LREC 2026

The Second International Workshop on Eye-Tracking Resources and Evaluation for Human-Aligned NLP (Gaze4NLP 2026)

Workshop Proceedings

Editors
Cengiz Acartürk, Burcu Can, Jamal Nasir, Çağrı Çöltekin

May 12, 2026

Proceedings of The Second International Workshop on Eye-Tracking Resources and Evaluation for Human-Aligned NLP (Gaze4NLP 2026)

ISBN 978-2-493814-84-5

Preface

It is our great pleasure to present the proceedings of Gaze4NLP: The International Workshop on Eye-Tracking Resources and Evaluation for Human-Aligned NLP, held on May 12, 2026, in Palma de Mallorca, Spain, as part of the 2026 International Conference on Language Resources and Evaluation (LREC2026).

The first Gaze4NLP workshop (International Workshop on Gaze Data and Natural Language Processing) was held on September 12, 2025 in Varna, as part of RANLP (Recent Advances in Natural Language Processing). This second edition again brought together researchers from diverse backgrounds to discuss advances in the field, eye-tracking resources and evaluation methods for human-aligned NLP.

Each paper in the proceedings was reviewed by at least two members of our program committee. The contributions cover a wide range of topics, including:

- using gaze for machine translation evaluation;
- part-of-speech data to study individual differences in reading Portuguese;
- analysing the relationship between visual perception and language production;
- analysing a low-resource language eye-movement corpus in Arabic;
- tracking facial expressions and eye-contact in VR environments;
- conducting automatic text simplification using eye-tracking measures in English;
- evaluating the impact of text simplification in French;
- comparing mouse and eye tracking in reading Romanian texts; and
- estimating on-screen gaze location using mouse data.

The organizers also present a survey of research on incorporating gaze data in NLP models and applications.

We would like to express our gratitude to the authors for their high-quality submissions and to the program committee members for their contributions to the reviewing process.

Cengiz Acartürk, Burcu Can, Jamal Nasir and Çağrı Çöltekin

Organizing Committee

Cengiz Acartürk, Jagiellonian University, Poland
Burcu Can, University of Stirling, Scotland, UK
Çağrı Çöltekin, University of Tübingen, Germany
Jamal Nasir, University of Galway, Ireland

Programme Committee

Ana Matić Škorić, University of Zagreb, Croatia
Fariz Ikhwantri, Simula Research Laboratory, Norway
Irina Temnikova, Big Data for Smart Society Institute (GATE), Bulgaria
Joseph Lemley, University of Galway, Ireland
Melike Caglayan, Jagiellonian University, Poland
Mila Vulchanova, Norwegian University of Science & Technology, Norway
Natalia Grabar, CNRS & Université de Lille, France
Noam Siegelman, Hebrew University of Jerusalem, Haskins Laboratories, Israel
Oksana Ivchenko, University of Lille, France
Özge Alaçam, University of Bielefeld, Germany
Sergiu Nisioi, University of Bucharest, Romania

Table of Contents

Workshop Program

Tuesday, May 12

09:00 **Session 1: Eye-Tracking Data in Reading and Language Processing**

Eye tracking for Machine Translation Quality Evaluation
Natalia Glazyrina and Ondřej Bojar

Cross-Linguistic Analysis of Eye Movement Patterns: Insights from the First Arabic Eye-Tracking Corpus for NLP
Ibtehal Baazeem, Hend Al-Khalifa and Abdulmalik AlSalman

Exploring Cognitively Informed Sentence Simplification with Gaze-Guided Text Generation
Andreas Säuberli, Diego Frassinelli and Barbara Plank

Impact of Text Simplification on Eye-Tracking-Based Reading Profiles Across Domains
Oksana Ivchenko and Natalia Grabar

11:00 **Session 2: Eye-Tracking Data in Multimodal Context**

Parts of Speech Shape Reading-Time Variability in Brazilian Portuguese
Diego Alves

CoordiMap: Conceptual Proposition of a new Framework for the Annotation of Verbal Elicitation Paths on Visual Experiment Stimuli and Introduction of the Associated Annotation Tool
Carmen Schacht

A Comparative Study Between Mouse and Eye Tracking Signals for Long Romanian Texts
Bogdan Alexandru Gheorghe and Sergiu Nisioi

Eye-Contact and Facial Expression Tracking for Assertiveness Training in VR-Based Anti-Bullying Education
Lubomir Ivanov, Anabel Nolasco and Mary Vrahimis

Predicting Gaze Location without Camera or Eye-Tracker
Saman Rezapoor, Sajad Shirali-Shahreza and Gerald Penn

A Survey of Incorporating Gaze Data into Natural Language Processing Models and Applications
Cengiz Acarturk, Burcu Can, Melike Caglayan, Jamal Abdul Nasir and Cagri Coltekin

Eye Tracking for Machine Translation Quality Evaluation

Natalia Glazyrina, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic
nglazyrinav@gmail.com, bojar@ufal.mff.cuni.cz

Abstract

Eye tracking offers unique insights into cognitive processes, making it a promising tool for evaluating machine translation (MT). This study explores the feasibility of using an iPhone 12 camera-based eye tracker with a 14-inch laptop display for conducting translation evaluation in personal workspaces, offering a more accessible and cost-effective alternative to traditional setups. Participants evaluated source sentences, selected translations, and identified problematic words while their gaze metrics were recorded and analyzed. Our findings reveal statistically significant correlations between gaze patterns and preferred translations, as well as increased visual attention to problematic words. These results demonstrate that home-based eye tracking systems are technically sufficient for capturing gaze behavior accurately enough for MT evaluation purposes. A potential practical application is to speed up translation proof-reading using eye tracking technique to automatically mark portions of text that should be attended to and improved based on the gaze pattern during a quick reading.

Keywords: eye tracking, machine translation, correlation, fixation, saccade

1. Introduction

Human evaluation of machine translation (MT) quality remains a crucial aspect in the advancement of translation technology. However, the subjectivity inherent in human judgment poses a challenge in achieving a reliable and consistent assessment. In recent years, eye tracking technology has emerged as a promising tool to delve into the cognitive processes that underlie translation evaluation (Stymne et al., 2012; Sajjad et al., 2016). By capturing individuals' reading patterns during the evaluation of translation options, eye tracking provides insights into the linguistic cues that influence decision-making, such as incorrect word order, morphological disagreement, and semantic ambiguity, thus offering a more objective lens to complement traditional subjective and self-reported evaluation methods.

Previous studies have demonstrated the potential of eye tracking in predicting the preferred translation among multiple options (Sajjad et al., 2016; Doherty et al., 2010) or machine translation error analysis (Stymne et al., 2012). These investigations have employed standalone eye tracking systems to monitor participants' gaze movements, revealing that poorly translated text causes readers to frequently jump back while reading, which serves as a measurable marker of processing difficulty. Furthermore, research demonstrates that "bad" sentences result in significantly higher gaze times and fixation counts compared to high-quality ones.

In this paper, we contribute to the evolving landscape of MT evaluation using an iPhone camera-based eye tracking approach. This approach was chosen over webcam-based due to gaze tracing qualities revealed during the comparison of eye tracking systems. Unlike conventional standalone systems, this methodology offers a pragmatic alternative, avoiding the need for dedicated eye tracking hardware and enabling broader accessibility. This approach embraces real-world scenarios, where users can employ their own devices for evaluation. By lowering technical and financial barriers, this method aims to democratize access to eye tracking technology, allowing researchers and practitioners to integrate cognitive insights into MT evaluation without relying on specialized equipment. Such an accessible solution has the potential to expand the reach of eye tracking research to diverse, including non-specialist, environments, making the evaluation process more inclusive and practical.

Despite the broad use and advances in automatic evaluation of machine translation, see e.g. Lavie et al. (2025), human evaluation remains the gold standard in the field of machine translation. Automatic metrics, while scalable, often struggle to capture semantic nuances, stylistic consistency, and the actual cognitive load experienced by a reader. This enduring importance is evidenced by the annual Conference on Machine Translation (WMT; Kocmi et al., 2025), where human judgment serves as the benchmark for validating the accuracy of automated systems.

However, traditional human evaluation often treats the translator's or rater's decision as a "black box," focusing on the final output rather than the process. By integrating eye tracking, we can move beyond simple preference scores to observe the cognitive effort involved in processing translation errors.

We implement an experimental design in which participants are presented with a source sentence

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 1–9
May 12, 2026.

in English and two target candidate translations in Russian. Participants are tasked with selecting a better option, while also identifying problematic words within the suboptimal choice.

To evaluate the efficacy of the approach, we analyze the correlations between eye movement metrics (such as fixation/saccade count and time spent on each sentence) and participants' translation choices. We hypothesize that the utilization of an iPhone camera-based eye tracker can be used to assess that correlation and to substantiate that problematic words within suboptimal translations are associated with a higher concentration of gaze fixations and gaze saccades.

2. Related Works

Although human judgment remains the gold standard, as seen in the annual Conference on Machine Translation (WMT),[1] human evaluation is not without flaws; it is resource-intensive and prone to high inter-annotator variability and subjectivity (Graham et al., 2013; Lommel et al., 2014).

To bridge the gap between automated scores and subjective human ratings, researchers have turned to eye tracking. The foundational assumption of eye tracking is the "eye-mind hypothesis" (Just and Carpenter, 1980), which suggests a link between gaze fixation and cognitive processing of linguistic content. Based on this theory, Doherty et al. (2010) aimed to explore whether eye tracking data can reflect the quality of MT output as rated by human evaluators and whether eye tracking could be used as a semi-automated tool for evaluating MT quality. This study analyzed various eye tracking metrics, including gaze time, fixation count, fixation duration, and pupil dilation. The results indicated correlations between eye tracking metrics and the quality of MT output as rated by evaluators. Specifically, "bad" sentences had longer gaze times and more fixations compared to "good" sentences. The duration of fixation and pupil dilation showed less consistent correlations.

Building on this, Sajjad et al. (2016) utilized eye tracking data to address the subjectivity and low inter-annotator agreement often found in traditional human judgments. The authors demonstrated that specific reading patterns, such as the number of regressions and the total reading time, effectively distinguish between high- and low-quality translations. They found that combining eye tracking features with BLEU scores (Papineni et al., 2002) yielded promising results in predicting translation quality, indicating that reading patterns capture more than just fluency. This suggests that gaze data capture cognitive nuances, such as semantic processing effort, that surface-level n-gram overlap metrics like BLEU inherently overlook.

Furthermore, Bojar et al. (2016) investigated the cognitive drivers of inter-annotator disagreement within the WMT Shared Translation Task. Using a high-precision EyeLink II tracker in a controlled laboratory setting, the authors found that inconsistent rankings often stemmed from specific error types – mainly in translations that displayed high fluency but low adequacy. Their gaze data revealed that these "deceptive" translations caused significant uncertainty and longer processing times. The study also highlighted the cognitive burden of the source text, noting that annotators focused more on source sentences than references, which was expected because the participants were native speakers of the target language but only second-language learners for the source.

Despite the established benefits of eye tracking metrics, their integration into large-scale MT evaluation has been hindered by a reliance on expensive, lab-bound hardware. Lately, several studies have compared webcam-based eye tracking systems with traditional in-lab systems, evaluating their viability across different research domains. In psycholinguistics, webcam-based systems have been used to study language processing in naturalistic environments, providing accessibility to diverse populations and geographically dispersed participants. For instance, Özsoy et al. (Özsoy et al., 2023) investigated heritage language processing using webcam-based eye tracking, demonstrating that data collected in such settings was largely consistent with in-lab systems. This approach facilitated the inclusion of heritage speakers who otherwise might not have access to laboratory facilities. Other studies have replicated psycholinguistic effects, such as the verb semantic constraint and lexical interference effects, using webcam-based tracking, confirming its ability to capture both robust and subtle phenomena (Prystauka et al., 2023). Similarly, a recent replication of a Visual World study on verb aspect processing showed that webcam-based eye tracking, even with off-the-shelf tools, can achieve comparable results to infrared systems, offering a cost-effective and accessible alternative (Vos et al., 2022). These findings suggest that while webcam-based systems can reliably replicate key effects, careful attention must be given to factors like calibration, lighting, and participant guidance to ensure data quality.

Our work contributes to this shift by exploring the efficacy of iPhone-based eye tracking specifically for MT quality evaluation. By moving the experimental environment from the controlled laboratory to a home-based setting, we aim to lower the financial and technical barriers to high-quality, "process-oriented" human evaluation. This approach not only

[1] https://www.statmt.org/wmt25/

democratizes access to cognitive data but also introduces a new layer of quality control, allowing researchers to filter human annotation based on real-time cognitive engagement and attentional focus.

3. Methodology

3.1. Tool Selection

The experimental configuration was finalized after a comparative pilot of webcam-based systems. We initially evaluated jsPsych[2] with the WebGazer.js library. While jsPsych is a well-established tool for behavioral experiments, it presented several limitations. To validate the accuracy of each system, we conducted a controlled reading task where the researcher read the stimulus text slowly and linearly, line-by-line. As shown in Figure 1, the resulting gaze trace for the webcam-based system was highly distorted and failed to follow the horizontal progression of the lines. Furthermore, the gaze coordinates collected during trial runs were challenging to interpret, complicating the analysis.

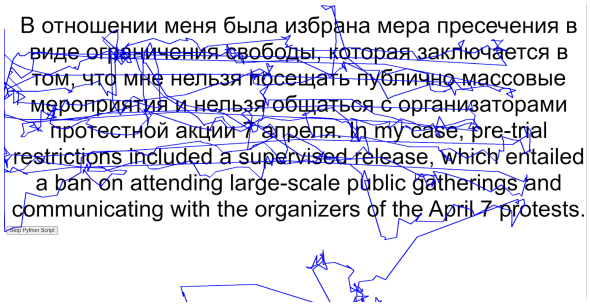


Figure 1: jsPsych eye trace mapped on the screen.

Consequently, we selected Eyeware Beam for data collection. This software supports high-precision tracking via an iPhone 12 mini camera (connected via USB) at a sampling rate of 90 Hz. The iPhone-based approach was chosen over standard webcams due to its far more accurate mapping, demonstrated in Figure 2. The camera was positioned horizontally at the base of a 14-inch laptop screen on the left side. Unlike previous studies, the participants' head positions were not fixed, and the exact screen-to-eye distance was not rigorously controlled. However, participants adhered to the eye tracker's recommended distance of approximately 50-60 cm.

3.2. Participants

Participants were recruited via convenience sampling from a pool of graduate-level volunteers within

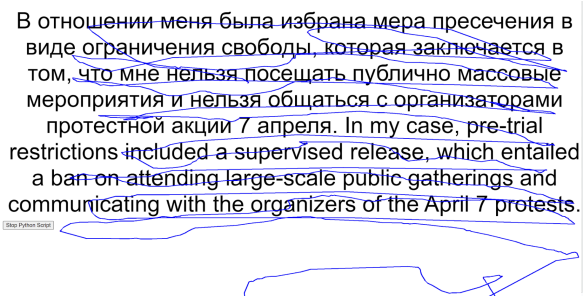


Figure 2: Eyeware Beam eye trace mapped on the screen.

a university environment. Participation was entirely voluntary, and no financial compensation was provided. All participants were informed of the study's objectives and the nature of the eye tracking data being collected prior to the start of the trial. For this pilot study we recruited 8 participants (4 male, 4 female) with the following profiles:

- **Language**: Native Russian speakers with B2+ English proficiency.
- **Age**: 25–30 years
- **Education**: Graduate-level or higher.
- **Vision**: Normal or corrected-to-normal vision (no glasses were worn during this specific trial to ensure maximum tracker stability).

