

## 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 naturalistic reading (e.g., [1, 2, 3]).
- ▶ But benchmarks based on controlled stimuli (e.g., [7, 17]) are also necessary for robust model evaluation.

### Our Work (in Progress)

- ▶ We collect self-paced reading benchmark data for a battery of postulated effects in German (10 phenomena).
- ▶ In parallel, we also collect eye-tracking data on the same materials.
- ▶ So far (December 2025):
  - ▶ 950 Prolific participants have been tested with SPR. 63 were excluded due to low accuracy on comprehension questions.
  - ▶ 195 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, compared to qualitative and surprisal-based [4, 10, 15] predictions.

Pre-Registration Protocol (SPR)



osf.io/wpra9

## Phenomena, Demographics, Predictions, and Estimates

### 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 [5]

**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]

Method	Level	Measure	Mean ± 95% CI
Eye Tracking	Word	Single Fixation <sup>1</sup>	238.1 ± 4.3
		First Fixation <sup>1</sup>	233.9 ± 3.8
		Gaze Duration <sup>1</sup>	306.0 ± 6.8
		Total Fixation <sup>1</sup>	558.6 ± 20.4
		Number of Fixations <sup>2</sup>	2.2 ± 0.1
		Skip Rate <sup>3</sup>	0.24 ± 0.01
		Regression Rate <sup>3</sup>	0.19 ± 0.01
SPR	Segment	Reading Time <sup>1</sup>	689.5 ± 14.3

<sup>1</sup>In milliseconds. <sup>2</sup>Average number of fixations per word.

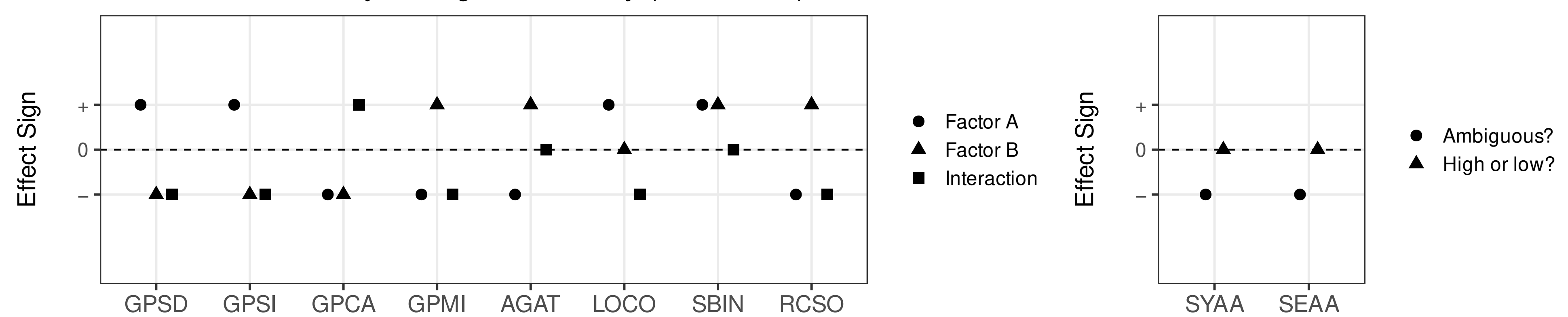
<sup>3</sup>Proportion of words.

### Highlights

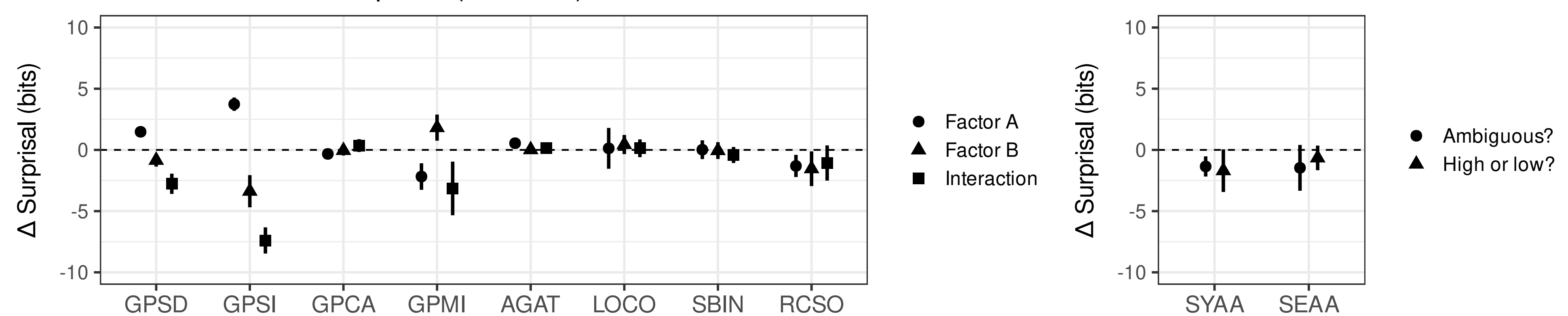
- ▶ **GEPPU (German Evaluation Benchmark for Psycholinguistics from Potsdam University)**
- ▶ A novel dataset with parallel SPR and eye-tracking data, based on controlled experimental designs
- ▶ Once published as a paper, the dataset will be made openly available!

