Eye Tracking Based Cognitive Evaluation of Automatic Readability Assessment Measures

Keren Gruteke Klein[1], Shachar Frenkel[1], Omer Shubi[1], Yevgeni Berzak[1,2]

[1]Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology, Haifa, Israel

[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA

{gkeren,fshachar,shubi}@campus.technion.ac.il, berzak@technion.ac.il

Introduction Automatic Readability Assessment (ARA) aims to evaluate text readability for diverse populations, supporting applications in education, accessibility, and text simplification. Traditional ARA systems rely on comprehension outcomes or human-labeled readability levels as ground truth, but both face key limitations: comprehension scores depend on task design rather than reading ease, while annotations suffer from low agreement, limited granularity, and topic confounds [1, 2, 3, 4, 5]. We propose a cognitive framework that evaluates readability directly from behavioral evidence using eye movements reflecting reading ease.

Design We evaluate readability measures using the OneStop Eye Movements dataset [6], which contains eye-tracking data from 360 English L1 adults reading original and human-simplified *Guardian* news paragraphs. Each participant read 54 paragraphs (27 original, 27 simplified), with different reader groups assigned to each version, and 60 readers per paragraph. We test traditional formulas, modern NLP-based measures, LLMs, commercial tools (Lexile, TextEvaluator), and psycholinguistic predictors including idea density, integration cost, embedding depth, word length, frequency, entropy and surprisal. Our main evaluation criterion is reading facilitation from simplification, quantified by differences in reading ease measures (e.g., Total Fixation Duration, averaged across participants) between the original and simplified versions. For each textual item, we compute (1) $\Delta\text{ReadabilityScore}_{M,T}$, the difference in measure M between the two versions of text T, and (2) ΔRT_T, the difference in reading ease between the same text versions. Then, the Pearson r correlation between $\Delta\text{ReadabilityScore}_{M,T}$ and ΔRT_T indicates the predictive quality for measure M.

Results Figure 1 presents the Pearson r correlations for all readability measures at the sentence level. We find that despite their wide adoption, existing ARA measures which are tuned on comprehension outcomes and readability level annotations are poor predictors of reading ease, outperformed by entropy and the big three predictors of reading times, and in particular by surprisal. The advantage of these measures is especially clear in Regression Rate, where surprisal yields the highest correlations, while all other measure groups show mostly non-significant effects. The results also hold at the paragraph level and when using Spearman ρ correlation.

Discussion We propose a new evaluation framework for readability assessment that emphasizes reading ease, using eye-tracking data over parallel corpora of original and simplified texts. This controlled design enables a principled evaluation of existing readability measures and commercial scoring systems. Our findings show that widely used ARA metrics, designed around comprehension outcomes and level annotations, are weak predictors of reading ease, while entropy and the big three predictors of reading times, especially surprisal, perform substantially better. In future work, we aim to extend this framework to additional reader populations and languages, and to use it for developing new cognitively grounded readability measures.

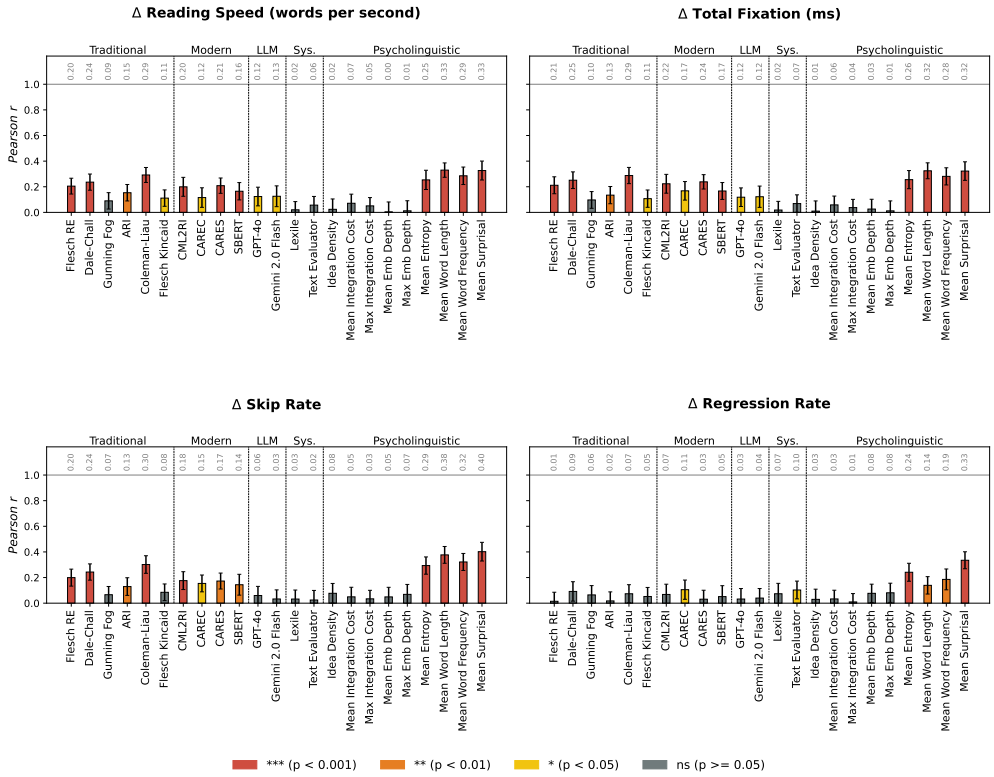

Figure 1: Evaluation of readability and psycholinguistic measures using reading facilitation, at the sentence level. Depicted are Pearson r correlations between $\Delta\text{ReadabilityScore}_{M,T}$, the readability difference between original and simplified texts according to measure M, and ΔRT_T, the average reading measure difference on the same pairs of texts. Error bars are 95% confidence intervals. Colors represent the statistical significance level of the correlation.

References

[1] Vajjala, S., & Lučić, I. (2019). On understanding the relation between expert annotations of text readability and target reader comprehension. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 349–359.

[2] Berzak, Y., Malmaud, J., & Levy, R. (2020). STARC: Structured annotations for reading comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5726–5735. https://doi.org/10.18653/v1/2020.acl-main.507

[3] Gruteke Klein, K., Shubi, O., Frenkel, S., & Berzak, Y. (2025, February). The effect of text simplification on reading fluency and reading comprehension in l1 english speakers. https://doi.org/10.31234/osf.io/dhk8c

[4] Vajjala, S. (2022). Trends, limitations and open challenges in automatic readability assessment research. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5366–5377.

[5] Brunato, D., De Mattei, L., Dell'Orletta, F., Iavarone, B., Venturi, G., et al. (2018). Is this sentence difficult? do you agree? *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2690–2699.

[6] Berzak, Y., Malmaud, J., Shubi, O., Meiri, Y., Lion, E., & Levy, R. (2025). Onestop: A 360-participant english eye-tracking dataset with different reading regimes. *PsyArXiv preprint*.