The experimental protocol was designed following the principle of data minimization. The utilized software processes the camera feed locally in real-time to calculate gaze vectors. Crucially, no raw video or photographic data of participants was stored at any point during the study. The exported data consisted exclusively of numerical logs containing temporal markers (timestamps) and spatial gaze coordinates relative to the screen. Since no personally identifiable information (PII) was linked to the gaze logs, the dataset is inherently anonymized.

3.3. Research Materials

The test stimuli consisted of sentence pairs extracted from the WMT Metric Task (2021[3] and 2022[4]) datasets.

- **Structure**: 10 distinct screen sets, each containing 10 experimental screens.
- **Layout**: A standardized interface displaying one English source sentence at the top and

[2] https://www.jspsych.org/7.3/

[3] https://drive.google.com/drive/folders/1TNIeXirfNMa6WV7LlS3Z51UxNNCgGcmS

[4] https://drive.google.com/file/d/1I0O-NzOLCxrO6noub2pY81BtWxp42A46/view

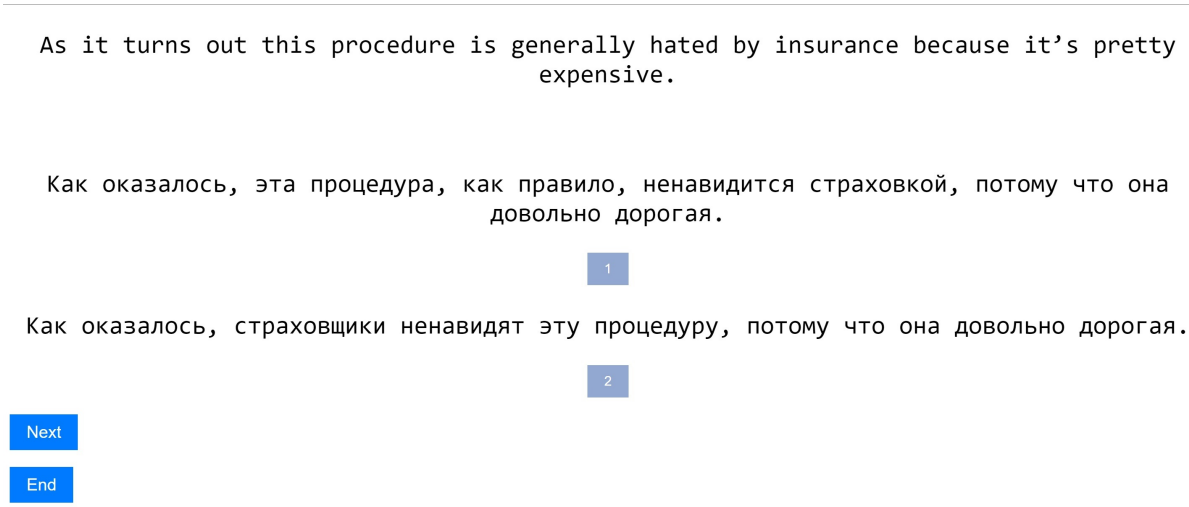


Figure 3: Example of the screen layout.

two candidate Russian translations (labeled "1" and "2") below (see Figure 3).

- **Calibration**: A warm-up set was provided to familiarize users with the interface, and calibration was verified at the start of each session and after breaks.

3.4. Experimental Task and Procedure

Participants were asked to perform a dual-stage evaluation task designed to capture both preference and cognitive load:

1. **Comparative Judgment**: Participants read the source and both translations, then selected the superior candidate by clicking a corresponding button. To prevent positional bias, the order of the translation candidates was randomized. Consequently, the 'better' translation appeared as either the first or second option with equal frequency throughout the experiment.

2. **Error Span Identification**: In the suboptimal translation, participants were instructed to click on specific words or phrases they perceived as problematic. This design follows Maja Popović's research (Popović, 2020) on identifying challenging sentence segments in machine translation, though no distinctions were made between different types of errors in this study.

To ensure data integrity, gaze data recorded during the clicking action (identification phase) was excluded from the cognitive load analysis. This allows us to isolate the uninterrupted reading process from the manual task of error marking.

3.5. Quality Control

A key contribution of this methodology is the use of gaze data as a quality control layer. By analyzing fixation density and saccadic movements, we can identify "inattentive" trials where the participant may have skimmed the text without full cognitive engagement. This enables the exclusion of unreliable human data that are typical for remote, home-based annotation tasks.

4. Analysis

4.1. Gaze Data Post-Processing

For the extraction of fixations and saccades for further analysis, a post-processing procedure was employed on the collected data. Notably, the collected traces exhibited a noticeable shift along the y-axis, possibly attributable to inaccuracies in the calibration process or head movement during the experiment. This phenomenon is illustrated in Figure 4. Consequently, a manual adjustment was required, using a constant addition to the y-coordinate across the entire trace for each screen. The modified, post-processed trace is shown in Figure 5.

It is worth noting that during the experiment a few times participants misclicked on the "Next" button and accidentally skipped a screen without noticing it. These occurrences were infrequent (5 times) and pointed to drawbacks of the technical implementation of the experiment. Those 5 screens are skipped in the analysis.

To analyze gaze behavior, we extracted features related to gaze fixations and saccades using the Velocity-Threshold Identification algorithm (Salvucci et al., 2000), with a velocity threshold set to 100. This algorithm identifies fixations and saccades based on point-to-point velocity. Our analy-

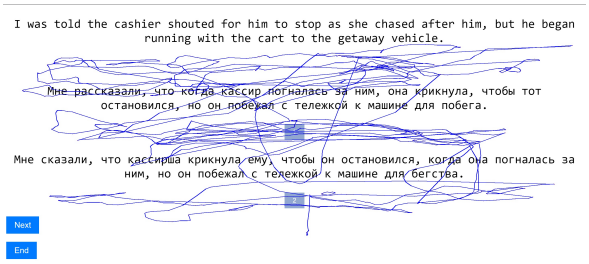


Figure 4: Mapping of originally collected trace of gaze.

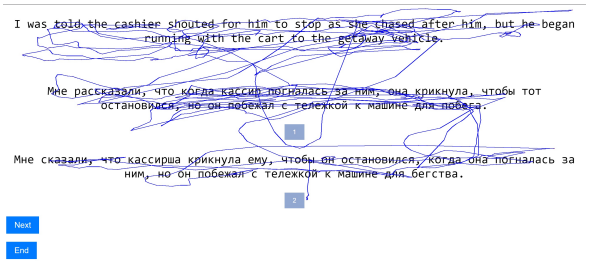


Figure 5: Mapping of shifted along y-axis trace of gaze.

sis includes two levels: word-level and sentence-level.

4.2. Word-level Analysis: The Cost of Errors

For the word-level analysis, we prepared two datasets, each containing three features, where each row represented data for a single screen and participant:

1. Dataset 1: Number of identified problematic words, number of fixations on these words, and total number of fixations on the screen.

2. Dataset 2: Number of identified problematic words, number of saccades on these words, and total number of saccades on the screen.

In both datasets, we conducted correlation analysis by calculating the Pearson correlation coefficient (PCC) between the number of words marked as problematic and the relative share of visual attention (fixations and saccades) those words received.

The scatter plots (Figures 6 and 7) illustrate these relationships. While there is a high density of points at low error counts, a clear upward trend is visible:

- **Fixation Proportion**: PCC=0.30 ($p < 0.001$)

- **Saccade Proportion**: PCC=0.44 ($p < 0.001$)

While the correlation coefficients indicate a low-to-moderate relationship, they are highly statistically significant. The higher correlation for saccades (0.44) suggests that problematic segments do not merely cause the eye to linger; they are more strongly associated with re-scanning behaviors as participants repeatedly glance back at the source sentence to check the original meaning whenever they run into a problematic translation. This finding implies that problematic words attract a disproportionate share of visual attention.

4.3. Sentence-level analysis: Predicting Preference

For the sentence-level analysis, we derived the following features: number of fixations per sentence, number of saccades per sentence, time spent on each sentence. Using these features, we modeled the participants' final translation choice (Sentence 1 vs. Sentence 2) using both a Logistic Regression (LR) model for statistical significance and a Decision Tree (DT) for behavioral interpretability.

4.3.1. Logistic Regression

The LR model was implemented using the statsmodels library (Seabold and Perktold, 2010) with default parameters, except for the maximum iteration parameter, which was set to 100.

The model summary revealed time spent and fixation counts on the second sentence as significant predictors ($p < 0.05$) with the following coefficient signs:

- Time spent on sentence 2: Negative

- Fixations/saccades on sentence 2: Positive

The coefficient signs reveal a "comparative pressure" effect: an increase in time spent on Sentence 1 significantly increases the probability of the user choosing Sentence 2. This suggests that the time metric captures the "struggle" to find meaning; when one candidate is difficult to parse, the user's preference shifts to the alternative. Additionally, a higher number of fixations or saccades on sentence 2 indicates tendency to choose that sentence.

4.3.2. Decision Tree Interpretation

To derive actionable thresholds for these behaviors, we trained a Decision Tree classifier (depth=3) using scikit-learn library (Pedregosa et al., 2011). Unlike the LR model, which provides a probability gradient, the DT identifies the exact points in the decision-making process.

The model's Feature Importance (Table 1) indicates that the decision-making process is primarily driven by metrics associated with the second

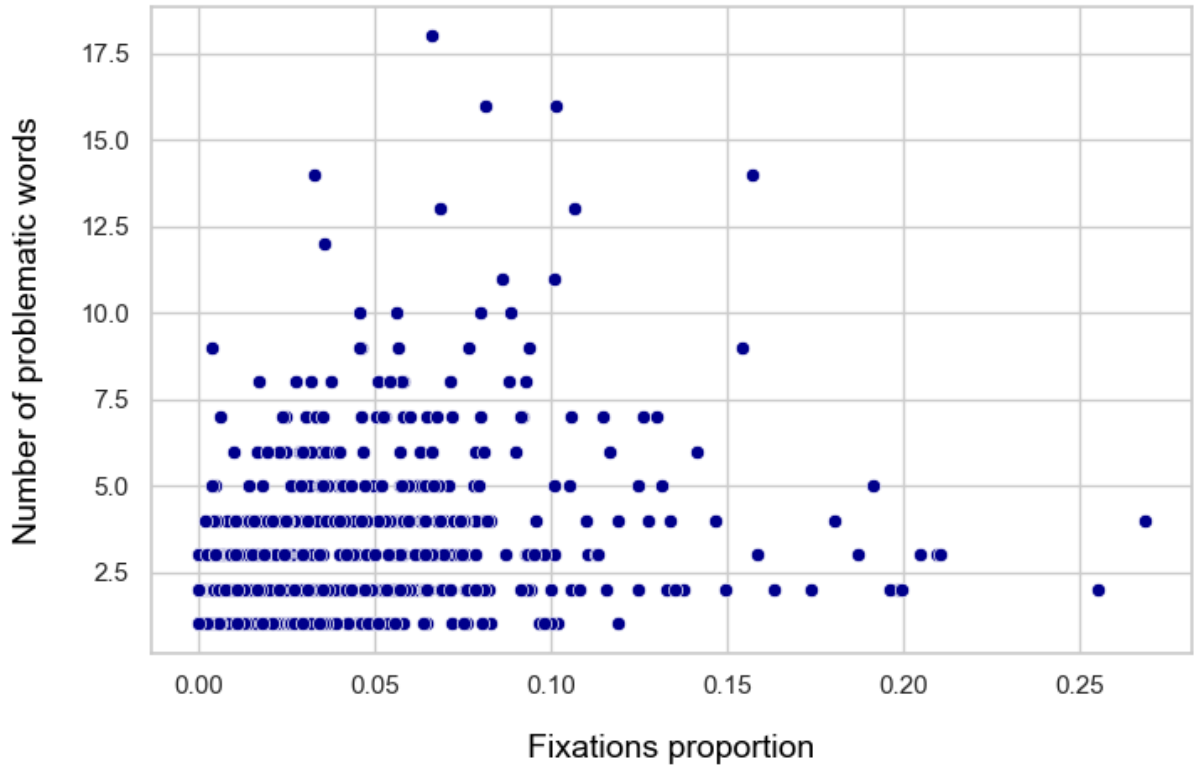


Figure 6: Number of problematic words vs. fixation proportion. Pearson correlation coefficient 0.30.

Feature	Importance
Time spent sentence 2	0.39
Saccades on sentence 2	0.32
Fixations on sentence 2	0.20
Saccades on sentence 1	0.08

Table 1: Feature Importance extracted from Decision Tree.

translation candidate. Specifically, time spent on sentence 2 (39.1%) and saccades on sentence 2 (32.3%) were the most influential factors, while metrics for sentence 1 provided significantly less predictive power.

The tree structure (Figure 8) revealed highly interpretable behavioral "thresholds." For instance, a specific path in the tree identified a high-certainty node (Gini = 0.188) where a low number of saccades (≤ 8.5) combined with a limited time investment (≤ 81 gaze units) on Sentence 2 led to a consistent selection of that candidate. This suggests that "fluency" – characterized by rapid, linear processing – is a stronger predictor of preference than simply the total amount of attention paid to a sentence.

5. Discussion

The results of this pilot study suggest that iPhone-based eye tracking is a viable, low-cost method for capturing cognitive effort in MT evaluation. The correlation found between visual attention – specifically gaze duration and fixation counts – and the final translation choice aligns with the "eye-mind hypothesis" (Just and Carpenter, 1980), suggesting that participants spend significantly more time processing suboptimal segments.

5.1. Asymmetry in Sentence Correlation

Interestingly, our analysis showed a stronger correlation between visual attention to the "second" translation candidate and the final choice than for the first. We hypothesize that this asymmetry does not reflect a lack of cognitive engagement with the first sentence, but rather a limitation in our current manual gaze-trace processing. Because the second sentence is often the final piece of information processed before a decision is made, the "recency effect" may make its associated gaze data more distinct.

