A Benchmark Dataset for Evaluating Sentence Processing Models in German

Sentence Processing Workshop in Golm

May 27, 2025

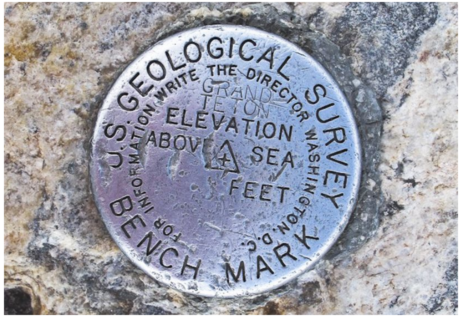
Michael Vrazitulis

1 Introduction

Benchmark data are essential for theory building and model testing.

In sentence processing, there is an **urgent need** for datasets covering multiple experimental designs in a single large participant sample.

One such dataset was created by Huang et al. (2024), covering several syntactic ambiguity phenomena in English (SAP Benchmark).



1 Introduction

Here, I present **ongoing work** on the creation of a **new benchmark dataset** for sentence processing in **German.**

It includes both **self-paced reading** and **eye-tracking** data.

2 The German Benchmark Dataset

osf.io/wpra9

GPSD (2×2): Garden Paths From Subject-vs.-Direct-Object Ambiguity Ambiguous/Unambiguous × S–O/O–S — closely replicating Meng & Bader (2000a)
GPSI (2×2): Garden Paths From Subject-vs.-Indirect-Object Ambiguity Ambiguous/Unambiguous × Active/Passive — loosely replicating Meng & Bader (2000b)
GPCA (2×2): Garden Paths From Coordination Ambiguity NP-/VP-Coordination × AP-/PP-Modifier — closely replicating Konieczny et al. (2000)
GPMI (2×2): Garden Paths From Modifier-vs.-Indirect-Object Ambiguity Modifier/No-Modifier × Ambiguous/Unambiguous — closely replicating van Kampen (2001)

2 The German Benchmark Dataset

osf.io/wpra9

AGAT (2×2): Agreement Attraction in Grammatical Sentences
Singular-/Plural-Controller × Match/Mismatch — closely replicating Häussler (2009)

LOCO (2×2): Local Coherence
Coherent/Incoherent × Intervener/No-Intervener — closely replicating Paape & Vasishth (2016)

SBIN (2×2): Similarity-Based Interference
Subject-Cue [Yes/No] × Animacy-Cue [Yes/No] — closely replicating Schoknecht et al. (2025)

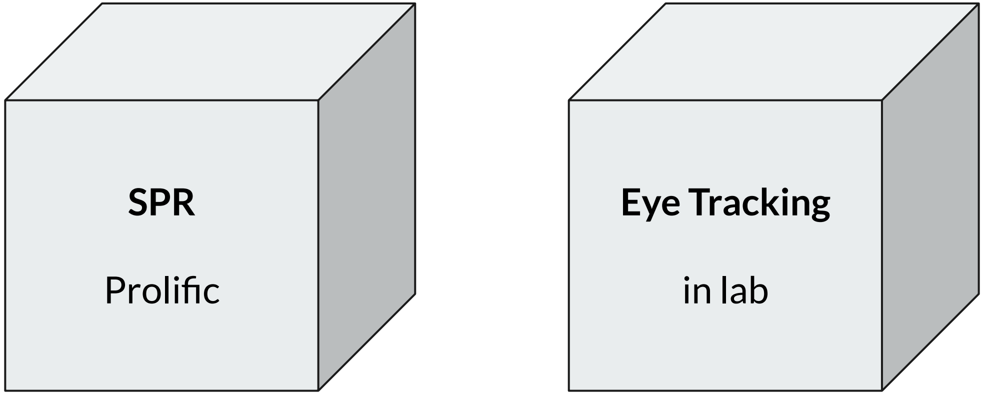
RCSO (2×2): Subject vs. Object Relative Clauses
Subject/Object × Double-/Single-Embedding — German adaptation of Hsiao & Gibson (2003)

2 The German Benchmark Dataset

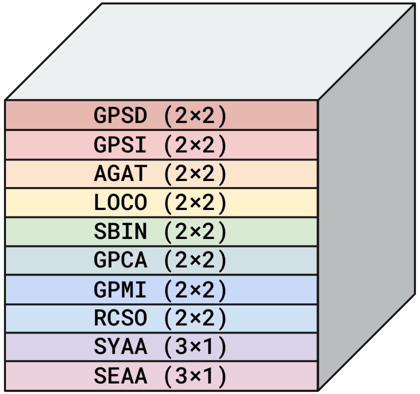
osf.io/wpra9

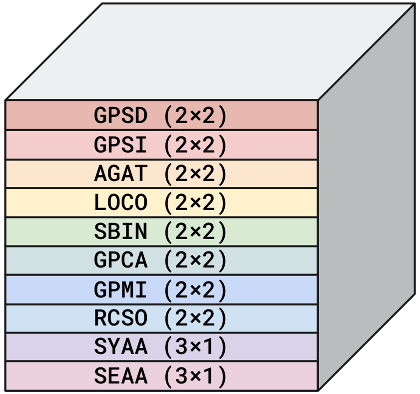
SYAA (3×1): Syntax-Based Attachment Ambiguity	
High-/Low-/Ambiguous-Attachment — closely replicating Logačev (2023)	
SEAA (3×1): Semantics-Based Attachment Ambiguity	
High-/Low-/Ambiguous-Attachment — German adaptation of Traxler et al. (2023)	