Method	L1	N	Gender			Age (SD)	Comprehension Accuracy (%)	Trials per Participant
			Female	Male	Other			
Eye Tracking	German	195	147	46	2	23.3 (4.5)	82.6 %	114
SPR	German	950	425	524	1	30.9 (8.9)	76.3 %	114

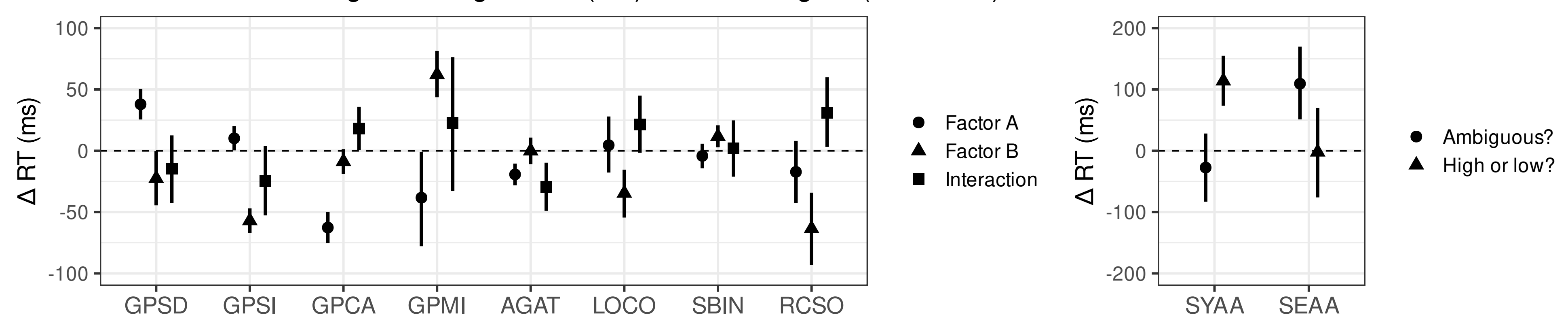
### Predictions From Psycholinguistic Theory (Qualitative)



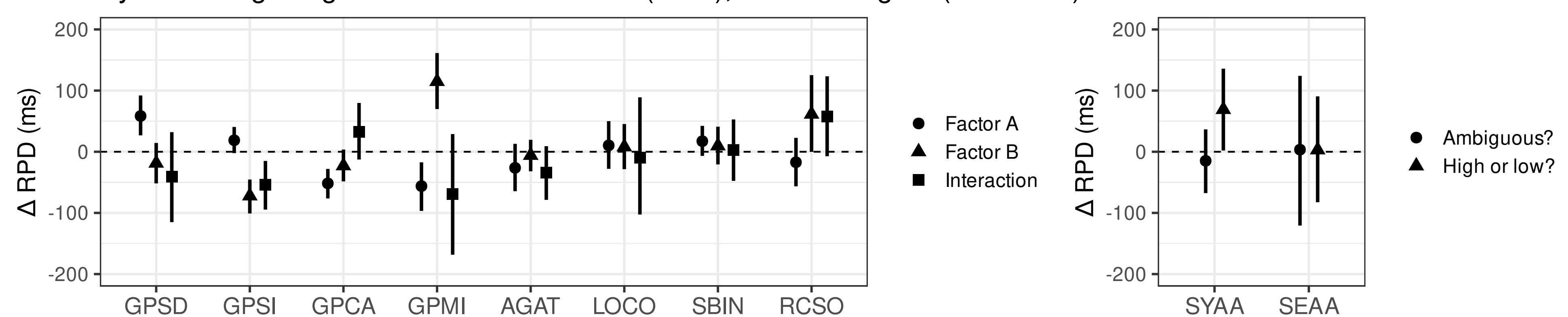
### Predictions From Surprisal (95% CIs)



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



### Eye Tracking: Regression Path Durations (RPD), Critical Region (95% CIs)



## References

- [1] Y. Berzak et al. In: *Open Mind* 6 (2022), pp. 41–50. [2] J. Chromý, M. Ceháková, and J. Brand. In: *Behavior Research Methods* 57.12 (2025), p. 345. [3] R. Futrell et al. In: *Language Resources and Evaluation* 55 (2021), pp. 63–77. [4] J. T. Hale. In: *Proceedings of the Second Meeting of the NAACL*. Pittsburgh, PA, 2001. [5] J. Häussler. PhD thesis. University of Konstanz, 2009. [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] W. Timkey et al. Unpublished manuscript. 2025. [18] M. J. Traxler, M. J. Pickering, and C. Clifton Jr. In: *Journal of Memory and Language* 39.4 (1998), pp. 558–592.