5.2. Challenges in Home-Based Calibration

Our home-based approach offers an alternative to the traditional lab-based setup described in (Bojar

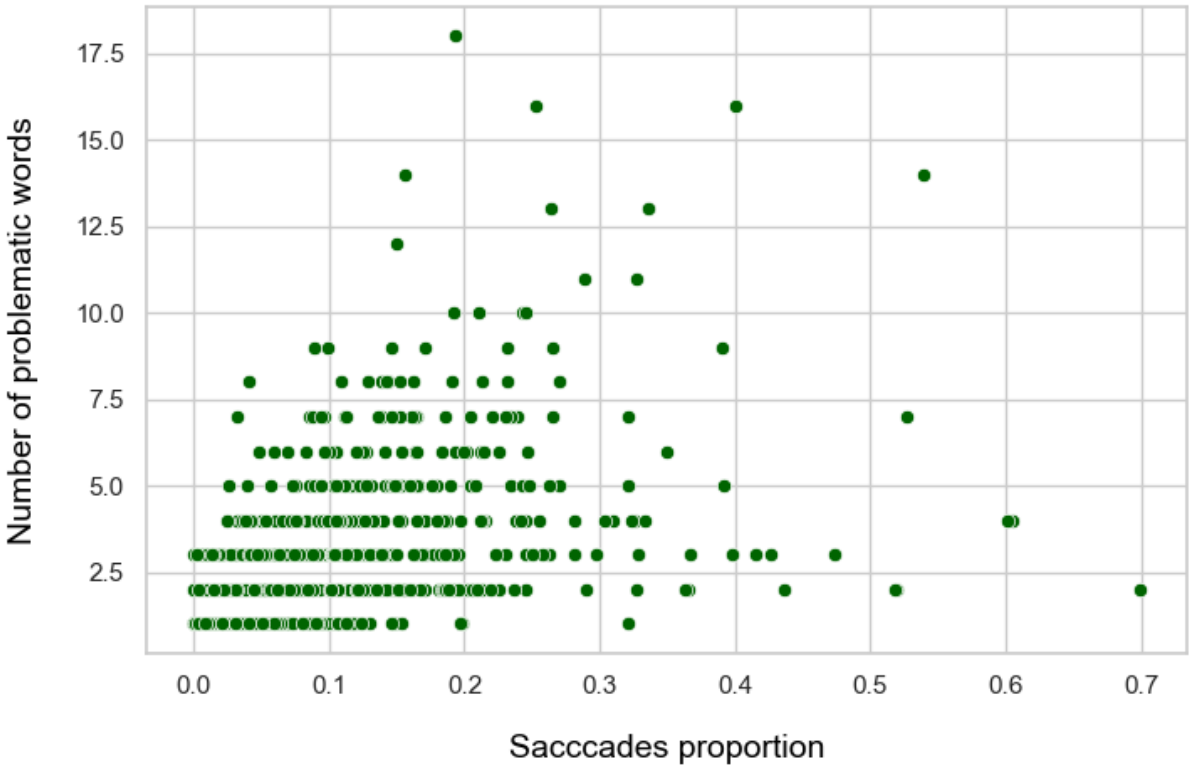


Figure 7: Number of problematic words vs. saccade proportion. Pearson correlation coefficient 0.44.

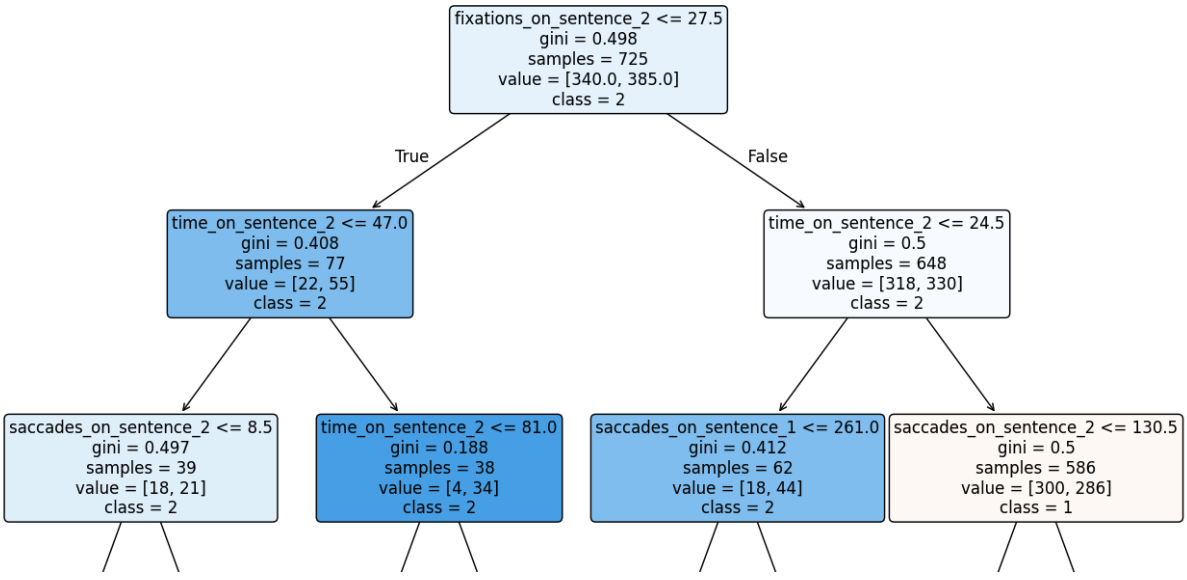


Figure 8: The decision tree analysis (depth=3) showing behavioral thresholds for MT selection.

et al., 2016), which utilized a high-precision EyeLink II tracker (250 Hz) and a chin rest to minimize noise. While the lab setup performed calibration before every screen and excluded data based on pupil size and blinks (removing 30 ms around each blink), our setup relied on an iPhone 12 mini (90 Hz) with calibration at session starts and breaks. Regarding trial exclusion, the lab study manually adjusted areas of interest due to non-linear distortions, while we excluded only 0.6% of screens due to technical misclicks and applied a constant y-axis shift to correct for calibration drift caused by the lack of head restraints. While the iPhone 12 mini provided sufficient precision for broad sentence-level analysis, the lack of head-restraints introduced “noise” during manual processing. This highlights a critical trade-

off: home-based environments offer higher accessibility but require more robust, automated post-processing scripts to handle natural head movements and slight calibration drifts.

6. Conclusion

While this study serves as a preliminary proof-of-concept, our findings suggest that eye tracking provides a 'process-oriented' layer that complements the 'black box' of traditional direct assessment. Unlike error span annotation (Kocmi et al., 2024), which only identifies the location of a flaw, gaze metrics - specifically saccade proportions - reveal the re-scanning behavior, when annotators double-check the source sentence when they encounter a translation that is hard to follow. This allows us to observe cognitive nuances, such as semantic processing effort, that traditional automatic metrics like BLEU or COMET inherently overlook. Furthermore, while the current requirement for manual alignment remains a technical bottleneck, the effort is justified by the potential to use gaze data as a quality control layer; this enables researchers to filter out 'inattentive' trials where participants may have skimmed the text without full cognitive engagement—a critical need for remote, home-based annotation.

6.1. Limitations and Future Work

While this study serves as a proof-of-concept for the technical viability of mobile-based tracking, we acknowledge that our participant pool was limited to a convenience sample of eight volunteers. This initial trial focused on demonstrating the workflow and technical feasibility rather than providing a large-scale demographic analysis.

Future research will focus on:

- **Scaling**: Expanding to a larger, more diverse group to validate the applicability of these metrics.
- **Linguistic Diversity and Cross-Family Pairs**: Our current study focused exclusively on an English–Russian language pair. Future iterations should expand to non-Indo-European languages, such as logographic systems (e.g., Chinese) or right-to-left scripts (e.g., Arabic). Investigating these diverse language pairs will help determine if the cognitive metrics identified here remain robust across different orthographies and reading directions.
- **Hardware**: Exploring higher-frequency sensors that could improve temporal resolution.

7. Acknowledgements

We would like to express our gratitude to Professor Krzysztof Krejtz for his invaluable comments and guidance throughout this study. His expertise in eye tracking research was instrumental in shaping our approach and ensuring the rigor of our analysis.

The work on this project was supported by the grant CZ.02.01.01/00/23_020/0008518 ("Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím").

8. Bibliographical References

Ondřej Bojar, Filip Děchterenko, and Maria Zelenina. 2016. A pilot eye-tracking study of wmt-style ranking evaluation. pages 20–26.

Stephen Doherty, Sharon O'Brien, and Michael Carl. 2010. Eye tracking as an automatic mt evaluation technique. *Machine Translation*, pages 1–13.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. pages 33–41.

M. A. Just and P. A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, pages 329–354.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. pages 1440–1453, Miami, Florida, USA.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-kiu Lo, Vilém Zouhar,

Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Dattatray Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the wmt25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 414–461, Suzhou, China. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. pages 165–172.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Maja Popović. 2020. Informative manual evaluation of machine translation output. pages 5059–5069.

Yanina Prystauka, Gerry T. M. Altmann, and Jason Rothman. 2023. Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavior Research Methods*, 56:3504—-3522.

Hassan Sajjad, Francisco Guzmán, Nadir Durrani, Ahmed Abdelali, Houda Bouamor, Irina Temnikova, , and Stephan Vogel. 2016. Eyes don't lie: Predicting machine translation quality using eye movement. pages 1082–1088.

Salvucci, Dario D., and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. pages 71—-78.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python.

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye tracking as a tool for machine translation error analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1121–1126, Istanbul, Turkey. European Language Resources Association (ELRA).

Myrte Vos, Serge Minor, and Gillian Catriona Ramchand. 2022. Comparing infrared and webcam eye tracking in the visual world paradigm. *FGlossa Psycholinguistics*, 1.

Onur Özsoy, Büsra Çiçek, Zeynep Özal, Natalia Gagarina, and Irina A. Sekerina. 2023. Turkish-german heritage speakers' predictive use of case: webcam-based vs. in-lab eye-tracking. *Frontiers in Psychology*, 14.

Cross-Linguistic Analysis of Eye Movement Patterns: Insights from the First Arabic Eye-Tracking Corpus for NLP

Ibtehal Baazeem[1], Hend Al-Khalifa[2], Abdulmalik Al-Salman[2]

[1]Artificial Intelligence and Robotics Institute, [2]College of Computer and Information Sciences
[1]King Abdulaziz City for Science and Technology, Riyadh 13523, Saudi Arabia
[2] King Saud University, Riyadh 11543, Saudi Arabia
ibaazeem@kacst.gov.sa
{hendk, salman}@ksu.edu.sa

Abstract

Eye-tracking corpora have become valuable resources for understanding human reading behavior and developing cognitively-informed NLP models. However, existing resources predominantly focus on left-to-right Latin script languages, leaving a significant gap for morphologically rich, right-to-left languages like Arabic. This paper presents a cross-linguistic analysis of eye movement patterns using the AraEyebility corpus, the first Arabic eye-tracking corpus comprising 57,617 words read by 15 native speakers. We systematically compare gaze metrics across Arabic and established English corpora. Our analysis identifies distinct patterns in fixation and regression durations, reflecting the unique orthographic characteristics of Arabic: cursive script, diacritization, bidirectional reading (text right-to-left, numbers left-to-right), and morphological complexity. The findings indicate that Arabic readers exhibit relatively longer mean fixation and regression durations than English readers, suggesting higher cognitive processing demands. We discuss implications for developing cognitively-aligned NLP models and provide recommendations for future multilingual eye-tracking research. The AraEyebility corpus is publicly available to support Arabic NLP research.

Keywords: eye-tracking, Arabic NLP, cross-linguistic analysis

1. Introduction

Eye-tracking technology has emerged as a powerful tool for investigating the cognitive processes underlying human reading. By capturing real-time gaze patterns, researchers can examine how readers process text at both word and sentence levels, providing insights that complement traditional linguistic analysis (Rayner, 1998). This connection between eye movements and cognitive processing, formalized in Just and Carpenter's (1980) eye-mind hypothesis, has motivated the development of eye-tracking corpora that serve as valuable resources for natural language processing (NLP) research.

Several landmark eye-tracking corpora have been established for left-to-right Latin script languages. The Dundee Corpus (Kennedy et al., 2003) contains eye movement data from English and French newspaper reading. The GECO corpus (Cop et al., 2017) provides bilingual English-Dutch reading data. The Provo Corpus (Luke and Christianson, 2017) and ZuCo (Hollenstein et al., 2018, 2020) offer English reading data with predictability norms and combined EEG signals. Additional resources exist for German, Portuguese, Chinese, and Danish, enabling cross-linguistic investigations of reading behavior.

However, a notable gap exists for Arabic, a morphologically rich language with unique orthographic properties that distinguish it from previously studied languages. Arabic is written in a cursive, right-to-left script; it uses diacritical marks (tashkeel) to indicate vowels; exhibits context-dependent letter shapes; and processes numbers left-to-right within right-to-left text. These characteristics suggest that Arabic reading may involve distinct cognitive demands that merit dedicated investigation.

This paper addresses this gap by presenting a cross-linguistic analysis using the AraEyebility corpus, the first comprehensive Arabic eye-tracking resource for NLP. We systematically compare eye movement patterns across Arabic and multiple other languages, examining how script-specific features influence reading behavior. Our contributions include: (1) the first systematic cross-linguistic comparison involving Arabic eye movement data; (2) quantitative analysis of how Arabic's orthographic properties affect gaze patterns; and (3) implications for developing cognitively-aligned Arabic NLP models.

The remainder of this paper is organized as follows. Section 1 introduces the study. Section 2 provides background on Arabic orthography and reading. Sections 3 and 4 review related eye-tracking corpora and present the AraEyebility corpus, respectively. Section 5 presents the cross-linguistic analysis and results. Section 6 discusses the findings and their implications for Arabic NLP and eye-tracking research, and Section 7 concludes the paper and outlines directions for future research.

2. Background: Arabic Orthography and Reading

2.1 Unique Properties of Arabic Script

Arabic is a Semitic language with a distinct writing system and linguistic structure. It is commonly

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 10–15
May 12, 2026.

classified into three categories: Classical Arabic (CA), Modern Standard Arabic (MSA), and dialects. CA includes the Holy Qur'an and early classical texts, while MSA, derived from CA, is used in contemporary formal writing such as books, newspapers, and digital media. In contrast, dialects are primarily spoken and vary across regions (El-Haj et al., 2015).

Arabic orthography exhibits several properties that fundamentally differentiate it from Latin-based writing systems and have direct implications for reading behavior. First, Arabic is written from right to left and consists of 28 consonantal letters that change shape depending on their position within a word (initial, medial, final, or isolated). The script is inherently cursive, requiring letters to connect within words and resulting in continuous visual word forms that differ substantially from non-cursive scripts (AlJassmi et al., 2021; Paterson et al., 2015). Additionally, many Arabic letters are distinguished solely by the presence and placement of dots, increasing visual similarity across letter forms and adding further demands on fine-grained visual processing (Paterson et al., 2015).

A third important feature is diacritization. Arabic diacritics (harakat) are supplementary marks placed above or below letters to indicate short vowels and other phonetic information. While these marks support disambiguation and pronunciation accuracy, they may also introduce additional visual complexity, potentially increasing fixation durations (Hermena et al., 2015). In practice, MSA texts are typically partially diacritized or undiacritized, requiring readers to rely on contextual and morphological cues for accurate interpretation.

The morphological richness of Arabic further compounds processing demands. Arabic employs a root-and-pattern system in which most words are derived from three-consonant roots combined with different morphological patterns. Unlike concatenative morphology in languages such as English, this structure distributes semantic and grammatical information across the word, requiring readers to integrate information from multiple letter positions during lexical access.

Finally, Arabic text exhibits bidirectionality when numbers are embedded within text. While Arabic words are read from right to left, numbers are processed from left to right. This shift in directionality within the same line introduces additional cognitive processing demands and has been associated with increased reading complexity and occasional inversion errors (Blanken et al., 1997).

Taken together, these orthographic and linguistic properties suggest that, compared to English, Arabic reading involves distinct visual, linguistic, and cognitive processes, which are expected to be reflected in eye movement behavior.

