A German Eye-Movement Benchmark Dataset

Michael Vrazitulis, Pia Schoknecht, Shravan Vasishth

SFB 1287 Limits of Variability in Language
Cognitive, Computational, and Grammatical Aspects

Introduction

- Benchmark data are an important tool for developing theories and evaluating model predictions.
- The majority of benchmark data in sentence processing are limited to English (e.g., [2, 7]).

Our Work (in Progress)

- We collect eye-tracking benchmark data for a battery of postulated effects in German (10 phenomena).
- In parallel, we also collect self-paced reading (SPR) data on the same materials.
- So far (April 25, 2025):
  - 659 Prolific participants have been tested with SPR. 44 were excluded due to low accuracy on comprehension questions.
  - 119 in-lab participants have been tested with eye tracking. 1 was excluded due to low accuracy on comprehension questions.
- We show the results so far, next to predictions based on qualitative theories, GPT-2 surprisal [3, 10, 15, 17], and lossy-context GPT-2 surprisal [1, 2, 5, 15].
- Model comparisons using Pareto-smoothed importance sampling [19] assess out-of-sample predictiveness.

Pre-Registration Protocol (SPR)

osf.io/wpra9

Predictions, Results, and Model Comparison

Experimental Designs

GPSD (2×2): Garden Paths From Subject-vs.-Direct-Object Ambiguity
Ambiguous/Unambiguous × S–O/O–S — closely replicating [12]

GPSI (2×2): Garden Paths From Subject-vs.-Indirect-Object Ambiguity
Ambiguous/Unambiguous × Active/Passive — loosely replicating [13]

GPCA (2×2): Garden Paths From Coordination Ambiguity
NP-/VP-Coordination × AP-/PP-Modifier — closely replicating [9]

GPMI (2×2): Garden Paths From Modifier-vs.-Indirect-Object Ambiguity
Modifier/No-Modifier × Ambiguous/Unambiguous — closely replicating [8]

AGAT (2×2): Agreement Attraction in Grammatical Sentences
Singular-/Plural-Controller × Match/Mismatch — closely replicating [4]

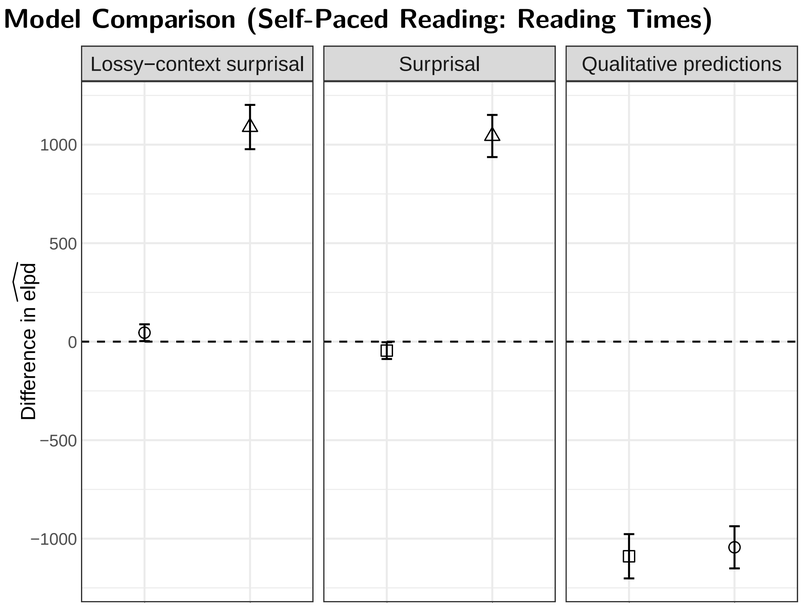
LOCO (2×2): Local Coherence
Coherent/Incoherent × Intervener/No-Intervener — closely replicating [14]

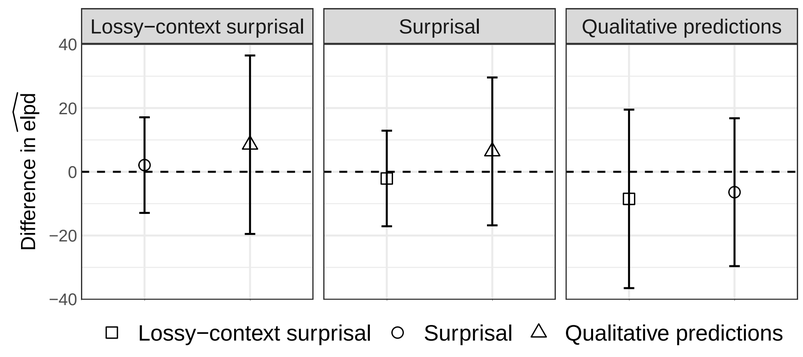
SBIN (2×2): Similarity-Based Interference
Subject-Cue [Yes/No] × Animacy-Cue [Yes/No] — closely replicating [16]