2 The German Benchmark Dataset

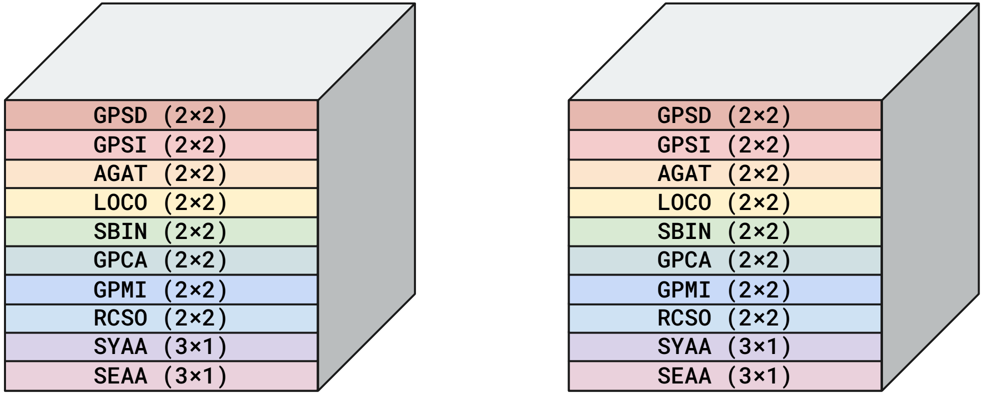


2 The German Benchmark Dataset





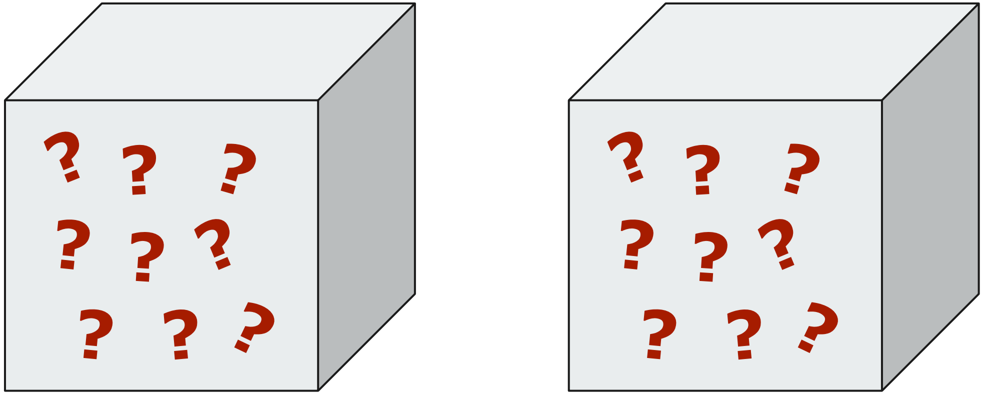
2 The German Benchmark Dataset



Three trials per condition per participant in a Latin square design

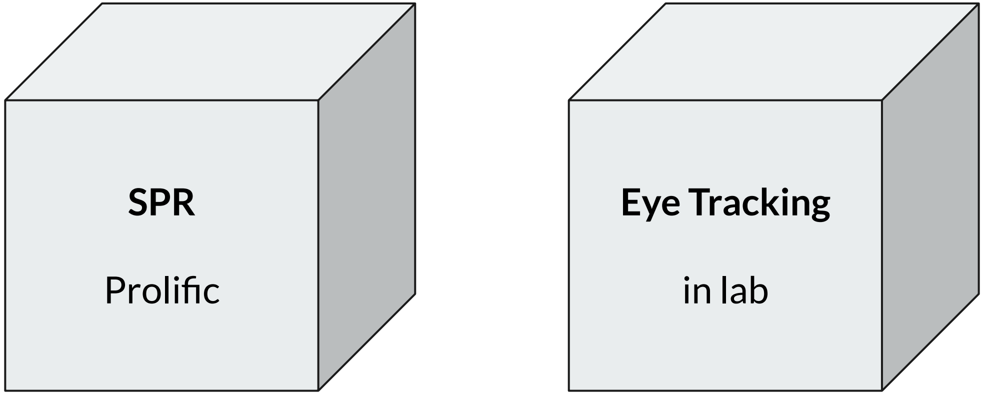
Overall trials per participant: **114** [= 8×3×(2×2) + 2×3×(3×1)]

2 The German Benchmark Dataset



Every trial is followed by a **comprehension question** with two response options, targeting the key syntactic dependency.