2.2 Eye Movement Characteristics in Arabic Reading

Previous research has identified several ways in which Arabic reading differs from Latin language reading. The perceptual span, the region from which useful information is extracted during a fixation, extends asymmetrically to the left in Arabic (the direction of upcoming text), contrasting with the rightward asymmetry in left-to-right languages (AlJassmi et al., 2021). Studies suggest that optimal viewing position in Arabic words tends toward the center, unlike the beginning-center position typical for English words, possibly reflecting morphological structure where root information is distributed across the word.

Arabic's informational density creates additional processing demands. Research indicates that Arabic reading is more time-intensive than Latin language reading, with word identification presenting greater challenges (AlJassmi et al., 2021). The impact of word frequency on skipping rates appears less pronounced in Arabic compared to English, suggesting different utilization of lexical information for reading decisions (AlJassmi et al., 2021). Furthermore, the bidirectional nature of Arabic text, where numbers are read left-to-right within otherwise right-to-left text, can introduce processing complications and occasional inversion errors (Blanken et al., 1997).

3. Related Eye-Tracking Corpora

Eye-tracking corpora have been developed across multiple languages, each contributing to our understanding of reading behavior. The Dundee Corpus (Kennedy et al., 2003) pioneered naturalistic eye-tracking data collection, recording 10 native speakers reading English newspaper texts (56, 212 words) and 10 speakers reading French texts (52,173 words). This corpus enabled investigation of parafoveal-on-foveal effects and established benchmarks for eye movement research.

The GECO corpus (Cop et al., 2017) expanded bilingual eye-tracking research by recording monolingual and bilingual readers navigating an English novel (54,364 words) and its Dutch translation (59,716 words). GECO provides 46 pre-extracted gaze features and has become a standard resource for computational modeling of reading. The Provo Corpus (Luke and Christianson, 2017) focused on predictability effects, collecting data from 84 native English speakers reading brief paragraphs with associated cloze task norms.

The ZuCo corpora (Hollenstein et al., 2018, 2020) are multimodal resources that combine eye-tracking with EEG data, enabling the investigation

Corpus	Language	Words	Participants	Script	Direction
AraEyebility	Arabic	57,617	15	Cursive	RTL
Dundee (EN)	English	56,212	10	Non-cursive	LTR
Dundee (FR)	French	52,173	10	Non-cursive	LTR
GECO (EN)	English	54,364	14	Non-cursive	LTR
GECO (NL)	Dutch	59,716	19	Non-cursive	LTR
ZuCo 1.0	English	21,629	12	Non-cursive	LTR
Provo	English	2,689	84	Non-cursive	LTR

Table 1: Comparison of eye-tracking corpora across languages, where RTL denotes right-to-left scripts and LTR denotes left-to-right scripts.

of both behavioral and neural correlates of reading. Table 1 presents a comparison of eye-tracking corpora across languages. Additional eye-tracking corpora exist for Portuguese (Leal et al., 2018, 2022), Chinese (Zhang et al., 2022), Danish (Hollenstein et al., 2022), German, and Japanese, thus providing cross-linguistic perspectives on reading behavior.

Despite this progress, right-to-left scripts remain significantly underrepresented in eye-tracking research. Prior Arabic studies on eye-tracking have primarily focused on specific linguistic phenomena, such as diacritization effects (Hermena et al., 2015), morphological processing (Khateb et al., 2013), or reading difficulties in special populations (Al-Wabil and Al-Sheaha, 2010). No comprehensive Arabic eye-tracking corpus for NLP applications existed until the development of AraEyebility.

4. The AraEyebility Corpus

4.1 Corpus Design and Data Collection

The AraEyebility corpus[1] (Baazeem et al., 2025) was developed to address the absence of Arabic eye-tracking resources for NLP research. The corpus comprises eye movement data collected from 15 native Arabic speakers (7 male, 8 female; ages 20-45) reading 587 paragraphs totaling 57,617 words. Participants were healthy adults with normal or corrected-to-normal vision, holding or pursuing degrees from Arab countries, and representing diverse professional backgrounds and Arabic-speaking regions to ensure representative data collection.

Texts were drawn from Arabic books covering 13 topics, including grammar, literature, health, politics, geography, and biography.The corpus includes both MSA texts from contemporary sources and CA texts from historical works, spanning authors from the 8th to the 21st centuries. Texts were partially diacritized following consultation with linguists, balancing disambiguation benefits against visual noise concerns. Each text was segmented into coherent paragraphs expressing single ideas, enabling paragraph-level analysis that balances contextual richness with experimental traceability.

Eye movements were recorded using a Tobii X120 eye-tracker operating at 120 Hz with 0.5-degree precision. Participants read silently at their own pace while seated approximately 60-65 cm from a 1920x1080 monitor. Texts were displayed in black traditional Arabic font (size 18) on a white background with appropriate line spacing. Each session included five-point calibration procedures, and recordings with gaze sample quality below 80% were repeated. The final dataset achieved an average gaze sample quality of 93%.

4.2 Extracted Features

The corpus includes 98 features categorized into text-based features (69), capturing linguistic properties and gaze-based features (29), capturing eye movement metrics. Gaze features encompass standard reading metrics, including time to first fixation, first fixation duration, single fixation duration, total fixation duration, fixation count, saccade metrics, regression measures, visit duration, and pupil measurements. Text-based features include character counts, word counts, syllable metrics, sentence statistics, readability scores, and Arabic-specific measures such as diacritization density and morphological complexity indicators.

5. Cross-Linguistic Analysis

5.1 Methodology

To examine cross-linguistic differences in reading behavior, we compared key eye movement metrics from the AraEyebility corpus with reported values from well-established English corpora.

We focused on metrics that are consistently reported across corpora and that reflect the fundamental aspects of reading: fixation duration (first fixation, single fixation, and total fixation) and regression duration. Where possible, we also examined reading time distributions and their relationship to text complexity.

Direct statistical comparison across corpora requires caution due to substantial differences in language, writing systems, experimental design,

[1] AraEyebility is publicly available at https://doi.org/10.7910/DVN/P5WPNS under a CC BY-NC 4.0 license.

text materials, participant populations, annotation conventions, and available metrics. While we acknowledge that adding inferential statistical tests would strengthen a strictly matched comparison, such analyses are not methodologically reliable in the current study because the comparison is based on corpus-level summary patterns rather than harmonized participant-level data under matched conditions. Accordingly, applying formal cross-corpus statistical tests could be misleading, as the comparison is based on reported plots and summary distributions rather than harmonized raw data. In this context, normalization or sensitivity analyses (e.g., restricting comparisons to matched genres or text lengths) were also not feasible. Therefore, the analysis emphasizes qualitative patterns from these plots rather than precise quantitative comparisons.

5.2 Fixation Duration Patterns

Analysis of the AraEyebility corpus reveals that Arabic readers exhibit longer mean fixation durations compared to English readers in the Dundee and GECO corpora. This pattern aligns with the hypothesis that Arabic's cursive script and morphological complexity impose additional processing demands. The distribution of total fixation duration shows positive skewness, with most readings being relatively brief but with a notable tail of longer reading times, particularly for morphologically complex or low-frequency words. The extended fixation durations in Arabic reading may reflect several factors: the need to process diacritical marks when present; the cognitive demands of letter-form identification given context-dependent shapes; the integration of visual information from a cursive script where word boundaries are less distinct; and the lexical access processes specific to root-and-pattern morphology. These findings are consistent with prior research suggesting that Arabic's informational density makes reading more time-intensive than for Latin languages (AlJassmi et al., 2021).

5.3 Regression Patterns

Regression patterns show that Arabic readers exhibit longer backward eye movements compared to English readers. Regressions in reading typically indicate comprehension difficulties, ambiguity resolution, or verification processes.

The elevated regression duration in Arabic reading may stem from lexical ambiguity in undiacritized text, where readers must sometimes revisit words to confirm their interpretation based on subsequent context. Additionally, the morphological richness of Arabic means that a single orthographic form can correspond to multiple morphological analyses. Readers may engage in more extensive reanalysis processes when initial parsing proves inconsistent with subsequent material. The bidirectional nature of Arabic text (with embedded left-to-right numbers) may also contribute to regression patterns, as readers navigate between different reading directions within the same line.

5.4 Reading Time Distributions

Examination of reading time distributions in AraEyebility reveals patterns consistent with those in other eye-tracking corpora, yet with Arabic-specific characteristics. First fixation duration and single fixation duration follow approximate normal distributions, whereas total visit duration and total fixation duration exhibit pronounced positive skewness. This pattern, also observed in GECO and ZuCo, reflects a mixture of rapid reading of familiar content and extended processing of challenging material.

Correlation analysis between eye movement metrics and text readability levels (Easy, Medium, Difficult) confirms that more complex texts elicit longer fixation durations and more visits. This relationship validates the corpus annotation and demonstrates that gaze patterns meaningfully reflect text processing difficulty. The correlation between participant-assigned readability levels and Open Source Metric for Measuring Arabic Narratives (OSMAN) (El-Haj and Rayson, 2016) readability scores further supports the reliability of the cognitive annotations.

6. Discussion

6.1 Implications for Arabic NLP

The cross-linguistic differences observed in this analysis suggest potential implications for Arabic NLP. First, the observed variation in processing times and rereading patterns may reflect differences in how textual information is processed across languages. The reported differences in fixation durations and regression behavior can be interpreted in light of established Arabic properties, including right-to-left reading, cursive connectivity, context-dependent letter forms, optional diacritics, root-and-pattern morphology, and bidirectional processing with embedded numbers. As such, models developed for English may not transfer directly to Arabic without considering language-specific characteristics, potentially motivating adjustments to feature representations or model design.

Second, the findings support the value of cognitively-informed approaches to Arabic NLP. Eye movement data can serve as training signals or evaluation criteria for models designed to predict text difficulty, generate simplified text, or assess text quality. The correlation between gaze patterns and readability levels in AraEyebility demonstrates that human processing difficulty is measurable and can inform computational models.

Third, the analysis highlights the importance of script-specific considerations in multilingual NLP.

The distinctive properties of Arabic script, including cursive writing, diacritization, and bidirectionality, create processing demands not found in Latin-script languages. Models aiming for cross-linguistic generalization must account for these fundamental differences in how readers process text.

6.2 Implications for Eye-Tracking Research

Our analysis also contributes to eye-tracking methodology. The development of AraEyebility demonstrates that comprehensive eye-tracking corpora can be constructed for right-to-left scripts, despite the technical challenges involved. The corpus design decisions, including paragraph-level segmentation, partial diacritization, and multi-genre text selection, provide a template for future eye-tracking corpus development in underrepresented languages.

The cross-linguistic patterns identified here suggest that theoretical models of reading developed primarily from English data may need to be revised to accommodate the full range of human writing systems. Arabic represents just one of many scripts that differ fundamentally from the Latin alphabet; similar investigations of other writing systems (Hebrew, Persian, Urdu, and various Indic scripts) would further advance our understanding of reading universals and specificities.

6.3 Limitations

Several limitations should be acknowledged. First, the cross-linguistic comparison relies on heterogeneous corpora that differ in language, writing system, participant characteristics, stimuli, methods, and measures. Consequently, the analysis is based on corpus-level patterns rather than matched participant-level data, which limits the validity of direct statistical testing and precludes strong causal or generalizable claims. Second, the AraEyebility corpus, while substantial, has a limited participant pool (15 readers) and exhibits class imbalance across readability levels, which may affect generalizability. Third, the corpus focuses on MSA and CA, dialectal Arabic, which millions of speakers use daily, is not represented.

7. Conclusion

This paper presents the first systematic cross-linguistic analysis of eye movement patterns in Arabic reading data. Using the AraEyebility corpus, we have demonstrated that Arabic reading exhibits distinctive characteristics, including longer fixation durations, elevated regression frequencies, and different optimal viewing positions, reflecting the unique cognitive demands of processing Arabic script. These findings contribute to both the theoretical understanding of reading across writing systems and the practical development of cognitively-informed Arabic NLP.

The AraEyebility corpus addresses a significant gap in eye-tracking resources and opens new avenues for Arabic NLP research. Future work should expand the corpus to include additional participants and dialectal Arabic, develop computational models that leverage gaze patterns for Arabic text processing, and extend cross-linguistic investigations to other underrepresented writing systems. As NLP increasingly addresses the world's linguistic diversity, cognitively-grounded resources like AraEyebility will be essential for developing models that reflect how humans actually process language.

8. Bibliographical References

AlJassmi, M. A., Hermena, E. W., and Paterson, k. B. (2021). Eye movements in Arabic reading. *Experimental Arabic Linguistics,* 10:85–108.

Al-Wabil, A. and Al-Sheaha, M. (2010). Towards an interactive screening program for developmental dyslexia: Eye movement analysis in reading Arabic texts. In *Proceedings of the 12th International Conference on Computers Helping People with Special Needs*, pages 25–32, Vienna, Austria, July 14–16. Springer.

Baazeem, I., Al-Khalifa, H., and Al-Salman, A. (2025). AraEyebility: Eye-tracking data for Arabic text readability. *Computation*, 13(5):108.

Blanken, G., Dorn, M., and Sinn, H. (1997). Inversion errors in Arabic number reading: Is there a nonsemantic route? *Brain and Cognition*, 34(3):404–423.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

El-Haj, M., Kruschwitz, U., and Fox, C. (2015). Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation*, 49(3):549–580.

El-Haj, M. and Rayson, P. (2016). OSMAN—A novel Arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255, Portorož, Slovenia. May 23–28. European Language Resources Association (ELRA).

Hermena, E. W., Drieghe, D., Hellmuth, S., and Liversedge, S. P. (2015). Processing of Arabic diacritical marks: Phonological–syntactic disambiguation of homographic verbs and visual crowding effects. *Journal of Experimental Psychology: Human Perception and Performance*, 41(2):494.

Hollenstein, N., Barrett, M., and Björnsdóttir, M. (2022). The Copenhagen corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language*

Resources and Evaluation Conference, pages 1712–1720, Marseille, France, June 20–25. European Language Resources Association (ELRA).

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1):180291.

Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. (2020). ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. May 11–16. European Language Resources Association (ELRA).

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Kennedy, A., Hill, R., and Pynte, J. (2003). The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movements*, Dundee, UK.

Khateb, A., Taha, H. Y., Elias, I., and Ibrahim, R. (2013). The effect of the internal orthographic connectivity of written Arabic words on the process of the visual recognition: A comparison between skilled and dyslexic readers. *Writing Systems Research*, 5(2):214–233.

Leal, S. E., Duran, M. S., and Aluísio, S. (2018). A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413, Santa Fe, New Mexico, USA, August 20–26. Association for Computational Linguistics.

Leal, S. E., Lukasova, K., Carthery-Goulart, M. T., and Aluísio, S. M. (2022). RastrOS project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese. *Language Resources and Evaluation*, 56(4):1333–1372.

Luke, S. G. and Christianson, K. (2017). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Paterson, K. B., Almabruk, A. A. A., McGowan, V. A., White, S. J., and Jordan, T. R. (2015). Effects of word length on eye movement control: The evidence from Arabic. *Psychonomic Bulletin & Review*, 22(5):1443–1450.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Zhang, G., Yao, P., Ma, G., Wang, J., Zhou, J., Huang, L., Xu, P., Chen, L., Chen, S., Gu, J., Wei, W., Cheng, X., Hua, H., Liu, P., Lou, Y., Shen, W., Bao, Y., Liu, J., Lin, N., and Li, X. (2022). The database of eye-movement measures on words in Chinese reading. *Scientific Data*, 9(1):411.

