A Progress Report on Ongoing Benchmark Data Collection for German Sentence Processing: Eye-Tracking and Self-Paced Reading

Michael Vrazitulis, Pia Schoknecht, Shravan Vasishth
University of Potsdam
{vrazitulis, schoknecht, vasishth}@uni-potsdam.de

In sentence processing research, theories are often developed to explain specific phenomena, such as the subject–object relative clause asymmetry. However, these theories are rarely evaluated against a broader range of empirical findings. A standard benchmark dataset covering multiple phenomena for model evaluation and comparison is currently missing.

Huang et al. [4] took a first step by creating a large-scale self-paced reading benchmark on English syntactic ambiguities (e.g., garden paths). They used this dataset to evaluate predictions from the surprisal metric [1, 7]. Their results revealed important gaps in the explanatory power of surprisal. More broadly, benchmark datasets allow researchers to quantify model predictions and systematically identify where models succeed and where they fail.

There is a pressing need for benchmark datasets based on eye-tracking measures and for languages other than English. It is also important to extend benchmarking beyond just effects of syntactic disambiguation. To address this, we are creating a benchmark dataset based on a large-sample eye-tracking study in German. The study covers a range of postulated effects, including garden-path ambiguities, agreement attraction, local coherence, interference effects, attachment ambiguities, and the relative clause asymmetry (for details, see Table 1). Each experimental design comprises three to four conditions, with three items per condition arranged in a Latin square. Trials are randomized individually for each participant. Each trial is followed by a binary-choice comprehension question targeting the critical dependency of the sentence. Participants whose accuracy on comprehension questions falls below chance level are excluded from analysis.

A complementary study is being conducted using self-paced reading (SPR) on the same materials. The eye-tracking data are collected in the lab, while the SPR data are collected online via Prolific.[1] By collecting both eye-tracking and SPR data for the same materials, we aim to enable direct comparisons across methods and study the relationship between different reading measures.

Data collection is ongoing. For eye-tracking, the current sample size (as of April 25, 2025) is 119 participants (pre-exclusion). For SPR, the current sample size is 659 participants (pre-exclusion). For eye-tracking, we plan to continue until all main effects and interactions across the tested phenomena reach 95% credible intervals of ±50 ms or narrower, based on total fixation times. The effect estimates are derived from a log-normal hierarchical model and backtransformed to milliseconds. For SPR, we will continue until 1,100 participants have been collected, as preregistered at `https://osf.io/wpra9?view_only=2945b83dddfe4731bd60d0103559d1b4` (anonymized link).

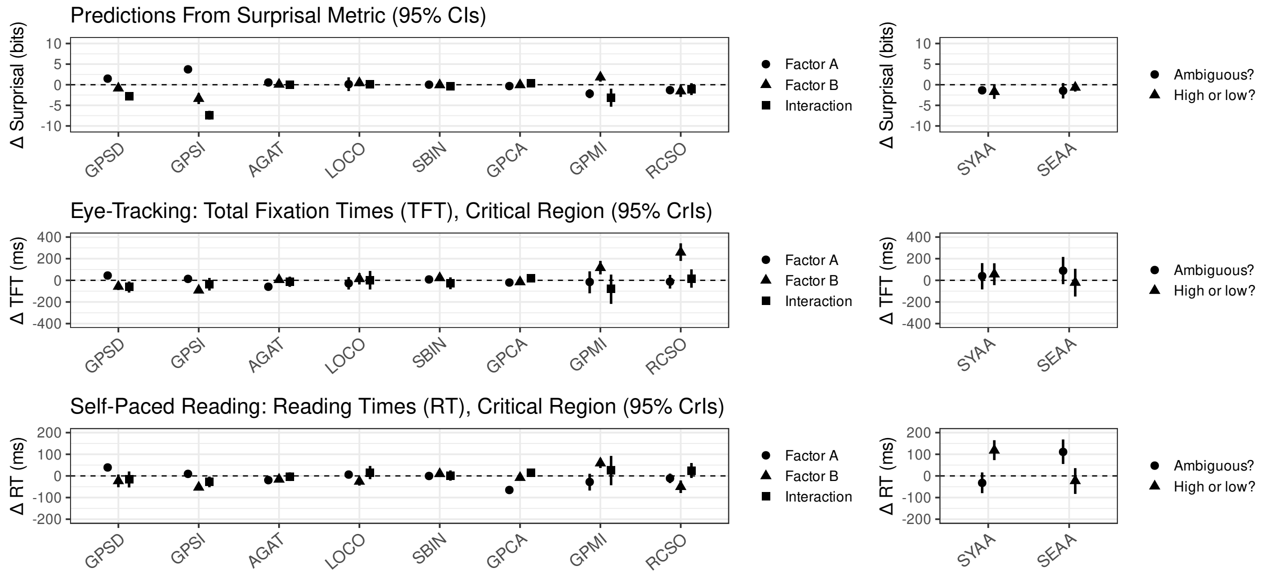
Preliminary analyses suggest that surprisal explains many of the observed effect patterns well, although some effects remain unexplained (see Figure 1). Word-by-word surprisal values were derived from a version of GPT-2 [12] pretrained on German corpora [14].