3 Data Collection Status and Goal



Already collected as of April 25, 2025:

N = 659

N = 119

3 Data Collection Status and Goal

SPR

Prolific

Eye Tracking

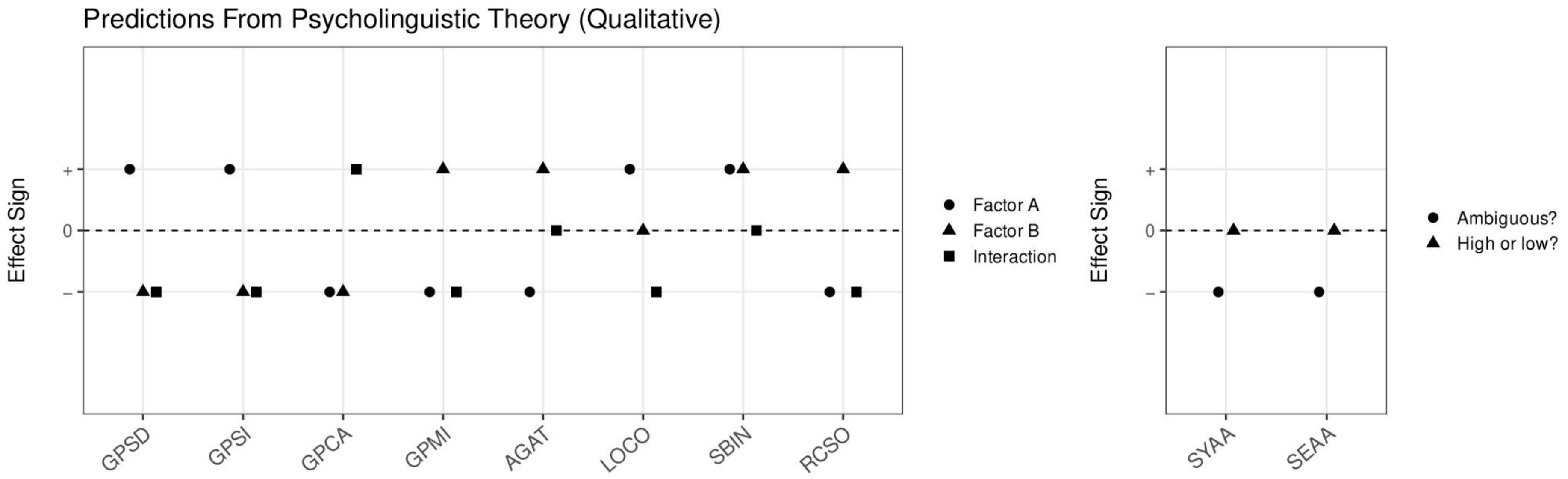
in lab

Goal:

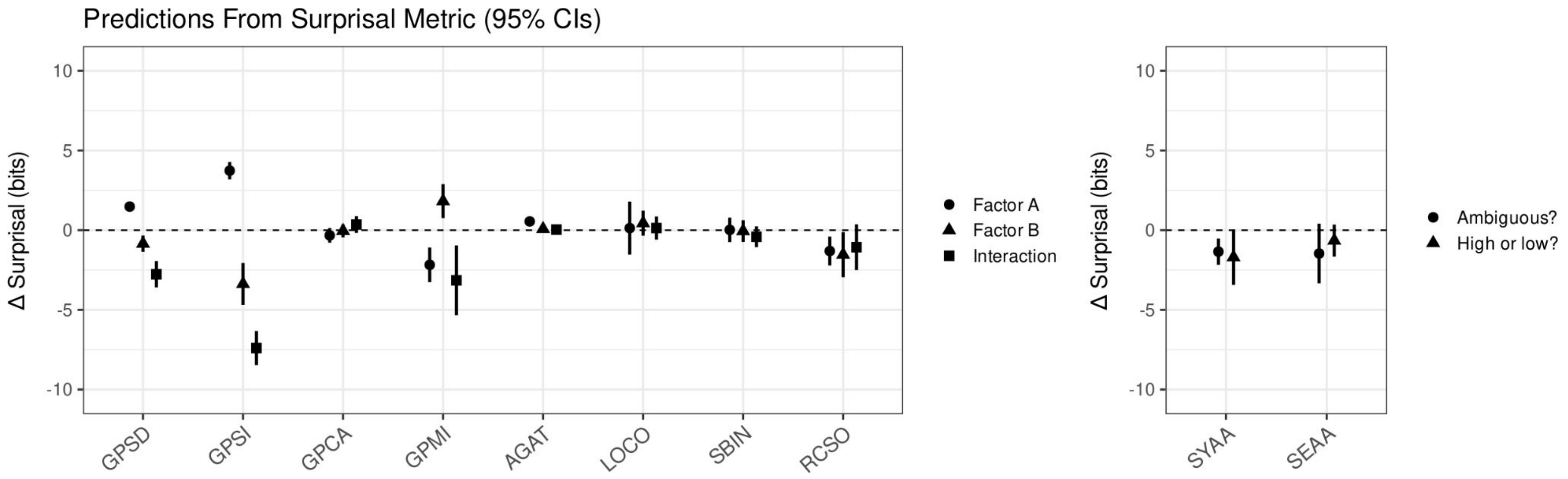
***N* = 1,100**

Stop when *all* effect CrIs on TFTs range ±50ms or less.