Exploring Cognitively Informed Sentence Simplification with Gaze-Guided Text Generation

Andreas Säuberli[1,2] **Diego Frassinelli**[1] **Barbara Plank**[1,2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
{andreas.saeuberli, diego.frassinelli, b.plank}@lmu.de

Abstract

Automatic text simplification has mostly relied on human judgments when it comes to what is considered easy or difficult to read. Eye movements while reading can offer a more direct and objective signal of processing effort and reading ease. In this paper, we explore gaze-guided text generation (GGTG), an approach to control reading ease in generated texts, and assess its use for sentence simplification. GGTG employs a gaze model that is trained to predict eye-tracking measures such as reading times or regression rates, which are then used to rerank next-token probabilities generated by a language model. We evaluated the approach on an English sentence simplification benchmark and found gains in automatic evaluation metrics, although the simplification operations are mostly limited to the lexical level. Its modular nature also allows GGTG to be combined with other simplification techniques such as prompting or fine-tuning.

Keywords: text simplification, eye tracking, cognitive modeling, controlled text generation

1. Introduction

Most research in automatic text simplification has relied on human judgments of simplicity or manual reference simplifications that are often crowdsourced (Coster and Kauchak, 2011; Paetzold and Specia, 2017; Alva-Manchego et al., 2020; Grabar and Saggion, 2022). This means that the human intuition of what *should* be considered easy to read is taken as a proxy of what is *actually* easy to read. However, recent research has shown that there can be a substantial discrepancy between subjective perception of difficulty and actual comprehension, and that both aspects can vary between different user groups (Alonzo et al., 2021; Carrer et al., 2024). In contrast, methods such as eye tracking can provide a more direct and objective signal of the processing effort required to read and understand a text (Just and Carpenter, 1980). While eye-movement data has been used to *evaluate* simplified texts, using this cognitive signal directly to *generate* simplifications remains underexplored. At the same time, the increasing amount of available eye-tracking-while-reading data, including multilingual corpora (Siegelman et al., 2022; Jakobi et al., 2025) and datasets involving diverse readers such as non-native or dyslexic readers (Kuperman et al., 2023; Siegelman et al., 2025; Hollenstein et al., 2022; Reich et al., 2024) makes it more feasible now to use such data for natural language processing (NLP) applications like text simplification.

Recent work by Säuberli et al. (2026) proposed **gaze-guided text generation** (GGTG) as a simple yet effective way to integrate eye-tracking data into the text generation process. It works by training a gaze model to predict eye-tracking measures and using these predictions to modify the next-token probabilities generated by a language model (LM). Their findings suggest that the level of control that can be achieved with this approach may be limited to shallow features affecting lexical processing, such as word length and frequency. In this paper, we explore to what extent GGTG can be used to simplify sentences, and whether the method can be pushed to induce more complex simplification operations at the syntactic level.

We build on and extend the experiments in Säuberli et al. (2026) in several ways:

- We explore five eye-tracking measures associated with different levels of processing and their ability to capture text complexity.
- We train gaze models that ignore word length and frequency and focus on higher-level aspects of text complexity.
- We evaluate GGTG on the ASSET benchmark for sentence simplification.

2. Related work

Controllable text simplification has been approached from several angles. Most prominently, models have been trained with special tokens or feature vectors to control characteristics like word frequency, dependency tree depth, and readability level (Scarton and Specia, 2018; Martin et al., 2020, 2022; Agrawal and Carpuat, 2023). Nishihara et al. (2019) used a lexical constraint loss to control lexical complexity. Kew and Ebling (2022) is the most similar to our approach. Instead of using eye-tracking data, they trained classifiers that predict the level of difficulty for next token candidates and modified the token probabilities accordingly.

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 16–23
May 12, 2026.

While cognitive data such as eye movements have been used to *evaluate* simplified texts or predicting readability (Rello et al., 2013; Singh et al., 2016; Vajjala et al., 2016; Ivchenko and Grabar, 2024; Gruteke Klein et al., 2025a,b), the present work is, to the best of our knowledge, the first to use gaze data directly in the simplification process.

3. Methods

3.1. Gaze-guided text generation (GGTG)

At its core, GGTG involves an ensemble of an off-the-shelf **language model** and a **gaze model**, which we train to predict word-level eye-tracking measures (e.g., first fixation time or regression rate). The LM predicts candidates for the next token, which are then re-ranked by the gaze model. The strength of the influence of the gaze model can be controlled via a **gaze weight**. This way, the LM output can be steered towards eliciting specific reading behaviors (e.g., longer/shorter fixation times or higher/lower regression rates), thereby manipulating reading ease.

We applied beam search with a beam size of 8 to decode simplified texts. We used the implementation by Säuberli et al. (2026) and refer to the corresponding paper for more details.

3.2. Language model and prompts

We chose the instruction-tuned Llama 3.2 model (3B; Grattafiori et al., 2024), as it is a small and efficient model with strong instruction-following performance. For the simplification task, we experiment with two different prompts. The first prompt instructs the LM to *paraphrase* the source sentence without changing its meaning, allowing us to assess the simplifying effect of the gaze model alone. The second prompt instructs the LM to *simplify*, to test whether GGTG still has a simplifying effect. See Appendix A for the precise wording.

3.3. Gaze models

3.3.1. Predicted eye-tracking measures

We selected five word-level eye-tracking measures that can plausibly be predicted from preceding context only (which is a requirement for autoregressive generation):

- **First fixation duration**: the duration of the first fixation on a word during the first pass (0 if the word is skipped in the first pass).
- **First-pass reading time**: the sum of all fixation durations on a word during the first pass (0 if the word is skipped in the first pass).
- **Go-past time**: the sum of all fixation durations from when the word was first fixated until the gaze moves past the word for the first time. This includes regression paths initiated on the word during first-pass reading.
- **First-pass skipping rate**: 1 if the word was skipped in the first pass, 0 otherwise.[1]
- **First-pass regression rate**: 1 if there was a regression during the first pass, 0 otherwise.

All measures are computed for readers individually and then averaged across readers for each word. The final measures are normalized to have a mean of 0 and a standard deviation of 1.

According to psycholinguistic research, some of these measures are associated with earlier cognitive processes such as word recognition, while others reflect later processing such as syntactic integration (Rayner, 1998; Godfroid, 2019). For example, first fixation duration is measured when the word is first looked at, while go-past time also includes time spent *after* the first fixation. Therefore, we expect earlier measures to mainly enable lexical simplification, while later measures may allow more syntactic simplification.

We fine-tuned the large variant of GPT-2 (774M; Radford et al., 2019) to predict the eye-tracking measures listed in Section 3.3.1. We trained separate models for each eye-tracking measure.

3.3.2. Model training

All models are trained on a mix of four publicly available eye-tracking corpora of naturalistic reading that cover a range of genres and difficulty levels, listed in Table 1. All of these datasets contain texts that span multiple sentences, so we perform automatic sentence splitting after calculating eye-tracking measures and train the gaze model on individual sentences.

We trained the models on 90% of each dataset and used the remaining 10% as a validation set for early stopping and to measure performance. We ensured that all sentences from the same text are assigned to the same data split to avoid data leakage. The remaining training procedure follows Säuberli et al. (2026). Performance on the validation set is reported in Table 2.

3.3.3. Residual models

Säuberli et al. (2026) found that their gaze model's reading time predictions were dominated by shal-

[1] First-pass skipping rate is the only measure for which higher values are associated with better reading ease. Therefore, for ease of interpretation, we flip the sign of the skipping rates, so that lower numbers can be considered better across all measures.

Dataset	Text genre/content	# words	# readers
EMTeC (Bolliger et al., 2025)	LLM-generated; various genres	50,871	(*)107
OneStopQA (Berzak et al., 2025)	Original and simplified news	35,164	(*)360
MECO-L1 English (Siegelman et al., 2022)	Encyclopedic information	2,107	46
Provo (Luke and Christianson, 2017)	Various genres	2,743	84

Table 1: Datasets used for training the gaze models. (*) means not every text is read by every reader.

Eye-tracking measure	GPT-2	LR	LR + GPT-2 residual
First fixation duration	**0.593**	0.546	0.587
First-pass reading time	**0.636**	0.596	0.633
Go-past time	**0.481**	0.367	0.458
First-pass skipping rate	0.610	0.562	**0.619**
First-pass regression rate	**0.226**	0.140	**0.226**

Table 2: Explained variance (R^2) for each gaze model, averaged across the four validation datasets. LR = linear regression.

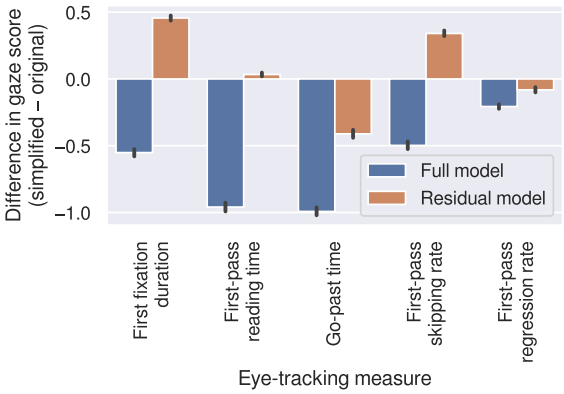


Figure 1: Mean gaze score differences between original and simplified reference texts in the ASSET validation set. A negative value means that the gaze model predicts lower eye-tracking measures for the simplified version (which is the expectation). Error bars show 95% confidence intervals.

low features such as word length and frequency. To avoid this, we trained a second version of each model that does *not* capture the variance associated with these features. We did this by first fitting a linear regression model to predict the eye-tracking measures based word length and word frequency alone.[2] Next, we computed the residuals of the linear regression model on the training data, normalized them to mean 0 and standard deviation 1, and trained the GPT-2-based gaze model to predict these residuals.

3.4. Evaluation

We evaluated our approach on ASSET (Alva-Manchego et al., 2020), an established sentence simplification benchmark for English with multiple crowdsourced references. We report results on the validation set, which consists of 2,000 sentences with ten human reference simplifications each. We consider gaze weights in the range from 0 (gaze model deactivated) to −3 (decrease eye-tracking measure).

We report the reference-based evaluation metrics SARI (Xu et al., 2016) and LENS (Maddela et al., 2023) against all ten references, as well as the reference-free metric LENS-SALSA (Heineman et al., 2023). We also report the effects on lexical and syntactic features such as word frequency and dependency tree depth.[3]

[2]Word length was calculated as the number of characters. Word frequency is measured on the Zipf scale based on the *wordfreq* package (Speer, 2022). Linear regression models were fitted with *scikit-learn* (Pedregosa et al., 2011).

[3]Dependency trees were generated using Stanza (Qi et al., 2020).

4. Results and discussion

4.1. Do the gaze models capture sentence complexity?

The GGTG approach can only work for sentence simplification if the gaze model is able to discriminate simple from complex sentences. While the gaze models described in Section 3.3 are indirectly trained to capture complexity by predicting eye-tracking measures, this does not necessarily translate into a useful model of complexity in general. Moreover, it is unclear which eye-tracking measures are suitable for sentence simplification.

Therefore, as a first step, we assess whether the gaze scores predicted by our models differ between original and simplified sentences in ASSET. These differences are visualized in Figure 1. For the full models trained to predict eye-tracking measures, we observe lower gaze scores in the simplified versions on average, with the largest differences for first-pass reading time and go-past time. In contrast, the models that were trained on the linear regression residuals consistently predict smaller differences, or even differences in the opposite direction. This is expected, as these models do not have access to some of the most salient predictors

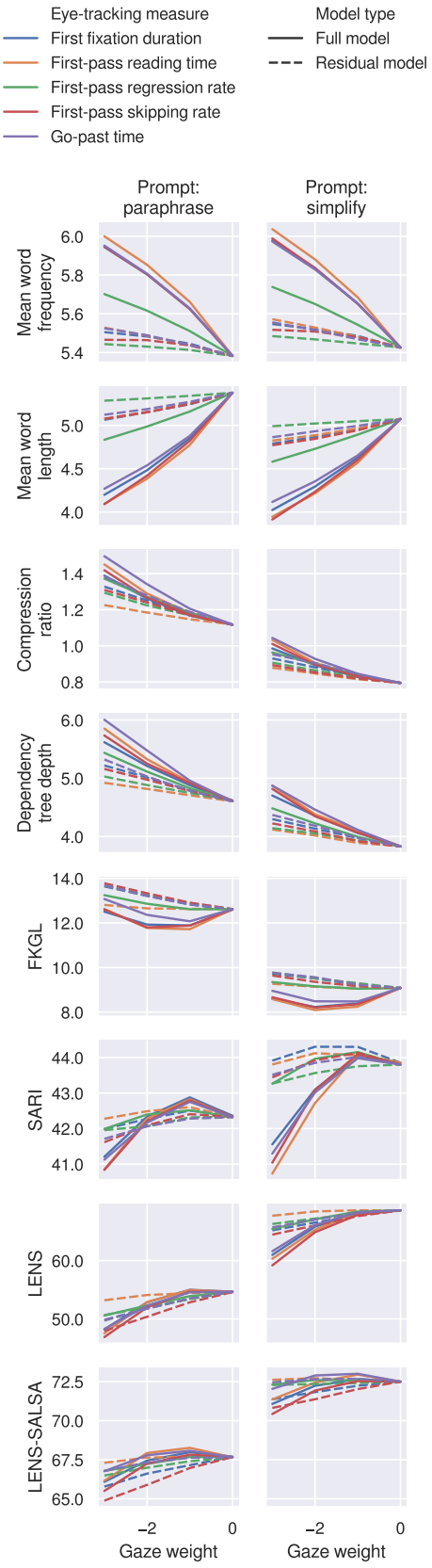


Figure 2: Effect of gaze models on generated texts and evaluation metrics. A gaze weight of 0 indicates generation without a gaze model, a negative gaze weight means steering the language model to decrease the corresponding eye-tracking measure.

of simplicity – word length and frequency. However, go-past time and first-pass regression rate – both associated with later cognitive processing – are still predicted to be lower in the simplified texts. This suggests that these two residual models capture some aspects of simplicity in the ASSET dataset that go beyond word length and frequency, possibly features at the syntactic or semantic level.

4.2. How does GGTG affect the generated texts?

Figure 2 shows how changing the gaze weight for the different gaze models affects the output texts, comparing full vs residual models (see Section 3.3.3) and the two prompt types (see Section **??**). As expected, word length and frequency (plots in first two rows) are strongly affected by all full models, but less so by the residual models. Compression ratio increases with stronger gaze weights, indicating that output sentences tend to become longer. This is due to the fact that during beam search, appending tokens that decrease the overall gaze score is preferred over stopping the generation (which would mean that the gaze score remains unchanged). Dependency trees also tend to increase in depth, likely due to the increased sentence length. This suggests that syntactic nesting is not reduced in the simplified sentences.

Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) estimates a text's readability using the number of sentences, words, and syllables as surface-level proxies. We observe (plots in fifth row) that almost all full models decrease FKGL (i.e., improve readability) with gaze weights -1 and -2, but this effect is negated by the increase in sentence length around gaze weight -3. SARI and LENS-SALSA slightly improve with gaze weight -1, but quickly decrease with lower weights.

