

In the SPR data ($N = 615$), lossy-context surprisal yielded the best predictive accuracy, slightly outperforming standard surprisal ($\Delta\widehat{\text{elpd}}$ = 45.5, SE = 21.3) and vastly outperforming qualitative predictions ($\Delta\widehat{\text{elpd}}$ = 1089.5, SE = 56.2). Standard surprisal also clearly outperformed qualitative predictions ($\Delta\widehat{\text{elpd}}$ = 1044.0, SE = 53.5). However, no clear model advantage was observed in the eye-tracking data ($N = 118$); all models achieved similar predictive performance (lossy-context surprisal vs. surprisal: $\Delta\widehat{\text{elpd}}$ = 2.1, SE = 7.5; lossy-context surprisal vs. qualitative predictions: $\Delta\widehat{\text{elpd}}$ = 8.5, SE = 14.0; surprisal vs. qualitative predictions: $\Delta\widehat{\text{elpd}}$ = 6.4, SE = 11.6). See Figure 1 for a visual summary of the model comparison results.

Discussion

For SPR, both surprisal-based predictors clearly outperformed theory-driven qualitative predictions, supporting probabilistic accounts of sentence processing. Lossy-context surprisal showed a small advantage over standard surprisal, but further evidence is needed to establish the robustness of this difference. For eye tracking, no reliable model differences were found, possibly due to higher noise or the smaller sample size. These results underscore the importance of evaluating models across diverse reading measures.

[1] As preregistered and justified here (anonymized link): https://osf.io/wpra9?view_only=2945b83dddfe4731bd60d0103559d1b4

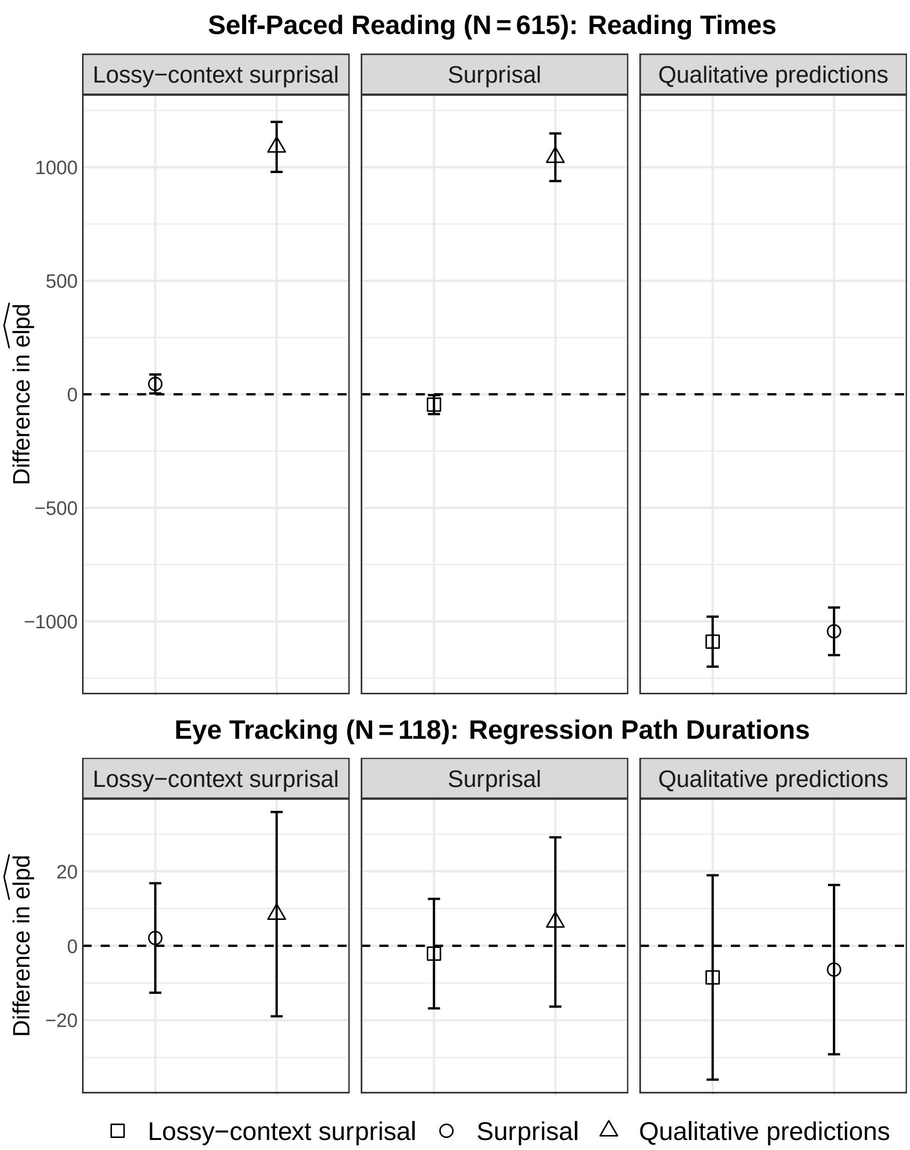

Figure 1: Model comparison based on expected log predictive density (elpd) differences from PSIS-LOO, for SPR (top) and eye tracking (bottom). Each panel shows how the model named in the title compares to the others. Positive values indicate better performance of the titled model. Error bars represent 95% CIs.

References

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

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

[3] Bayerische Staatsbibliothek. (2020). German GPT-2 Model [https://huggingface.co/dbmdz/german-gpt2].

[4] Futrell, R., Gibson, E., & Levy, R. P. (2021). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814.

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

[6] Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*, 1413–1432.