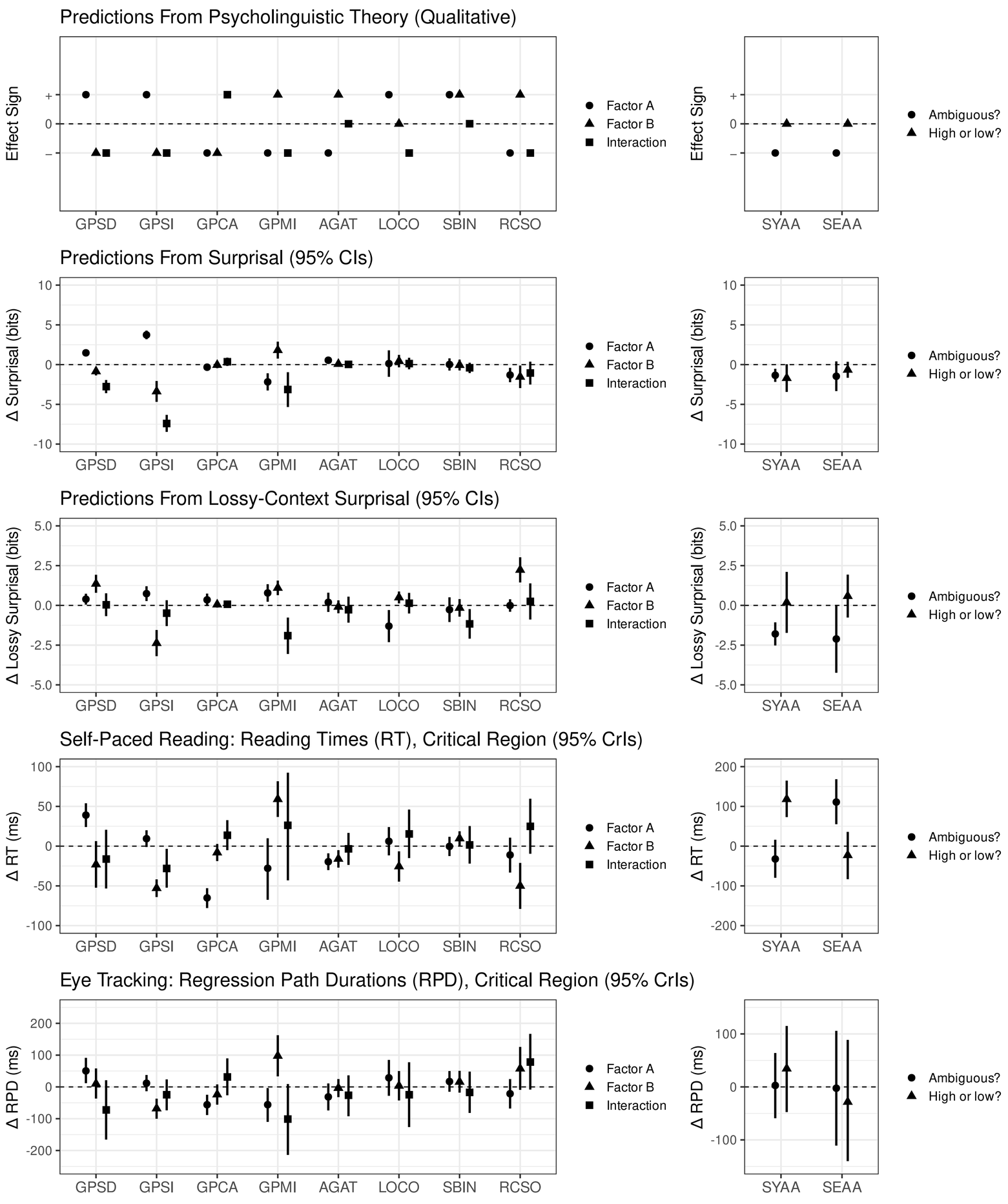
RCSO (2×2): Subject vs. Object Relative Clauses
Subject/Object × Double-/Single-Embedding — German adaptation of [6]

SYAA (3×1): Syntax-Based Attachment Ambiguity
High-/Low-/Ambiguous-Attachment — closely replicating [11]

SEAA (3×1): Semantics-Based Attachment Ambiguity
High-/Low-/Ambiguous-Attachment — German adaptation of [18]







References

[1] J. Devlin et al. In: *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (long and short papers)*. 2019, pp. 4171–4186. [2] R. Futrell, E. Gibson, and R. P. Levy. In: *Cognitive Science* 44.3 (2021), e12814. [3] J. T. Hale. In: *Proceedings of the Second Meeting of the NAACL*. Pittsburgh, PA, 2001. [4] J. Häussler. PhD thesis. University of Konstanz, 2009. [5] J. Hennert et al. Unpublished manuscript. 2025. [6] F. Hsiao and E. Gibson. In: *Cognition* 90.1 (2003), pp. 3–27. [7] K.-J. Huang et al. In: *Journal of Memory and Language* 137 (2024), p. 104510. [8] A. van Kampen. PhD thesis. Free University of Berlin, 2001. [9] L. Konieczny, B. Hemforth, and C. Scheepers. In: *German Sentence Processing*. Ed. by B. Hemforth and L. Konieczny. Springer, 2000, pp. 247–278. [10] R. Levy. In: *Cognition* 106.3 (2008), pp. 1126–1177. [11] P. Logačev. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 49.9 (2023), p. 1471. [12] M. Meng and M. Bader. In: *Language and Speech* 43.1 (2000), pp. 43–74. [13] M. Meng and M. Bader. In: *Language and Cognitive Processes* 15.6 (2000), pp. 615–666. [14] D. Paape and S. Vasishth. In: *Language and Speech* 59.3 (2016), pp. 387–403. [15] A. Radford et al. In: *OpenAI Blog* 1.8 (2019), p. 9. [16] P. Schoknecht, H. Yadav, and S. Vasishth. In: *Journal of Memory and Language* 141 (2025), p. 104599. [17] B. Staatsbibliothek. https://huggingface.co/dbmdz/german-gpt2. 2020. [18] M. J. Traxler, M. J. Pickering, and C. Clifton Jr. In: *Journal of Memory and Language* 39.4 (1998), pp. 558–592. [19] A. Vehtari, A. Gelman, and J. Gabry. In: *Statistics and Computing* 27 (2017), pp. 1413–1432.