Once completed, the full benchmark dataset will be made publicly available to support quantitative model evaluation and evidence-based theory development in sentence processing.

[1] https://www.prolific.com

Table 1: Sentence processing phenomena and corresponding experimental designs.

GPSD (2×2): Garden Paths From Subject-vs.-Direct-Object Ambiguity Ambiguous/Unambiguous × S–O/O–S — closely replicating [9]
GPSI (2×2): Garden Paths From Subject-vs.-Indirect-Object Ambiguity Ambiguous/Unambiguous × Active/Passive — loosely replicating [10]
AGAT (2×2): Agreement Attraction in Grammatical Sentences Singular-/Plural-Controller × Match/Mismatch — closely replicating [2]
LOCO (2×2): Local Coherence Coherent/Incoherent × Intervener/No-Intervener — closely replicating [11]
SBIN (2×2): Similarity-Based Interference Subject-Cue [Yes/No] × Animacy-Cue [Yes/No] — closely replicating [13]
GPCA (2×2): Garden Paths From Coordination Ambiguity NP-/VP-Coordination × AP-/PP-Modifier — closely replicating [6]
GPMI (2×2): Garden Paths From Modifier-vs.-Indirect-Object Ambiguity Modifier/No-Modifier × Ambiguous/Unambiguous — closely replicating [5]
RCSO (2×2): Subject vs. Object Relative Clauses Subject/Object × Double-/Single-Embedding — German adaptation of [3]
SYAA (3×1): Syntax-Based Attachment Ambiguity High-/Low-/Ambiguous-Attachment — closely replicating [8]
SEAA (3×1): Semantics-Based Attachment Ambiguity High-/Low-/Ambiguous-Attachment — German adaptation of [15]



Predictions From Surprisal Metric (95% CIs)

Eye-Tracking: Total Fixation Times (TFT), Critical Region (95% CrIs)

Self-Paced Reading: Reading Times (RT), Critical Region (95% CrIs)

Figure 1: Predictions and observed effects across 2×2 and 3×1 designs. Rows show surprisal-based predictions, eye-tracking results (total fixation times), and self-paced reading results.

References

[1] J. T. Hale. In: *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Pittsburgh, PA, 2001. [2] J. Häussler. PhD thesis. University of Konstanz, 2009. [3] F. Hsiao and E. Gibson. In: *Cognition* 90.1 (2003), pp. 3–27. [4] K.-J. Huang et al. In: *Journal of Memory and Language* 137 (2024), p. 104510. [5] A. van Kampen. PhD thesis. Free University of Berlin, 2001. [6] L. Konieczny, B. Hemforth, and C. Scheepers. In: *German Sentence Processing*. Springer, 2000, pp. 247–278. [7] R. Levy. In: *Cognition* 106.3 (2008), pp. 1126–1177. [8] P. Logačev. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 49.9 (2023), p. 1471. [9] M. Meng and M. Bader. In: *Language and Speech* 43.1 (2000), pp. 43–74. [10] M. Meng and M. Bader. In: *Language and Cognitive Processes* 15.6 (2000), pp. 615–666. [11] D. Paape and S. Vasishth. In: *Language and Speech* 59.3 (2016), pp. 387–403. [12] A. Radford et al. In: *OpenAI Blog* 1.8 (2019), p. 9. [13] P. Schoknecht, H. Yadav, and S. Vasishth. In: *Journal of Memory and Language* 141 (2025), p. 104599. [14] B. Staatsbibliothek. https://huggingface.co/dbmdz/german-gpt2. 2020. [15] M. J. Traxler, M. J. Pickering, and C. Clifton Jr. In: *Journal of Memory and Language* 39.4 (1998), pp. 558–592.