A Benchmark Self-Paced Reading Dataset of German Sentence Processing

Michael Vrazitulis (University of Potsdam) & Shravan Vasishth (University of Potsdam)

Recent work by [4] reports a large-scale self-paced reading benchmark dataset on English garden-path sentences that was used to evaluate the predictions of the surprisal metric [6]. This investigation revealed important gaps in the explanatory power of the surprisal account. More broadly, this kind of benchmark data has the potential to allow researchers to quantitatively evaluate the predictions of competing models, and to uncover potentially important gaps in the theoretical predictions of computational models of sentence processing.

Such a benchmark dataset is urgently needed for languages other than English, and for a broader range of phenomena than garden-pathing. Towards this end, we plan to create a benchmark dataset from a single large-sample experiment that will test a battery of theoretically postulated effects in German. The experiment will use self-paced reading (SPR) and will be administered online via Prolific (https://www.prolific.com).

We will collect data from a sample of 1,100 unimpaired native speakers of German. The study will consist of ten experimental designs with three to four conditions each (see Table 1 for a listing) arranged in the standard Latin square design (three items per condition). The order of presentation will be individually randomized for each participant. After every trial, the participant will provide a response to a binary-choice comprehension question that targets the key syntactic dependency of the sentence just presented. We will exclude from data analysis any participants that have accuracy below 70%.

We will report statistical analyses regarding the main effect or interaction in reading times (at the critical or spillover region) that is of most theoretical interest for each of the assessed phenomena. Reading times will be analyzed using Bayesian hierarchical models, and Bayes factors will be used to evaluate the evidence for the effect of interest being present.

Table 1: Sentence-processing phenomena and corresponding experimental designs.

No.	Phenomenon / Design
1	Garden paths from subject-vs.-direct-object ambiguity **2×2:** S–O/O–S × ambiguous/unambiguous; closely replicating [8]
2	Garden paths from subject-vs.-indirect-object ambiguity **2×2:** active/passive × ambiguous/unambiguous; loosely replicating [9]
3	Agreement attraction; across grammatical sentences only **2×2:** singular-/plural-control × match/mismatch; closely replicating [2]
4	Local coherence (LC) **2×2:** LC/no-LC × intervener/no-intervener; closely replicating [10]
5	Interference **2×2:** subject-cue [yes/no] × animacy-cue [yes/no]; closely replicating [11]
6	Morphosyntax-based attachment ambiguity **3×1:** high-/low-/ambiguous-attachment; closely replicating [7]
7	Semantics-based attachment ambiguity **3×1:** high-/low-/ambiguous-attachment; German adaptation of [12]
8	Garden paths from coordination ambiguity **2×2:** NP-/VP-coordination × AP-/PP-modifier; closely replicating [5]
9	Dative–genitive ambiguity **2×2:** dative/genitive × ambiguous/unambiguous; closely replicating [1]
10	Subject vs. object relative clauses (RCs) **2×2:** subject-/object-RC × double-/single-embedding; German adaptation of [3]

References

[1] J. Häussler. “Syntaktische und semantische Verarbeitungsprozesse bei der Analyse strukturell mehrdeutiger Verbfinalsätze im Deutschen: Eine empirische Untersuchung”. PhD thesis. Free University of Berlin, 2001. [2] J. Häussler. “The emergence of attraction errors during sentence comprehension”. PhD thesis. University of Konstanz, 2009. [3] F. Hsiao and E. Gibson. “Processing relative clauses in Chinese”. In: *Cognition* 90.1 (2003), pp. 3–27. [4] 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. [5] L. Konieczny, B. Hemforth, and C. Scheepers. “Head position and clause boundary effects in reanalysis”. In: *German Sentence Processing*. Springer, 2000, pp. 247–278. [6] R. Levy. “Expectation-based syntactic comprehension”. In: *Cognition* 106.3 (2008), pp. 1126–1177. [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] 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. [9] M. Meng and M. Bader. “Ungrammaticality detection and garden path strength: Evidence for serial parsing”. In: *Language and Cognitive Processes* 15.6 (2000), pp. 615–666. [10] D. Paape and S. Vasishth. “Local coherence and preemptive digging-in effects in German”. In: *Language and Speech* 59.3 (2016), pp. 387–403. [11] P. Schoknecht and S. Vasishth. “Do syntactic and semantic similarity lead to interference effects? Evidence from self-paced reading and event-related potentials using German”. Under review. 2023. [12] M. J. Traxler, M. J. Pickering, and C. Clifton Jr. “Adjunct attachment is not a form of lexical ambiguity resolution”. In: *Journal of Memory and Language* 39.4 (1998), pp. 558–592.