Overall, there is no clear pattern regarding the different eye-tracking measures. First-pass reading time appears to have the most consistent positive effect on evaluation metrics, which may be explained by this gaze model's relatively strong performance (see Table 2). Residual models generally have weaker effects, but in the case of first-fixation duration and first-pass reading time, improvements in SARI can still be observed. As expected, the *simplify* prompt yields better readability and evaluation metrics than the *paraphrase* prompt, but even here, GGTG can further improve evaluation metrics. See Table 3 for example outputs.

5. Conclusion

The appeal of using eye-tracking data for text simplification is that, as a cognitive signal, it reflects reading ease more directly than human judgments

Version	Text
Source	A Georgian inscription around the drum attests his name.
LM-only	An inscription on the drum confirms his name.
Reading time -1	There is an inscription on the drum with his name on it.
Reading time -2	The name of the person is written on a drum.
Regression rate -1	There is an inscription on the drum that confirms his name.
Regression rate -2	The name of the person is mentioned in an inscription on a drum.

Table 3: Example outputs with the *simplify* prompt. Reading time -1 indicates that the gaze model predicts first-pass reading time and a gaze weight of -1 is used.

or manually simplified texts. Our results show that some simplification operations can be achieved by applying GGTG, and that small improvements in automatic evaluation metrics can be achieved, even if the LM is already explicitly prompted to simplify.

However, more complex operations beyond the lexical surface level remain a challenge, even for the residual models, which are trained to focus on less superficial features. A reason for this challenge could be the training data for the gaze models. There is a growing amount of available eye-tracking data, but extracting more subtle effects and patterns from naturalistic reading corpora is difficult. In contrast, psycholinguistic research commonly uses minimal pairs to measure these effects. Leveraging these resources could also be helpful for text simplification.

In sum, while GGTG can help at least at a superficial, lexical level, applying it on its own is not yet sufficient in reality. However, thanks to the modularity of the approach, it is easily possible to combine it with other techniques, including prompting and fine-tuning.

Limitations

Automatic evaluation. Automatic evaluation metrics can only measure the adequacy and difficulty of simplified texts to a very limited degree, and human evaluation is usually recommended (Alva-Manchego et al., 2021; Grabar and Saggion, 2022; Carrer et al., 2024). Since our work is exploratory and the number of investigated parameters would have made a comprehensive human evaluation unfeasible, we decided to prioritize automatic evaluation metrics.

English only. Our evaluation is limited to English texts, limiting the generalizability of our results to other languages. The main reason for this is the scarcity of eye-tracking data in other languages.

Number of tested models. Finally, we only considered a single base model for both the language and gaze model, limiting generalizability to other models. While we initially experimented with several base models, the results were not substantially different, so our results only include one set of relatively small models for simplicity and reproducibility.

Ethical considerations

Trustworthiness of model outputs. Given the use of large language models and the nature of our approach, there is little control over the content and semantic accuracy of the generated texts. Therefore, our method should not be used without additional safeguards and manual inspection or post-editing. This is particularly important in accessibility scenarios with potentially vulnerable end users, which is among the most common use cases of text simplification.

Reproducibility. All datasets and code libraries used in this project are open-source and received due citations. The code and data for reproducing the results and figures in this paper is available from the accompanying repository: https://github.com/mainlp/gaze-guided-sentence-simplification/

Use of generative models. We used GitHub Copilot to accelerate programming tasks. All generated code was thoroughly checked and tested. We did not use generative models for ideation, results interpretation, or paper writing.

Acknowledgements

We thank the three anonymous reviewers for their valuable feedback. This research is in parts supported by the ERC Consolidator Grant DIALECT 101043235.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819. Association for Computational Linguistics.

Oliver Alonzo, Jessica Trussell, Becca Dingman, and Matt Huenerfauth. 2021. Comparison of methods for evaluating complexity of simplified texts among deaf and hard-of-hearing adults at different literacy levels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–12. ACM.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. OneStop: A 360-participant English eye tracking dataset with different reading regimes. *Scientific Data*, 12(1).

Lena S. Bolliger, Patrick Haller, Isabelle C. R. Cretton, David R. Reich, Tannon Kew, and Lena A. Jäger. 2025. EMTeC: A corpus of eye movements on machine-generated texts. *Behavior Research Methods*, 57(7).

Luisa Carrer, Andreas Säuberli, Martin Kappus, and Sarah Ebling. 2024. Towards holistic human evaluation of automatic text simplification. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 71–80, Torino, Italia. ELRA and ICCL.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Aline Godfroid. 2019. *Eye Tracking in Second Language Acquisition and Bilingualism: A Research Synthesis and Methodological Guide*. Routledge.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, et al. 2024. The Llama 3 herd of models. *arXiv*.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025a. Surprisal takes it all: Eye tracking based cognitive evaluation of text readability measures. *arXiv*.

Keren Gruteke Klein, Omer Shubi, Shachar Frenkel, and Yevgeni Berzak. 2025b. The effect of text simplification on reading fluency and reading comprehension in L1 English speakers. *OSF*.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495. Association for Computational Linguistics.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The copenhagen corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Oksana Ivchenko and Natalia Grabar. 2024. Study of medical text reading and comprehension through eye-tracking fixations. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 84–92, Torino, Italia. ELRA and ICCL.

Deborah Noemie Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusic Vilus, Stefan Frank, Marie-Luise Müller, Kristine M Jensen de López, Nik Kharlamov, Hanne B. Søndergaard Knudsen, Yevgeni Berzak, Ella Lion,

Irina A. Sekerina, Cengiz Acarturk, Mohd Faizan Ansari, Katarzyna Harezlak, Pawel Kasprowski, Ana Bautista, et al. 2025. MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, ETRA '25, pages 1–11. ACM.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Tannon Kew and Sarah Ebling. 2022. Target-level sentence simplification as controlled paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Technical report.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Kaidi Lõo, Marco Marelli, Kelly Nisbet, et al. 2023. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, 45(1):3–37.

Steven G. Luke and Kiel Christianson. 2017. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2017. Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40, Valencia, Spain. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

David R. Reich, Shuwen Deng, Marina Björnsdóttir, Lena Jäger, and Nora Hollenstein. 2024. Reading does not equal reading: Comparing, simulating and exploiting reading behavior across populations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13586–13594, Torino, Italia. ELRA and ICCL.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? Text simplification strategies for people with dyslexia.

In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 1–10. ACM.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718. Association for Computational Linguistics.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, Marco Marelli, et al. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-Movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

Noam Siegelman, Sascha Schroeder, Yaqian Borogjoon Bao, Cengiz Acartürk, Niket Agrawal, Lena S. Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, Sofya Goldina, Romualdo Ibáñez Orellana, Lena A. Jäger, Ómar I. Jóhannesson, Anurag Khare, Nik Kharlamov, Hanne B. S. Knudsen, Árni Kristjánsson, Charlotte E. Lee, Jun Ren Lee, et al. 2025. Wave 2 of the Multilingual Eye-Movement Corpus (meco): New text reading data across languages. *Scientific Data*, 12(1).

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Robyn Speer. 2022. wordfreq. Zenodo.

Andreas Säuberli, Darja Jepifanova, Diego Frassinelli, and Barbara Plank. 2026. Controlling reading ease with gaze-guided text generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2383–2397. Association for Computational Linguistics.

Sowmya Vajjala, Detmar Meurers, Alexander Eitel, and Katharina Scheiter. 2016. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 38–48, Osaka, Japan. The COLING 2016 Organizing Committee.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

A. Prompts

***paraphrase* prompt:**

```
Paraphrase the following text.  You
may change the wording and structure
of the text, but not its meaning.
Only respond with the paraphrased
text.  Here is the text:  [...]
```

***simplify* prompt:**

```
Simplify the following text.  You
may change the wording and structure
of the text, but not its meaning.
Only respond with the simplified
text.  Here is the text:
```

Impact of Text Simplification on Eye-Tracking-Based Reading Profiles Across Domains

Oksana Ivchenko Natalia Grabar
CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
oksana.ivchenko.etu@univ-lille.fr, natalia.grabar@univ-lille.fr

Abstract

Understanding how text readability affects reading behaviour is crucial for improving accessibility and health communication. We analyse sentence-level eye-tracking data from the French Eye-TrAcking (FETA) corpus, which includes original and manually simplified texts from three domains: general, medical, and clinical. Using clustering of fixation-based features, we identify recurrent processing patterns and examine how these patterns change under text simplification. Cluster quality is evaluated using silhouette scores and participant-level bootstrap stability. Simplification does not uniformly reduce reading effort but reorganises processing in domain-dependent ways. Medical texts show strong diversification, general texts moderate diversification, and clinical texts show a reduction in the number of distinct reading profiles. Hence, rather than uniformly facilitating reading, simplification redistributes effort across sentences, underscoring the need for domain-sensitive readability approaches.

Keywords: Readability, Eye-tracking, Text Simplification, Clustering, Reading profiles

1. Introduction

Understanding how text complexity affects reading behaviour is essential for improving accessibility and health communication. This issue is particularly important in specialised domains such as medicine, where texts are often dense and conceptually demanding (Eklics et al., 2024; Brown, 2008). Text simplification is commonly used to improve readability (François and Lefer, 2022), yet its effects on reading dynamics are not always uniform and may vary across domains. Identifying which textual elements drive reading difficulty is crucial for developing more effective simplifications (Cheng Sheang et al., 2022).

Eye-tracking provides an objective way to study reading processes by capturing fixation durations, regressions, and other indicators of cognitive effort (Salvucci and Goldberg, 2000; Duchowski, 2007; Radach and Kennedy, 2013). These measures reveal how readers process complex sentences and where difficulties arise. Reading effort is not homogeneous across sentences, and simplification may reorganise rather than uniformly reduce processing demands. Unsupervised methods, including clustering approaches, have been applied to eye-tracking data to identify patterns of reading behaviour and cognitive processing (Kucharský et al., 2020; Göbel and Martin, 2018; Guan et al., 2025), enabling the discovery of latent processing profiles without predefined categories.

We analyse sentence-level eye-tracking data from original and manually simplified texts in three domains–general, medical, and clinical–to identify recurrent processing profiles. Using clustering of fixation-based features, we examine how these profiles vary across domains and how they change under simplification. This approach allows us to characterise different configurations of reading effort and to assess whether simplification leads to more homogeneous or more diverse processing patterns.

This study addresses the following questions: (i) Can sentence-level eye-tracking features identify distinct processing profiles? (ii) Do these profiles vary across text domains? (iii) How does simplification reshape these profiles?

2. Method

Dataset. Experiments are conducted on the FETA (French Eye-TrAcking) corpus (Ivchenko and Grabar, 2025), a French eye-tracking dataset combining *original* and manually *simplified* texts from three domains: *general* Wikipedia articles, *medical* Wikipedia articles, and *clinical* case reports (toxicology and gastroenterology), thus enabling an analysis of reading behaviour across increasing conceptual complexity. The corpus contains gaze recordings from 46 native French participants (32 women, 14 men; age 18–43 years, $M = 23.3$, $SD = 6.7$) collected with a Tobii Pro Spectrum eye tracker. All participants had normal or corrected-to-normal vision, and no medical training. Texts are distributed across two counterbalanced sets, such that each participant reads only one version (original or simplified) of each text.

Simplification modifies text structure by increasing the number of sentences through syntactic segmentation. By domain, sentence counts increased from 32 to 42 in clinical texts (+31.3%), from 73 to 107 in general texts (+46.6%), and from 144 to 179 in medical texts (+24.3%). As a result, simplified texts contain more, shorter sentences than their original counterparts.

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 24–29

Eye-tracking data representation. We perform clustering on sentence-level observations derived from eye-tracking recordings, where word-level measures are aggregated into sentence representations by averaging across words within each sentence. Each observation corresponds to a *participant–sentence* pair, where sentence boundaries follow the experimental segmentation. For each observation, we use a vector of standardized eye-tracking features (z-scores) capturing early and late reading processes (Cook and Wei, 2019; Vasishth et al., 2013). The following set of five non-redundant features was selected for clustering:

- **First-pass first fixation duration** (ms): reflects early lexical processing effort upon first encountering a word;
- **Average duration of fixations** (ms): reflects overall processing effort per word;
- **First-pass regression** (proportion): indicates whether the reader made a backward eye movement from a word during its first encounter, reflecting early integration difficulty;
- **Number of fixations**: represents how many times the reader's gaze landed on the word or sentence (AOI);
- **Re-reading duration** (ms): measures the time spent revisiting a text region after the initial reading pass, calculated as the difference between regression-path duration and first-pass duration.

Observations are analysed separately for each condition defined by `text type` $\in$ {clinical, medical, general} and `version` $\in$ {original, simplified}.

Outlier removal. Eye-tracking measures are known to be heavy-tailed and sensitive to occasional tracking artifacts. To reduce the impact of extreme values on clustering solutions, we remove outlier observations within each condition. Within each `text type` × `version` condition, features are z-standardised and sentence-level outliers (|z| > 3 on any selected feature) are removed.

Clustering. We cluster observations with k-means (Lloyd's algorithm) using Euclidean distance in the standardized feature space. For each condition, we fit k-means with `n_init`=20 random initializations and select the solution with minimal within-cluster sum of squares. We report the silhouette score (Rousseeuw, 1987) as an internal validation measure. This metric compares how close each observation is to points within its assigned cluster relative to points in the nearest alternative cluster. Higher values indicate better separation and cohesion. In our work, k is fixed per condition based on stability and interpretability considerations.

Stability under participant bootstrap. To ensure that cluster structure is not driven by a small number of participants, we evaluate robustness using participant-level bootstrap resampling. For each condition and chosen k, we first fit k-means on the full dataset and obtain cluster centroids. We then perform $B = 30$ bootstrap iterations, sampling participants with replacement and refitting k-means on the aggregated sentence-level observations. Because cluster labels are arbitrary across runs, bootstrap centroids are aligned to the full-data solution via minimum-cost matching (Hungarian algorithm; Kuhn, 1955). Stability is defined as the mean Euclidean distance between matched centroids, averaged across bootstrap samples. Lower values indicate more stable clustering. Results are reported in table 1.

Text	Version	k	Silhouette	Stability	n
Clinical	Orig.	3	0.215	0.538	638
	Simpl.	2	0.272	0.168	1023
Medical	Orig.	2	0.247	0.194	3372
	Simpl.	5	0.251	0.203	4864
General	Orig.	2	0.256	0.214	1672
	Simpl.	3	0.219	0.202	2242

Table 1: Clustering diagnostics after outlier removal (k: number of clusters; higher silhouette scores and lower centroid distances indicate better clustering quality).

Hierarchical clustering validation. To assess the robustness of the K-means clustering solutions, we performed a cross-validation using agglomerative hierarchical clustering (Ward, 1963) with Ward's linkage and Euclidean distance. Unlike K-means, hierarchical clustering does not rely on random initialization and therefore provides an independent view of the underlying structure.

For each condition (text type $\times$ version), we fit hierarchical clustering with the same number of clusters k selected for the K-means analysis. Agreement between the two algorithms was quantified using the Adjusted Rand Index (ARI), which measures the similarity between two partitions while correcting for chance. The index ranges from -1 to 1, where 1 indicates identical partitions and 0 indicates agreement expected by chance.

