Slovenian Data Collection Within an International Eye-tracking Project

Lea Popović, Lamija Zahirović, Anastasija Žarevac, Kristian Elersič

University of Ljubljana
Kongresni trg 12, 1000 Ljubljana
lp58329@student.uni-lj.si, lz67616@student.uni-lj.si, az26388@student.uni-lj.si, kristian.elersic@ff.uni-lj.si

1 Introduction

Eye-tracking is one of the more reliable methods for studying written language processing, but the field is dominated by monolingual datasets, limiting comparability across different languages. To improve this, the MultiplEYE COST Action is developing an international multilingual open-access corpus of eye-tracking reading data using standardized data-collection and preprocessing procedures. Our contribution to this project is collecting quality eye-tracking and psychometric data from 100 native Slovenian speakers.

2 Methods

The experiment consists of two sessions. In the first session, participants read naturalistic texts in their native language, including two practice and ten experimental texts, translated to Slovenian within the MultiplEYE protocol, covering multiple genres and varying in complexity levels. The texts are followed by two questions assessing their familiarity and difficulty, as well as six comprehension questions.

Participants' eye movements are recorded using the EyeLink1000 eye-tracker. Following the first session, participants fill out a demographic questionnaire. In the second session, the participants take psychometric tests (WMC, RAN, Stroop, WikiVocab, Flanker, and PLAB) which assess their cognitive and language-related abilities.

MultiplEYE strives to achieve the FAIR (Findable, Accessible, Interoperable, and Reusable) principles through standardizing data collection, preprocessing and quality control, which ensures reproducibility and comparability across languages (Jäger et al. 2026; Müller 2024).

3 Goals and preliminary findings

MultiplEYE's main goal is creating an open-access database which will serve as a resource for research in psychology, linguistics, and machine learning (Jäger et al. 2026). The multilingual and standardized nature of the project will enable testing linguistic theories and comparing reading behaviour across a wide set of languages. Connecting eye-tracking data with psychometric test results could provide valuable insights into the relationship between an individual's cognitive abilities and their reading abilities. Additionally, the information about fixation durations and saccade patterns could be valuable in training natural language processing models (Čuber and Prinčič 2025).

We have started data quality checks in Python (PyCharm) to ensure data integrity, and conducted an exploratory analysis examining whether regression rates during reading (how often readers move their gaze back) are related to working memory capacity (WMC), which is to be confirmed on the full dataset.

The preliminary analysis conducted on data from 32 participants suggests a positive association between WMC and regression rates across all texts. Specifically, higher WMC was associated with more frequent regressions. Further analyses on the full dataset are needed to confirm and interpret these preliminary results.

Besides contributing to international multilingual research, we hope the data collected from the Slovene-speaking sample could inspire further investigations into the interindividual differences within the population.

References

Jäger, L. A., Hollenstein, N., Ana, M. , Jakobi, D. N., Stegenwallner-Schütz, M., Ding, C., Eva, P. V., Kasperė, R. and Müller, M.-L. (2026). Multipleye: Enabling multilingual eye-tracking data collection for human and machine language processing research. *Psycharchives.org*.

Müller, M.-L. (2024). Data management in eye-tracking research: Crucial steps and challenges. *Psycharchives.org*.

Čuber, A. and Prinčič, G. (2025). Standardizing multilingual eye-tracking for psycholinguistic and nlp research. *Proceedings of the MEi:CogSci Conference*, 19(1).