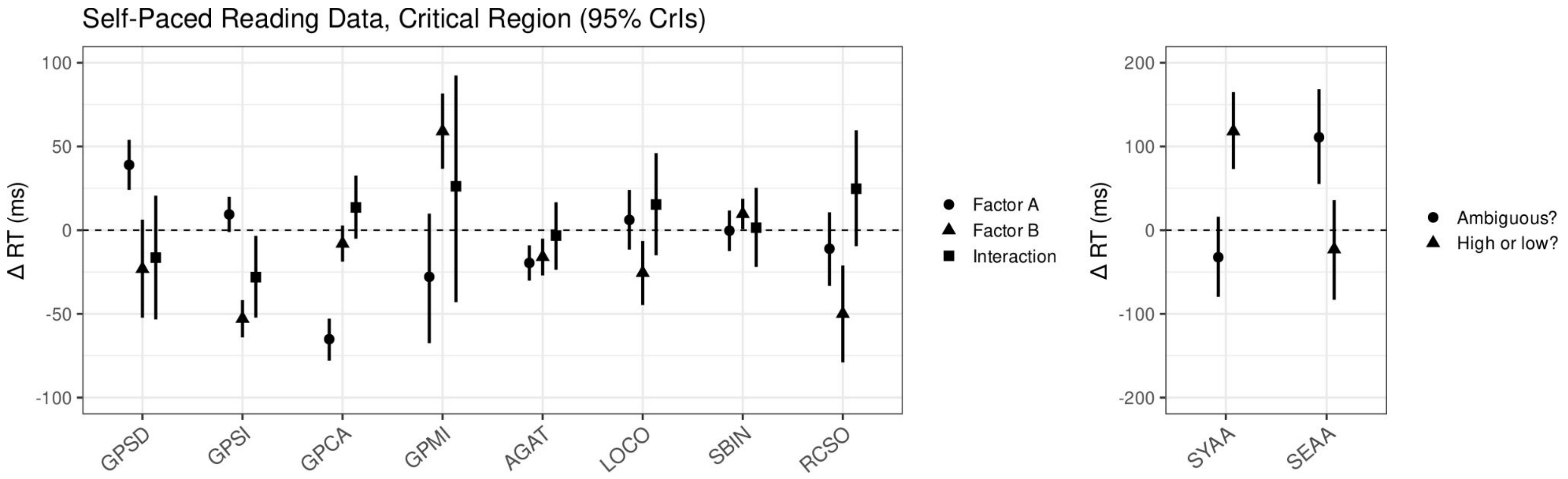
4 Predictions and Observed Effects



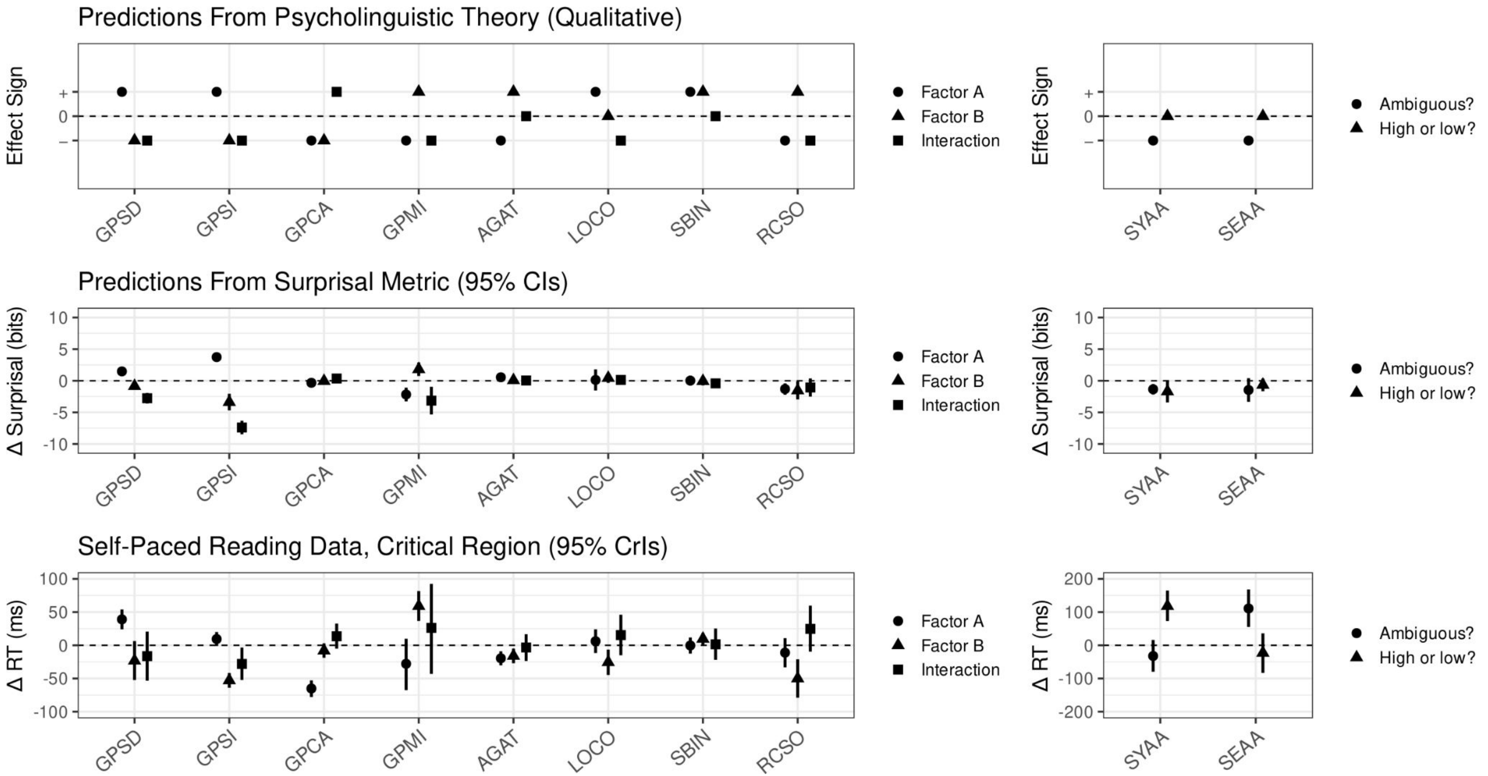
4 Predictions and Observed Effects



4 Predictions and Observed Effects

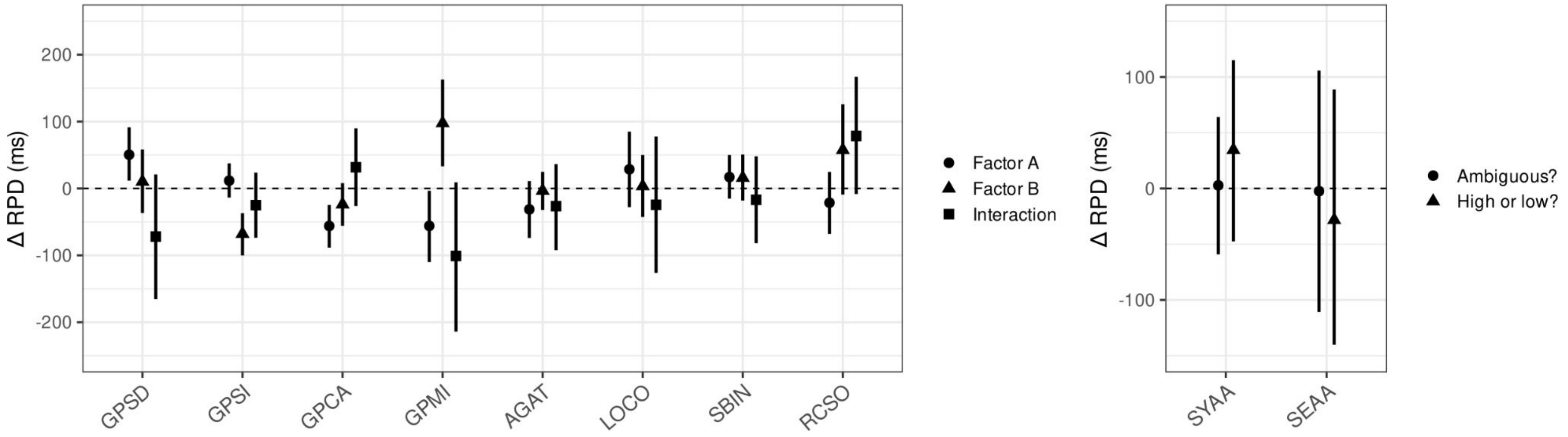


4 Predictions and Observed Effects

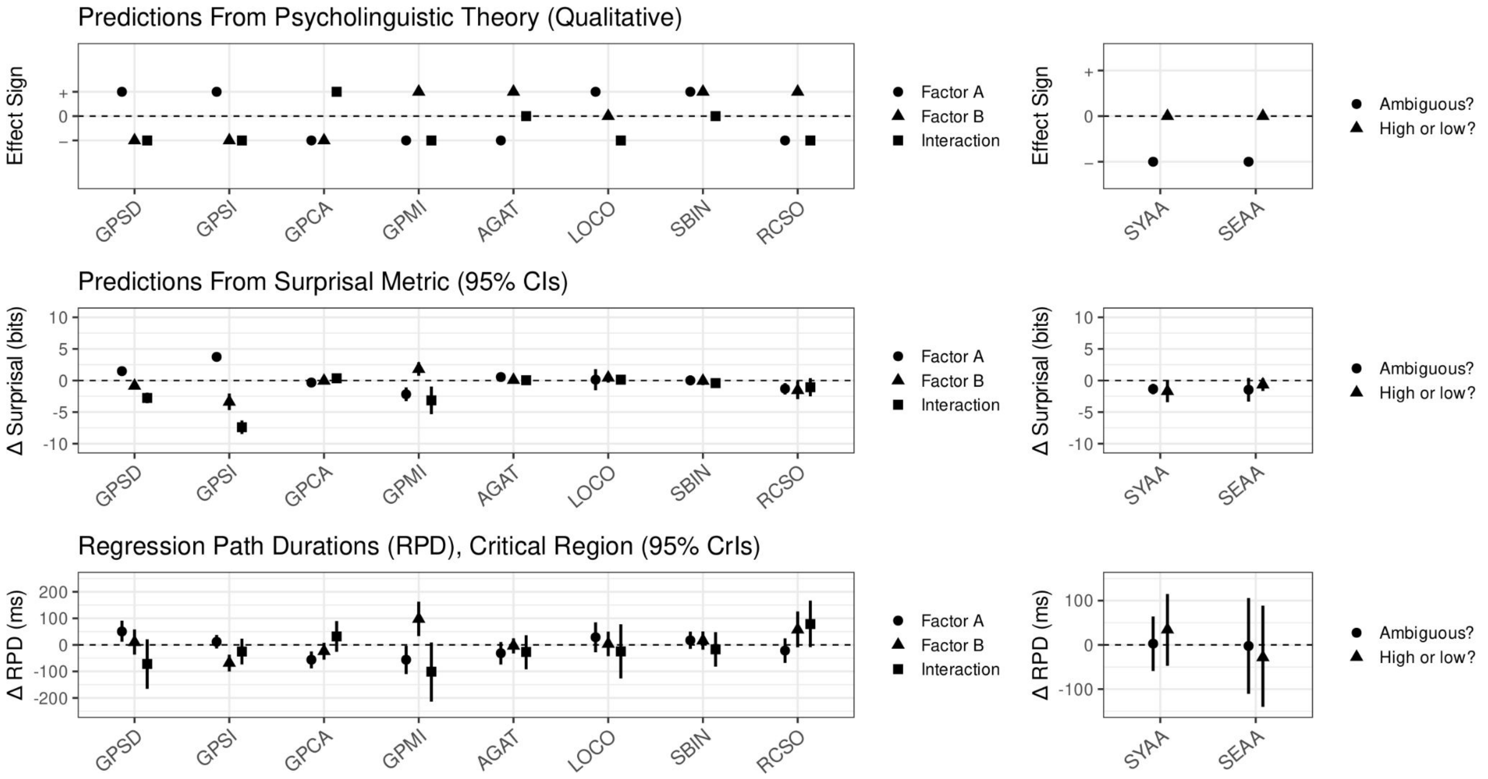


4 Predictions and Observed Effects

Regression Path Durations (RPD), Critical Region (95% CrIs)



4 Predictions and Observed Effects



