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

Michael Vrazitulis[1], Pia Schoknecht[1], Shravan Vasishth[1]

[1] University of Potsdam

vrazitulis@uni-potsdam.de

Background

Sentence processing models are often evaluated on narrow empirical domains, limiting claims about their generalizability. A benchmark dataset spanning multiple controlled experimental designs is needed for systematic, quantitative model evaluation. Huang et al. [1] took an important step by introducing a large-scale self-paced reading benchmark covering a range of syntactic disambiguation phenomena in English (e.g., garden paths, relative clauses, agreement attraction), revealing limitations in surprisal-based [2] accounts. However, there are currently no comparable resources for other languages, or for an even broader range of phenomena.

Method

We present GEPPU (**G**erman **E**valuation Benchmark for **P**sycholinguistics from **P**otsdam **U**niversity), a benchmark dataset for German sentence processing during reading, based on controlled experimental materials. The dataset includes both eye-tracking and self-paced reading data, drawn from two separate experiments using identical stimuli. Data collection is ongoing. The eye-tracking version of the experiment is being run in-lab with an EyeLink 1000 Plus eye tracker (1,000 Hz, right eye, N = 155 so far). The SPR version is being run online via Prolific (N = 714 so far).

Eye-tracking data collection will continue until all experimentally manipulated factors, and their interactions, yield effect estimates with 95% credible intervals within $\pm$50 ms. SPR data collection will continue up to 1,100 participants (as preregistered at https://osf.io/wpra9?view_only=2945b83dddfe4731bd60d0103559d1b4).

Each participant read 114 controlled sentences drawn from ten experimental designs, each probing a distinct psycholinguistic phenomenon: garden paths (four designs), local coherence, interference, agreement attraction, attachment ambiguities (two designs), and relative clause asymmetries.[3, 4, 5, 6, 7, 8] Sentences were followed by binary-choice comprehension questions targeting the key syntactic dependency. Items were arranged in a Latin square design with three trials per participant and experimental condition. Trial presentation order was randomized individually for each participant.

Results

Table 1 summarizes participant demographics, comprehension accuracy, and trial counts regarding the eye-tracking and self-paced reading data collected so far. Table 2 shows summary statistics for key eye-tracking and self-paced reading measures.

Discussion

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

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**	**Trials per**
			Female	Male	Other		**Accuracy (%)**	**Participant**
Eye Tracking	German	155	116	37	2	23.1 (4.4)	82.4	114
SPR	German	714	367	346	1	31.5 (9.2)	76.5	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]	236.4 $\pm$ 4.8
		First Fixation[1]	232.2 $\pm$ 4.3
		Gaze Duration[1]	303.1 $\pm$ 7.6
		Total Fixation[1]	552.0 $\pm$ 23.1
		Number of Fixations[2]	2.2 $\pm$ 0.1
		Skip Rate[3]	0.11 $\pm$ 0.01
		Regression Rate[3]	0.19 $\pm$ 0.01
SPR	Per Segment	Reading Time[1]	676.1 $\pm$ 15.3

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

References

[1] 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.

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

[3] Meng, M., & Bader, M. (2000). Mode of disambiguation and garden-path strength: An investigation of subject–object ambiguities in German. *Language and Speech*, *43*(1), 43–74.

[4] Paape, D., & Vasishth, S. (2016). Local coherence and preemptive digging-in effects in German. *Language and Speech*, *59*(3), 387–403.

[5] 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.

[6] Häussler, J. (2009). *The emergence of attraction errors during sentence comprehension* (Doctoral dissertation). University of Konstanz.

[7] 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.

[8] Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*(1), 3–27.