

A Multilingual Cross-Method Dataset for Evaluating the Measurement Reliability of Individual Differences in Naturalistic Reading

Cui Ding (cui.ding@uzh.ch)¹, Shan Gao¹, Lena A Jäger¹

¹Department of Computational Linguistics, University of Zurich

Background: Incremental reading paradigms are widely used to measure processing difficulty during reading. Prior comparisons of eye-tracking, self-paced reading (SPR), and maze tasks have shown considerable similarities in the overall effects of linguistic predictors, while also revealing differences in effect magnitude, timing, and localization [1]. Recent variants such as A-maze [2], bidirectional SPR [3], and Mouse Tracking for Reading (MoTR) [4] further expand this methodological toolkit, including for studies of individual differences. However, most cross-paradigm comparisons have focused on English, leaving open the question of how reliable these methods are for individual-difference studies across languages. We address this gap with a multilingual dataset in which participants read naturalistic texts using bidirectional SPR, the A-maze task, and MoTR across two sessions per method. We estimate both within-method cross-session reliability and cross-method reliability of individual differences in naturalistic reading.

Method: Native speakers of English, German, Chinese complete the experiment in two sessions. In each session, participants read naturalistic materials using the three paradigms, with the order of paradigms randomly assigned. Stimuli comprise two types of materials: out-of-context naturalistic sentences sampled from the same books or source types as the MultiPLEYE materials, but from non-overlapping passages [5], and longer texts composed of original MultiPLEYE passages. For each paradigm, participants first complete the sentence-reading task, then read one MultiPLEYE text, and then answer the original comprehension questions. To assess reliability, we use the two-task Bayesian hierarchical model [6] adapted to naturalistic reading by [7]. We estimate within-method cross-session reliability and cross-method reliability, focusing on individual differences in sensitivity to word length, lexical frequency, and predictability.

Results & Discussion: Data collection is scheduled to begin shortly. The resulting dataset will provide a multilingual resource for evaluating the suitability of web-deployable reading paradigms for individual-differences research across languages and writing systems. The planned analyses will assess how reliably SPR, maze, and MoTR capture participant-level sensitivity to word length, lexical frequency, and predictability. By estimating within-method cross-session reliability, the study will assess whether each paradigm captures stable individual-level variation over time. By estimating cross-method reliability, it will reveal whether different paradigms capture convergent individual differences in the effects of linguistic predictors during naturalistic reading.

References

- [1] Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, 41(2), 105–128.
- [2] Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111, 104082.
- [3] Paape, D., Vasishth, S., et al. (2022). Conscious rereading is confirmatory: Evidence from bidirectional self-paced reading. *Glossa Psycholinguistics*, 1(1).
- [4] Wilcox, E. G., Ding, C., Sachan, M., & Jäger, L. A. (2024). Mouse Tracking for Reading (MoTR): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138, 104534.
- [5] Jakobi, D. N., Stegenwallner-Schütz, M., Hollenstein, N., Ding, C., Kaspere, R., Matić Škorić, A., Pavlinusic Vilus, E., Frank, S., Müller, M.-L., Jensen de López, K. M., et al. (2025). MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications* (pp. 1–11).
- [6] Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, 26(2), 452–467.
- [7] Haller, P., Ding, C., Stegenwallner-Schütz, M., Reich, D. R., Koncic, I., Makowski, S., & Jäger, L. A. (2026). Replicate me if you can: Assessing measurement reliability of individual differences in reading across measurement occasions and methods. *Cognitive Science*, 50(1), e70121.