5 Model Evaluation and Comparison

Models with single linear predictor …

1. **Qualitative predictions** based on psycholinguistic theory (preregistered at osf.io/wpra9), encoding each predicted main effect / interaction as a 1-unit difference on the predictor

2. **Surprisal** (Hale 2001; Levy 2008) from GPT-2 (Radford et al. 2019)

3. **Lossy-context surprisal** (Futrell et al. 2021) from GPT-2, after probabilistic reconstruction of distorted context with BERT (Devlin et al. 2019) and Gibbs sampling

Evaluation method: Pareto-smoothed importance sampling (PSIS-LOO)

5 Model Evaluation and Comparison

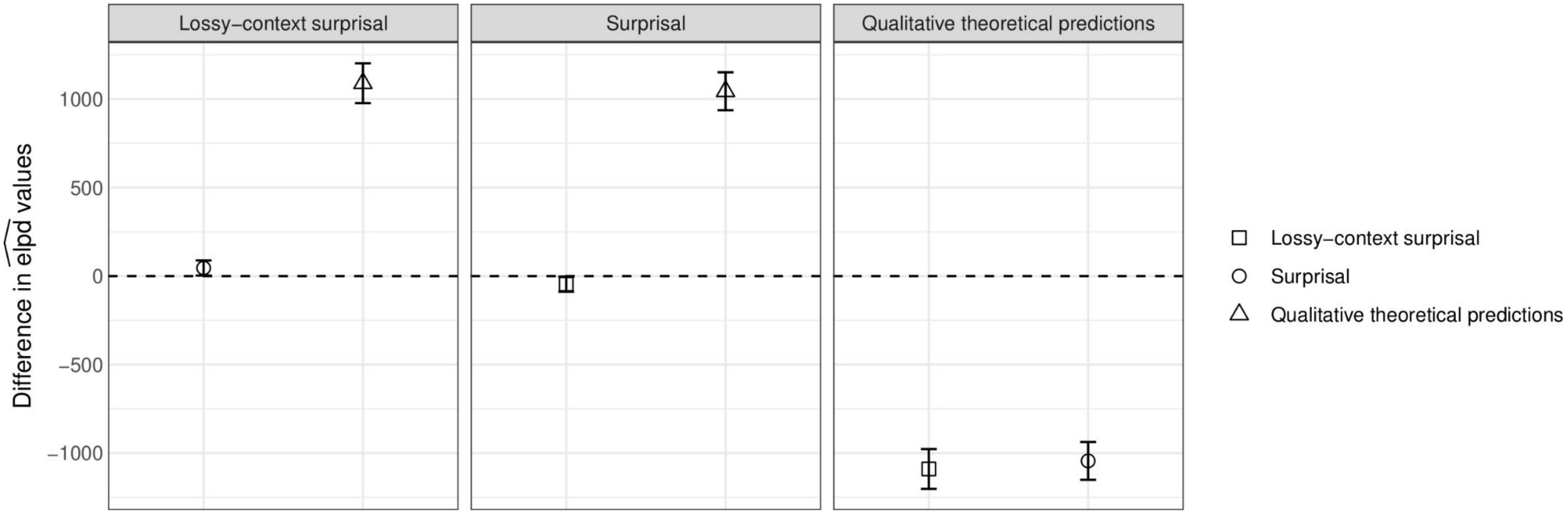
Self-paced reading ($N_{included}$ = 615), reading times (RT) on critical region:

```
log(critical_RT) ~ predictor + (predictor | subject) + (predictor | phenomenon) + (predictor | phenomenon:item)
```