The comparison revealed moderate-to-high agreement between K-means and hierarchical clustering across most conditions (Figure 1). For the medical and general texts, ARI values ranged from 0.45 to 0.58, indicating that the identified reading profiles are largely algorithm-independent. Lower

agreement was observed for the clinical simplified condition (ARI ≈ 0.25), suggesting that while distinct reading strategies exist, cluster boundaries in this condition are less sharply defined.

Overall, the consistency between two independent clustering approaches supports the structural validity of the reported reading profiles and indicates that the main low-effort vs. high-effort distinction is not specific to a single algorithm.

However, it should be noted that simplification in the corpus often involves sentence splitting or merging, which prevents a strict one-to-one alignment between original and simplified sentences. Consequently, rather than performing direct pairwise comparisons, the present approach adopts a condition-level perspective and analyses distributions of sentence-level processing patterns within each condition. As such, the results reflect global differences in reading behaviour across conditions rather than direct sentence-level effects of simplification.

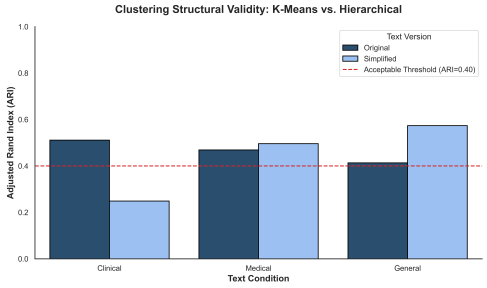


Figure 1: Agreement between k-means and hierarchical clustering across conditions, showing moderate-to-high consistency between solutions.

3. Interpretation of cluster profiles

General Texts. In **original** general texts (Figure 2), sentences cluster into two processing profiles, mainly distinguished by early lexical processing:

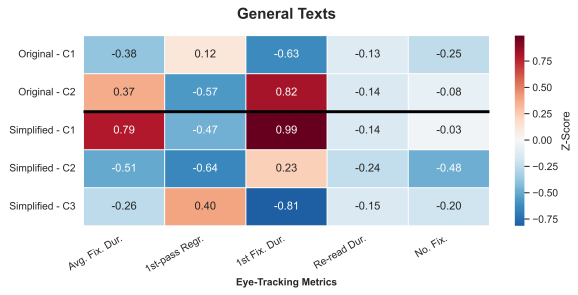


Figure 2: Cluster centroids for general texts (z-scored eye-tracking features).

- **C1.** Shorter fixation durations, reduced first-fixation duration, minimal re-reading, and fewer fixations overall, with slightly elevated early regressions. This profile reflects relatively fast processing with occasional brief corrective regressions.
- **C2.** Longer early fixation durations and higher average fixation durations, combined with few regressions and limited re-reading. This pattern indicates deeper first-pass lexical processing followed by stable linear reading.

In the **simplified** general texts, three distinct profiles emerge, indicating diversification rather than homogenisation of sentence-level processing:

- **C1.** Longer fixation durations and elevated first-fixation durations, with low regression and re-reading rates. This profile reflects deeper but stable first-pass processing.
- **C2.** Shorter fixation durations, reduced regressions, minimal re-reading, and fewer fixations overall, corresponding to fluent processing.
- **C3.** Relatively short fixation durations with increased first-pass regressions and limited re-reading, suggesting fast processing with occasional corrective regressions.

Overall, simplification does not uniformly reduce processing effort but increases the diversity of sentence-level processing patterns, enabling fluent processing for some sentences while others remain more effortful.

Medical Texts. In the **original** medical texts (Figure 3), sentences cluster into two primary processing profiles:

- **C1.** Elevated average and first-fixation durations combined with reduced regression rates. This pattern reflects substantial first-pass lexical integration with stable linear processing.
- **C2.** Shorter fixation durations and lower first-fixation durations, with slightly increased first-pass regressions. This profile corresponds to relatively fluent processing with occasional corrective eye movements.

In the **simplified** medical texts, five distinct sentence-level profiles emerge, indicating strong diversification of processing dynamics:

- **C1.** Elevated fixation durations and high first-fixation durations with low regression and re-reading rates, reflecting intensive but stable first-pass processing.
- **C2.** Very low first-fixation duration and reduced fixation counts, indicating highly fluent early lexical access.

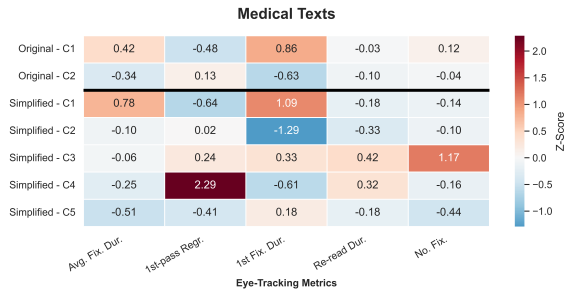


Figure 3: Cluster centroids for medical texts (z-scored eye-tracking features).

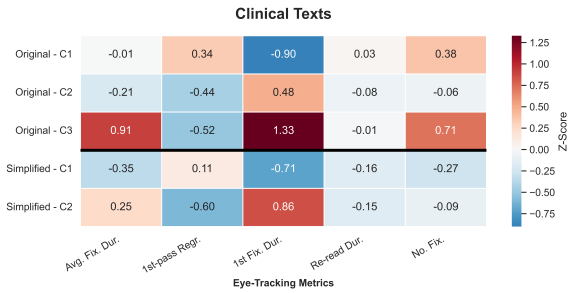


Figure 4: Cluster centroids for clinical texts (z-scored eye-tracking features).

- **C3.** Increased regressions, elevated re-reading duration, and high fixation counts, suggesting reanalysis and repeated inspection.
- **C4.** Extremely high early regressions with short fixation durations, indicating rapid initial processing followed by substantial corrective eye movements.
- **C5.** Globally reduced fixation durations and low regression rates, corresponding to fluent and stable processing.

Taken together, whereas original medical texts exhibit two main processing modes, simplification leads to marked diversification of sentence-level patterns. Rather than uniformly reducing effort, it redistributes processing across distinct profiles.

Clinical Texts. In **original** clinical texts (Figure 4), we observe three sentence-level reading profiles:

- **C1.** Sentences associated with short first-fixation durations and elevated regression rates, combined with increased fixation counts. This pattern suggests rapid initial processing followed by corrective eye movements and additional inspection.
- **C2.** Sentences characterised by longer first-fixation durations but reduced regression activity, indicating careful early lexical integration followed by stable linear reading.
- **C3.** Sentences showing elevated average fixation duration, strongly increased first-fixation duration, and higher fixation counts, reflecting substantial lexical integration effort during the first pass without extensive reanalysis.

In the **simplified** clinical texts, two processing profiles remain:

- **C1.** Sentences associated with reduced fixation durations, lower fixation counts, and minimal re-reading, indicating globally fluent processing.
- **C2.** Sentences characterised by elevated first-fixation durations and slightly increased average fixation duration but reduced regression rates, suggesting deeper initial lexical processing that remains stable and linear.

Overall, while original clinical texts exhibit multiple distinct processing regimes, simplification appears to reduce heterogeneity and promote more stable processing dynamics across sentences.

3.1. Cross-domain comparison of processing profiles

Sentence-level clustering reveals that the impact of simplification on reading dynamics varies across text domains.

In the **general texts**, the number of clusters increases from two in the original condition to three in the simplified condition, while within-cluster variance decreases (2.08 to 1.55). This indicates moderate diversification of processing profiles accompanied by greater internal consistency.

In the **medical texts**, the effect is considerably stronger: clusters increase from two to five, and within-cluster variance drops markedly (2.23 to 1.55). This suggests that simplification does not uniformly reduce processing effort but redistributes it across multiple internally coherent regimes.

In contrast, the **clinical texts** exhibit the opposite pattern. The original texts yield three clusters, whereas the simplified texts yield two, with a slight increase in within-cluster variance (1.83 to 2.02). This indicates a reduction in distinct processing configurations and more homogeneous reading dynamics across sentences.

This pattern suggests that simplification reorganises sentence-level processing in a domain-dependent manner: medical texts show strong diversification, general texts moderate diversification, and clinical texts relative homogenisation, reflecting how readers balance early decoding and later integration processes across domains. Detailed cluster proportions and variance values are reported in Table 2.

Text	Version	k	Avg. Var.	Cluster Sizes (%)	Total Inertia
Clinical	Original	3	1.83	35.3 / 41.8 / 22.9	1166.91
	Simplified	2	2.02	52.0 / 48.0	2066.74
Medical	Original	2	2.23	46.9 / 53.1	7508.05
	Simplified	5	1.55	22.3 / 24.8 / 12.1 / 10.1 / 30.8	7524.70
General	Original	2	2.08	54.1 / 45.9	3475.72
	Simplified	3	1.55	29.4 / 36.3 / 34.3	3481.84

Table 2: Detailed clustering diagnostics per condition.

4. Conclusion

We investigated whether sentence-level eye-tracking behaviour can be grouped into interpretable processing profiles across text domains and levels of simplification. Clustering analyses revealed stable and meaningful patterns distinguished by fixation duration, regressions, and overall processing effort.

Simplification did not produce a single homogeneous processing pattern. Instead, it reorganised reading dynamics in domain-dependent ways. Medical texts showed strong diversification of processing profiles, general texts showed a moderate diversification, and clinical texts showed a relative homogenisation. These findings suggest that simplification redistributes processing demands across sentences rather than uniformly reducing effort.

This work provides a basis for future research on predicting reading effort from textual features and on text simplification. Thus, identifying sentence-level processing profiles can help detect passages that remain effortful even after simplification and guide targeted revisions. Such profiles could also inform predictive models that estimate reading effort directly from text, enabling adaptive simplification and accessibility-oriented writing tools.

Acknowledgement

This work was partially funded by the French National Research Agency (ANR) through the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01, and by the French State under the France-2030 programme and the Initiative of Excellence of the University of Lille, which are acknowledged for the funding and support granted to the R-CDP-25-002-PRIME-NEXT-GEN.

We thank the reviewers for their helpful comments and questions, which improved the overall quality of the paper. We also thank Jamal Abdul Nasir (University of Galway) for valuable assistance with the methodological analysis.

5. Bibliographical References

Jo Brown. 2008. How clinical communication has become a core part of medical education in the UK. *Medical Education*, 42(3):271–278.

Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, and Horacio Saggion. 2022. Identification of complex words and passages in medical documents in French. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 116–125, Avignon, France. ATALA.

Anne E. Cook and Wei Wei. 2019. What can eye movements tell us about higher level comprehension? *Vision*, 3(3).

Andrew Duchowski. 2007. *Eye Tracking Methodology: Theory and Practice*. Computer Science. Springer London.

K. Eklics, A. Csongor, A. Hambuch, and J.-D. Fekete. 2024. Diverse integration of simulated patients in medical education for communication, language, and clinical skills in Hungary. *Advances in Medical Education and Practice*, 15:301–312.

Thomas François and Marie-Aude Lefer. 2022. Revisiting simplification in corpus-based translation studies: Insights from readability research. *Meta*, 67(1):50–70.

Zheng-Hong Guan, Sunny S. J. Lin, and Liwei Hsu. 2025. Profiling readers in multiple-text reading: Affective engagement, metacognition, and mediation effects. *2025 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 338–342.

Fabian Göbel and Henry Martin. 2018. Unsupervised clustering of eye tracking data. In *Spatial Big Data and Machine Learning in GIScience*, pages 25–28. Spatial Big Data. Workshop at the 10th International Conference on Geographic Information Science (GIScience 2018).

Šimon Kucharský, Ingmar Visser, George Ovidiu Truţescu, Patrick G Laurence, Maria Zaharieva, and Maartje EJ Raijmakers. 2020. Cognitive

strategies revealed by clustering eye movement transitions. *Journal of Eye Movement Research*, 13(1).

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Ralph Radach and Alan Kennedy. 2013. Eye movements in reading: Some theoretical context. *Quarterly journal of Experimental Psychology (2006)*, 66.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, page 71–78, New York, NY, USA. Association for Computing Machinery.

Shravan Vasishth, Titus von der Malsburg, and Felix Engelmann. 2013. What eye movements can tell us about sentence comprehension. *WIREs Cognitive Science*, 4(2):125–134.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

6. Language Resource References

Oksana Ivchenko and Natalia Grabar. 2025. A French eye-tracking corpus of original and simplified medical, clinical, and general texts - FETA. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 37–43, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Parts of Speech Shape Reading-Time Variability in Brazilian Portuguese

Diego Alves
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de

Abstract

This study uses regression analysis of Brazilian Portuguese eye-tracking data to examine variability in reading times across grammatical categories. Mixed-effects models reveal distinct patterns: numerals elicit high individual variability in early-stage reading, while function words (e.g., adpositions, determiners) drive differences in late-stage integration. In contrast, nouns show stable effects. These findings demonstrate that individual differences in reading are systematically linked to specific parts of speech, with numerals and function words as key loci of variability.

Keywords: reading time, Portuguese, parts-of-speech

1. Introduction

Theoretical models of reading have largely been built to explain systematic effects of linguistic input on eye movements, emphasising factors such as lexical frequency, predictability, and structural complexity.

Recent work have demonstrated that variation in cognitive resources, reading skill, and linguistic experience systematically modulates reading behaviour (e.g., Haeuser and Kray, 2024; Staub, 2021). These findings motivate models in which eye movements reflect not only properties of the text, but also enduring characteristics of individual readers.

Empirical investigations of individual variability in reading have typically focused on how readers differ in their sensitivity to specific psycholinguistic phenomena, such as lexical frequency, word predictability, syntactic complexity, ambiguity resolution, and integration costs (e.g., Kuperman et al., 2018; Nicenboim et al., 2016; Staub, 2021). Variability is usually assessed in relation to experimental manipulations or item-level properties, rather than to the grammatical units over which these effects are realised. As a result, it remains largely unexplored whether individual differences in reading behaviour are distributed uniformly across grammatical categories, or whether certain parts of speech are more prone to variability than others.

This paper investigates the role of part-of-speech (PoS) categories in shaping reading behaviour in Brazilian Portuguese. Using eye-tracking data, we examine how different PoS influence multiple stages of lexical processing, as reflected in first fixation duration, first-run dwell time, and total dwell time. To assess these effects, we analysed inter-participant variability using linear mixed-effects models applied to word-level eye-tracking measures, allowing us to estimate both average PoS effects and individual differences in sensitivity to grammatical category. By comparing these measures across readers, we aim to identify which PoS categories are most strongly associated with individual variability in reading times, thereby providing insight into the linguistic factors that drive individual differences in sentence processing.

2. Related Work

Individual variability in reading behaviour has become an important topic in psycholinguistic research, particularly in eye-tracking studies of sentence processing.

Previous work has shown that readers differ reliably in their sensitivity to specific perceptual and linguistic factors, including word frequency, predictability, visual contrast, and font difficulty, indicating that individual differences in eye movements reflect stable properties of the reading system rather than measurement noise (Staub, 2021). Other studies have linked variability in eye movements to reader characteristics, including age, demonstrating systematic differences in the use of contextual information during reading (Haeuser and Kray, 2024).

