GEPPU: An Eye-Tracking and Self-Paced Reading Benchmark Dataset for German Sentence Processing

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

Sentence processing models are typically evaluated on narrow empirical domains, limiting claims about generalizability. Broad benchmarks based on controlled experimental designs are needed for systematic, quantitative model evaluation. Large-scale naturalistic reading datasets (e.g., [1, 2]) have recently become available and capture real-world variability, but they cannot isolate the contrasts needed to test specific theoretical predictions. Huang et al. [3] and Timkey et al. [4] illustrated the utility of controlled designs by constructing a large-scale English benchmark covering diverse syntactic phenomena (e.g., garden paths, relative clauses, agreement attraction), revealing clear limitations of surprisal-based [5] accounts. This motivates similarly broad controlled benchmarks in other languages and for a wider range of theoretically relevant contrasts.

GEPPU (**G**erman **E**valuation Benchmark for **P**sycholinguistics from **P**otsdam **U**niversity) is a novel controlled benchmark for German sentence processing during reading. It includes eye-tracking and self-paced reading (SPR) data from two experiments using identical stimuli; data collection is ongoing. The eye-tracking study is being run in-lab with an EyeLink 1000 Plus (1,000 Hz, right eye, N = 195 so far). The SPR study is being run online via Prolific (N = 950 so far).

Eye-tracking data collection will continue until the 95% credible interval for each factor's effect falls within ±50 ms on total fixation time. SPR data will be collected up to 1,100 participants.[1]

Each participant read 114 controlled sentences across ten experimental designs probing distinct psycholinguistic phenomena: four garden-path designs, two attachment-ambiguity designs, local coherence, similarity-based interference, agreement attraction, and relative clause asymmetries.[6, 7, 8, 9, 10, 11] Sentences were followed by binary-choice comprehension questions. Items were arranged in a Latin-square design with three trials per participant per condition, and trial order was randomized for each participant.

Table 1 summarizes participant demographics, comprehension accuracy, and trial counts for the eye-tracking and SPR datasets collected so far. Table 2 reports summary statistics for the key eye-tracking and SPR measures.

The GEPPU dataset provides a resource for evaluating computational models of sentence processing across a range of controlled experimental manipulations. It provides both eye-tracking and self-paced reading data on identical linguistic materials. This dataset will facilitate systematic model testing, allowing researchers to assess the generalizability of computationally implemented models across reading measures. It will thereby support cumulative theory development in psycholinguistics and NLP. The dataset will be made publicly available once the paper introducing the data is published.

[1] As preregistered here: https://osf.io/wpra9?view_only=2945b83dddfe4731bd60d0103559d1b4

Table 1: Participant demographics, comprehension accuracy, and trial count per participant in the GEPPU dataset, split by data collection method (eye tracking or self-paced reading).

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

Table 2: Mean values and 95% between-subject confidence intervals for eye-tracking and self-paced reading measures in the GEPPU dataset.

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

[1]In milliseconds. [2]Average number of fixations per word. [3]Proportion of words.

References

[1] Y. Berzak et al. “CELER: A 365-participant corpus of eye movements in L1 and L2 English reading”. In: *Open Mind* 6 (2022), pp. 41–50. [2] J. Chromỳ, M. Cehánková, and J. Brand. “The HeCz corpus: A large, richly annotated reading corpus of newspaper headlines in Czech”. In: *Behavior Research Methods* 57.12 (2025), p. 345. [3] K.-J. Huang et al. “Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty”. In: *Journal of Memory and Language* 137 (2024), p. 104510. [4] W. Timkey et al. “Eye movements reveal a dissociation between prediction and structural processing in language comprehension”. 2025. [5] R. Levy. “Expectation-based syntactic comprehension”. In: *Cognition* 106.3 (2008), pp. 1126–1177. [6] M. Meng and M. Bader. “Mode of disambiguation and garden-path strength: An investigation of subject–object ambiguities in German”. In: *Language and Speech* 43.1 (2000), pp. 43–74. [7] P. Logačev. “The role of underspecification in relative clause attachment: Speed–accuracy tradeoff evidence.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 49.9 (2023), p. 1471. [8] D. Paape and S. Vasishth. “Local coherence and preemptive digging-in effects in German”. In: *Language and Speech* 59.3 (2016), pp. 387–403. [9] P. Schoknecht, H. Yadav, and S. Vasishth. “Do syntactic and semantic similarity lead to interference effects? Evidence from self-paced reading and event-related potentials using German”. In: *Journal of Memory and Language* 141 (2025), p. 104599. [10] J. Häussler. “The emergence of attraction errors during sentence comprehension”. PhD thesis. University of Konstanz, 2009. [11] F. Hsiao and E. Gibson. “Processing relative clauses in Chinese”. In: *Cognition* 90.1 (2003), pp. 3–27.