5 Model Evaluation and Comparison

Self-paced reading ($N_{included}$ = 615), reading times (RT) on critical region:

```
log(critical_RT) ~ predictor + (predictor | subject) + (predictor | phenomenon) + (predictor | phenomenon:item)
```



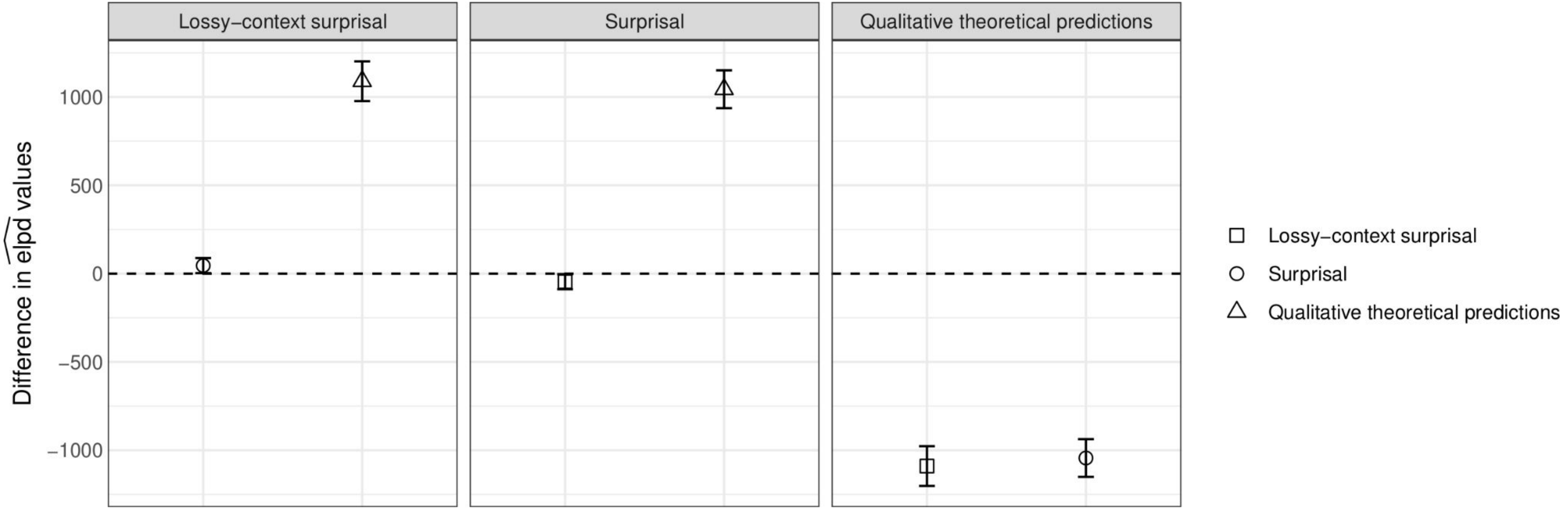
5 Model Evaluation and Comparison

Small, but significant edge of **lossy surprisal over plain surprisal.**

Huge edge of **both surprisal variants over classical qualitative theory.**

Self-paced reading ($N_{included}$ = 615), reading times (RT) on critical region:

```
log(critical_RT) ~ predictor + (predictor | subject) + (predictor | phenomenon) + (predictor | phenomenon:item)
```



5 Model Evaluation and Comparison

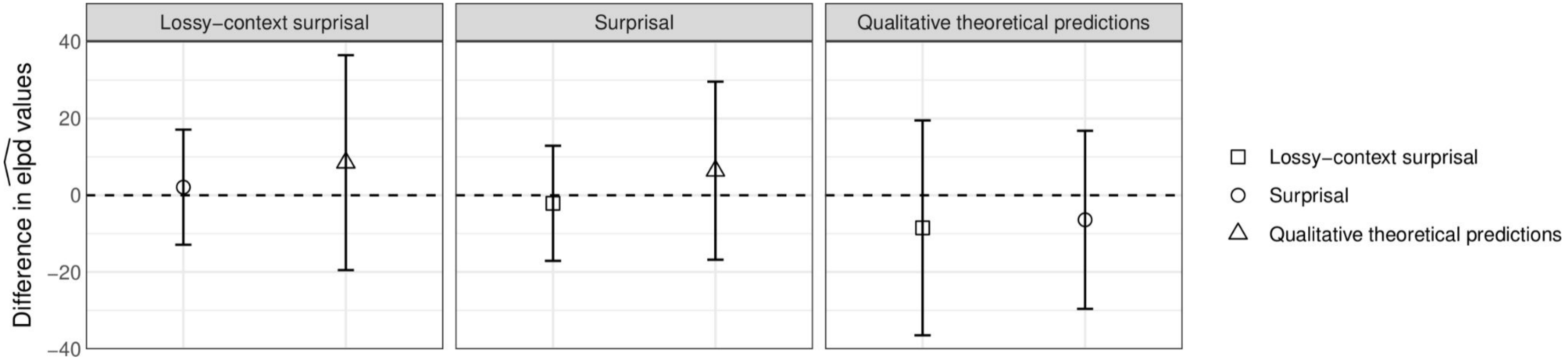
Eye tracking (N_{included} = 118), regression path durations (RPD) on critical region:

```
log(critical_RPD) ~ predictor + (predictor | subject) + (predictor | phenomenon) + (predictor | phenomenon:item)
```

