Abstract. *

Format: Up to 350 word. Markdown can be used for simple text formatting: You can emphasize text with bold or italics by using asterisks. You can also use lists, links, and backticks for code. See here for details. (Note that this form will not render Markdown, but it will be rendered in the workshop program.)

In sentence processing research, theories are often developed for explaining specific phenomena (e.g., subject vs. object relative clause processing), but these theories' explanatory capabilities are often not evaluated beyond a small range of empirical phenomena. Missing in the field is a standard benchmark data-set encompassing a range of phenomena that can be used for quantitative model evaluation and model comparison. A first step in this direction was taken by Huang et al. (2024), who created a large-scale self-paced reading benchmark data-set on English garden-path sentences. These data were used by Huang et al. to quantitatively evaluate the predictions of the surprisal metric (Levy, 2008, Hale, 2001). 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 data-set is urgently needed for eye-movements while reading and for languages other than English, as well as for a broader range of phenomena than garden path sentences. Towards this end, we are creating a benchmark data-set from a large-sample experiment (currently N=89, planned N=120) that tests a battery of theoretically postulated effects in German (e.g., different types of garden paths, agreement attraction, attachment ambiguities, similarity-based interference, subject vs. object relative clauses). An additional experiment is being run using web-based self-paced reading (SPR) methodology (currently N=586, planned N=1,100). By creating large-scale eye-movements and SPR data-sets for the same linguistic material, a further aim is to provide a potentially useful way to investigate the relationship between eye-tracking measures and self-paced reading times.
As an initial analysis, we will discuss the data in the light of predictions based on surprisal values calculated for the experimental items.
Once published, the full benchmark data-set will be freely available to be used for evidence-based model building in sentence processing.

Hale (2001). NAACL.
Huang et al. (2024). JML, 137, 104510.
Levy (2008). Cognition, 106(3), 1126-1177.