5 Model Evaluation and Comparison

Eye tracking ($N_{included}$ = 118), regression path durations (RPD) on critical region:

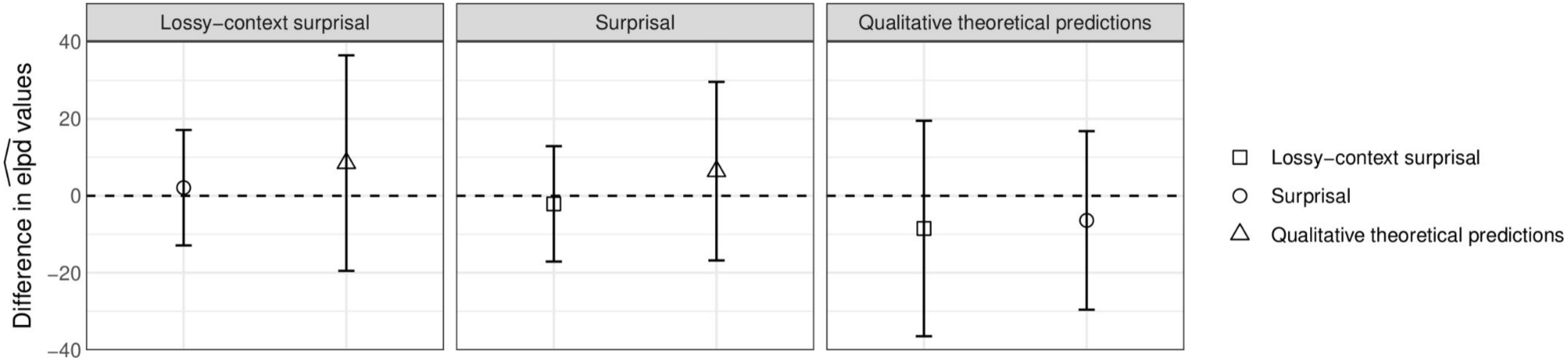
```
log(critical_RPD) ~ predictor + (predictor | subject) + (predictor | phenomenon) + (predictor | phenomenon:item)
```



5 Model Evaluation and Comparison

Eye tracking ($N_{included}$ = 118), regression path durations (RPD) on critical region:

```
log(critical_RPD) ~ predictor + (predictor | subject) + (predictor | phenomenon) + (predictor | phenomenon:item)
```

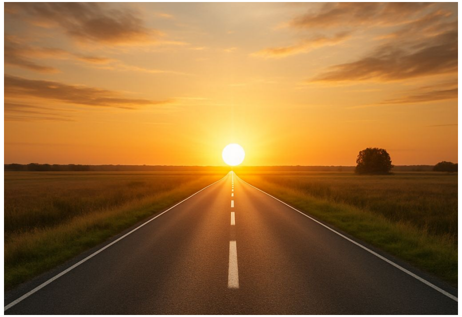


There is no evidence yet favoring one model over the others.

6 Future Directions

The **final dataset** will be **openly available.**

More **model evaluations** are planned, e.g, on resource-rational lossy surprisal (Hahn et al., 2022) or cognitive process models.



Thanks for your attention!

References [1/2]

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

- Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814.

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

- Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics.*

- Häussler, J. (2009). *The Emergence of Attraction Errors During Sentence Comprehension*. PhD thesis, University of Konstanz.

- Hsiao, F. and Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*(1), 3–27.

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

- Konieczny, L., Hemforth, B., and Scheepers, C. (2000). Head position and clause boundary effects in reanalysis. In Hemforth, B. and Konieczny, L., editors, *German Sentence Processing*, pages 247–278. Springer.

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

References [2/2]

- Logačev, P. (2023). The role of underspecification in relative clause attachment: Speed-accuracy tradeoff evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(9), 1471.
- Meng, M. and Bader, M. (2000a). Mode of disambiguation and garden-path strength: An investigation of subject–object ambiguities in German. *Language and Speech*, *43*(1), 43–74.
- Meng, M. and Bader, M. (2000b). Ungrammaticality detection and garden path strength: Evidence for serial parsing. *Language and Cognitive Processes*, *15*(6), 615–666.
- Paape, D. and Vasishth, S. (2016). Local coherence and preemptive digging-in effects in German. *Language and Speech*, *59*(3), 387–403.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.
- Schoknecht, P., Yadav, H., & Vasishth, S. (2025). Do syntactic and semantic similarity lead to interference effects? Evidence from self-paced reading and event-related potentials using German. *Journal of Memory and Language*, *141*, 104599.
- Traxler, M. J., Pickering, M. J., and Clifton Jr, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592.
- van Kampen, A (2001). *Syntaktische und semantische Verarbeitungsprozesse bei der Analyse strukturell mehrdeutiger Verbfinalsätze im Deutschen: Eine empirische Untersuchung*. PhD thesis, Free University of Berlin.