In parallel, psycholinguistic research has increasingly drawn on computational models of prediction, including surprisal-based approaches and large language models, to explain reading times across languages (Wilcox et al., 2023; Xu et al., 2023). Although this work has provided strong evidence for the role of predictability in sentence processing, analyses typically focus on average effects and offer limited insight into how predictive processing varies between individuals.

Beyond reader-level factors, eye-tracking studies have also shown that gaze patterns are sensitive to

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 30–34

grammatical structure. In particular, eye-movement features have been shown to reliably distinguish between major parts of speech, such as nouns and verbs, suggesting that grammatical categories are associated with distinct processing profiles (Barrett and Søgaard, 2015). However, this work has largely focused on classification and representation, rather than on whether different parts of speech differentially contribute to inter-participant variability in reading time.

Finally, most research on individual differences in reading has been conducted in English. Comparatively little psycholinguistic work has examined how inter-participant variability is distributed across grammatical categories in other languages, including Brazilian Portuguese, which differs from English in both syntactic and morphological structure. Recent eye-tracking resources for Brazilian Portuguese make it possible to address this gap (Sardinha, 2010). In addition, recent work has shown that parts of speech differ systematically in their reading-time profiles as a function of information content, as estimated by surprisal from large language models (Alves, 2025). However, it remains unclear whether such PoS-specific processing differences are also associated with differences in inter-participant variability during reading.

3. Methodology

3.1. Eye-tracking Data

The RastrOS corpus was created to support psycholinguistic research in Brazilian Portuguese (BP), with a particular emphasis on lexical predictability and sentence processing. It consists of two primary components: predictability norms obtained through a Cloze task and eye-tracking data collected during reading experiments.

The Cloze task was completed by 393 native speakers of BP recruited from six Brazilian universities, most of whom were undergraduate students. Each participant filled in five randomly selected paragraphs, balanced across three text genres: journalistic (40%), literary (20%), and popular science (40%). The dataset is annotated with PoS tags (generated with the Palavras parser; Bick 2000) and word frequency measures derived from Corpus Brasileiro (Sardinha, 2010) and BrWaC (Wagner Filho et al., 2018). In addition, the corpus includes surprisal and entropy-reduction values computed from the Cloze responses.

Eye-tracking data were collected from 37 undergraduate students using an EyeLink 1000 eye-tracker with a sampling rate of 1000 Hz. Participants read the same 120 sentences included in the Cloze corpus (2,494 words; 2,831 tokens including punctuation). Each sentence is annotated with 36 eye-movement measures, including first fixation duration, first-run dwell time, and total dwell time.

For our analysis, we parsed the RastrOS sentences using the Stanza parser (Qi et al., 2020) and assigned each word a Universal Part-of-Speech (UPOS) tag according to the Universal Dependencies guidelines (De Marneffe et al., 2021). This annotation scheme was chosen to facilitate future cross-linguistic comparisons.

As Portuguese has contractions (e.g., *da*, composed of the adposition *de* combined with the determiner *a*, equivalent to “of the” in English), an alignment phase was necessary. For each contraction in the eye-tracking data, we retained the PoS tag of the head of the contraction.

4. Evaluation Method

To examine inter-participant variability in reading behaviour across parts of speech (PoS), we analysed three standard eye-tracking measures: first fixation duration, first-run dwell time, and total dwell time.

1. First fixation duration - the duration of the first fixation on a word during its first pass. Annotated as IA_FIRST_FIXATION_DURATION in RastrOS.

2. First-run dwell time - the sum of all first-pass fixations on a word. IA_FIRST_RUN_DWELL_TIME in RastrOS.

3. Total dwell time - the sum of all fixations on a word during the trial. IA_DWELL_TIME in RastrOS.

First fixation reflects the initial processing of a word and is associated with early stages of lexical access. Gaze duration captures the time spent on a word during first-pass reading and is sensitive to lexical and syntactic processing. Total fixation time includes any regressions back to the word and reflects later stages of comprehension, such as reanalysis or integration difficulties (Rayner, 1998).

All reading-time measures were log-transformed after adding a constant to reduce skewness. Analyses were conducted at the word level.

To avoid unstable coefficient estimates driven by sparse data, only UPOS categories appearing in three or more distinct sentences were retained for analysis; all other categories were excluded prior to model estimation.

For each reading-time measure, we fitted a linear mixed-effects regression model with part of speech (UPOS) as the predictor of interest as shown in Equation 1. The models included established lexical and contextual control variables: word length,

word frequency (Freq, log-transformed), and word surprisal (Srp) estimated using the LLaMA-3.2-3B transformer language model[1] (Alves, 2025). To account for spillover effects and temporal dependencies in eye movements, we additionally included lagged predictors for reading time, word frequency, and surprisal from the one and two preceding words.

Participant was included as a random effect, with both a random intercept and random slopes for UPOS, allowing the effect of part of speech on reading time to vary across individuals. This hierarchical structure captures individual differences in baseline reading speed as well as in sensitivity to grammatical category.

$$\begin{aligned}\log(RT) \sim\ & \text{UPOS} + \text{WordLength} + \text{Freq} + \text{Srp} \\ & + \log(RT_{-1}) + \log(RT_{-2}) \\ & + \text{Freq}_{-1} + \text{Freq}_{-2} \\ & + \text{Srp}_{-1} + \text{Srp}_{-2} \\ & + (1 + \text{UPOS} \mid \text{Participant})\end{aligned} \tag{1}$$

To facilitate comparison of effect magnitudes across predictors, fixed-effect coefficients were fully standardized by refitting the models on variables that had been centred and scaled to unit variance, using the *effectsize* package in R. The resulting coefficients thus represent changes in reading time (in standard deviation units) associated with one standard deviation changes in the predictors.

Inter-individual variability in part-of-speech effects was quantified using the standard deviation of the corresponding random slopes estimated by the mixed-effects model. These random-slope standard deviations provide a model-based estimate of between-participant variability, reflecting individual differences in the strength of UPOS effects after partial pooling and accounting for noise in participant-specific estimates.

This modelling strategy allows us to simultaneously assess (i) the average effect of grammatical category on reading time and (ii) the extent to which these effects systematically vary across readers.

5. Results

Table 1 reports, for each UPOS category, the number of distinct Text–Sentence contexts in which it appears and the number of recording sessions in which it is attested.

As previously described, UPOS categories appearing in fewer than three sentences were not considered in the analysis; consequently, symbols (SYM) and interjections (INTJ) were excluded. All remaining UPOS categories occur in more than 25 distinct sentences and were read at least once by all participants.

UPOS	Sentences	Participants
NOUN	104	37
DET	99	37
VERB	97	37
ADP	91	37
ADJ	71	37
ADV	71	37
PRON	61	37
AUX	60	37
CCONJ	57	37
SCONJ	48	37
PROPN	40	37
NUM	27	37
SYM	2	35
INTJ	1	36

Table 1: Number of unique Text–Sentence pairs in which each UPOS appears, and number of unique recording sessions (participants) in which each UPOS is attested.

Regarding the predictors that are not UPOS in equation 1, as expected, across all three reading-time measures, word length showed a reliable positive effect and lexical frequency a reliable negative effect, indicating longer reading times for longer and less frequent words.

Surprisal also exhibited a consistent positive effect across measures. Autoregressive effects of previous reading times were robustly positive in all models. In contrast, spillover effects of frequency and surprisal varied across measures, with early measures showing facilitation from previous-word frequency, whereas total dwell time showed positive lagged frequency effects and largely absent surprisal spillover.

Figures 1,2, and 3 present the effects of UPOS on the three reading-time measures analysed: first fixation duration, first-run dwell time, and total dwell time, respectively. For each measure, bars show standardized fixed-effect estimates from the linear mixed-effects models, with error bars indicating 95% confidence intervals. Point size represents the standard deviation of participant-specific UPOS effects estimated by the model, reflecting the extent to which the influence of each UPOS category varies across readers.

The figures suggest that grammatical category exerts a stronger influence on total dwell time than on earlier reading-time measures. Eight UPOS categories show statistically reliable positive effects on total dwell time, compared to only three categories for first-fixation duration (two negative and one positive) and four for first-run dwell time (three positive and one negative).

[1] https://huggingface.co/meta-llama/Llama-3.2-3B

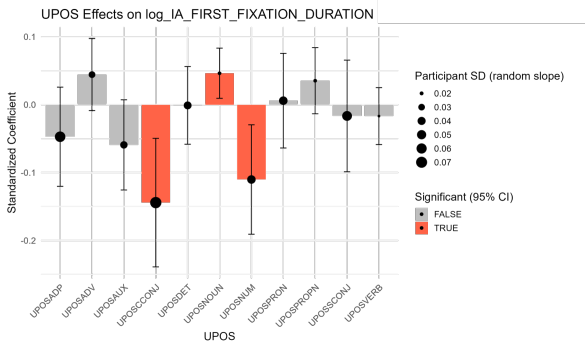


Figure 1: Standardized fixed effects of UPOS on log first fixation duration. Bars show fixed-effect estimates (±95% CI); point size indicates the standard deviation of participant-level random slopes.

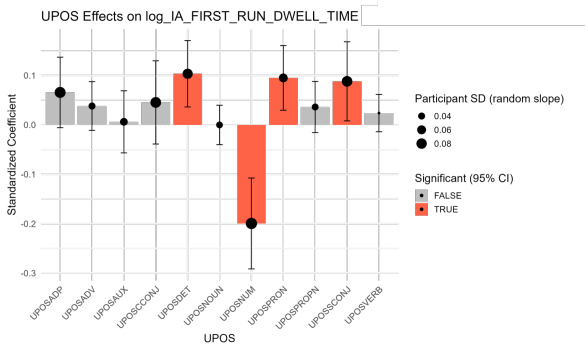


Figure 2: Standardized fixed effects of UPOS on log total dwell time. Bars show fixed-effect estimates (±95% CI); point size indicates the standard deviation of participant-level random slopes.

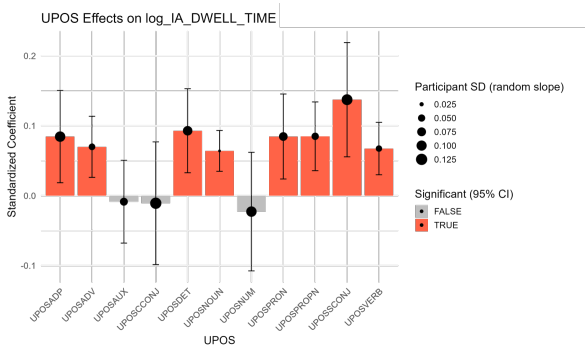


Figure 3: Standardized fixed effects of UPOS on log first-run dwell time. Bars show fixed-effect estimates (±95% CI); point size indicates the standard deviation of participant-level random slopes.

Across measures, the overall magnitude of inter-participant variability is broadly comparable, although it tends to reach higher levels for total dwell time, indicating greater individual differences at later stages of processing.

For first-fixation duration, reliable effects are limited to a small number of categories: numerals (NUM) and coordinating conjunctions (CCONJ) show negative effects accompanied by substantial inter-participant variability, whereas nouns (NOUN) exhibit a positive effect with comparatively low variability across readers.

A similar pattern of high variability is observed for numerals in first-run dwell time, where they again show a negative effect, although this effect does not extend to total dwell time. Additional effects on first-run dwell time are observed for subordinate conjunctions (SCONJ), pronouns (PRON), and determiners (DET), all of which display considerable between-participant variability.

In contrast, total dwell time reveals both a larger number of reliable UPOS effects and clearer differentiation in inter-participant variability. Adpositions (ADP), determiners (DET), and subordinate conjunctions show the greatest variability across readers, whereas nouns again emerge as the most stable category, exhibiting relatively low variability.

With the exception of numerals, the PoS categories showing greater variability are predominantly function words, which are typically short. These words frequently occur at the beginning of phrases or clauses, a position that is also commonly associated with the presence of filler particles. Filled pauses have been shown to occur preferentially at utterance- and phrase-initial positions, suggesting that they are linked to structural planning and boundary marking in speech production and comprehension (Maclay and Osgood, 1959).

Numerals occur in both numerical and written forms and fulfil a range of syntactic functions, including nominal subjects (nsubj), nominal modifiers (nmod), numeric modifiers (nummod), and oblique arguments (obl). Their effects appear to be confined to early stages of processing and show substantial variability across participants, suggesting that a more fine-grained analysis incorporating syntactic function may be informative.

6. Conclusion and Future Work

This paper investigated how part-of-speech (PoS) categories influence reading behaviour and its variability across individuals in Brazilian Portuguese.

Analysing eye-tracking data, we found that grammatical category has a stronger effect on late integration measures (total dwell time) than on early lexical access. Crucially, inter-participant variability was not uniform: function words (e.g., adpositions, determiners) and numerals showed the greatest individual differences, while content words such as nouns were more stable.

These results indicate that readers vary most in their processing of syntactic and structural elements, highlighting the need for models of reading

to account for how individual cognitive systems interact with the grammatical architecture of language.

Future work should extend this approach to other languages and incorporate finer-grained syntactic information, as individual UPOS categories can fulfil distinct syntactic functions that may differentially shape reading behaviour.

7. Limitations

The present study has limitations that should be considered when interpreting the results. First, the eye-tracking experiment was conducted exclusively with undergraduate students, resulting in a relatively homogeneous participant sample. This limits the extent to which the observed inter-individual variability can be generalised to broader populations with greater diversity in age, educational background, and reading experience. Future work including more heterogeneous reader groups may reveal different patterns of variability.

Second, the corpus used in the experiment was limited in terms of textual diversity, drawing from a restricted set of sources. Reading behaviour, and its variability, may differ across registers, genres, or communicative contexts, particularly with respect to the processing of function words and syntactic structure. Expanding the range of text types would therefore be important for assessing the robustness and generality of the observed effects.

8. Bibliographical References

Diego Alves. 2025. Benchmarking language model surprisal for eye-tracking predictions in Brazilian Portuguese. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 7–17.

Maria Barrett and Anders Søgaard. 2015. Reading behavior predicts syntactic categories. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 345–349.

Eckhard Bick. 2000. *The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework*. Aarhus Universitetsforlag.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Katja I Haeuser and Jutta Kray. 2024. Age differences in context use during reading and downstream effects on recognition memory. *Psychology and Aging*.

Victor Kuperman, Kazunaga Matsuki, and Julie A Van Dyke. 2018. Contributions of reader-and text-level characteristics to eye-movement patterns during passage reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(11):1687.

Howard Maclay and Charles E Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15(1):19–44.

Bruno Nicenboim, Pavel Logačev, Carolina Gattei, and Shravan Vasishth. 2016. When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in psychology*, 7:280.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Tony Berber Sardinha. 2010. Corpus brasileiro. *Informática*, 708:0–1.

Adrian Staub. 2021. How reliable are individual differences in eye movements in reading? *Journal of Memory and Language*, 116:104190.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.

CoordiMap: Conceptual Proposition of a new Framework for the Annotation of Verbal Elicitation Paths on Visual Experiment Stimuli and Introduction of the Associated Annotation Tool

Carmen Schacht
Ruhr-University Bochum, Germany
Department of Linguistics
carmen.schacht@ruhr-uni-bochum.de

Abstract

Consistent alignment of multi-modal experimental data—such as verbal utterances in elicitation tasks, (static) visual stimuli, and gaze data—presents a challenge in linguistic research. These elicitations often encode information about the visual perception strategies or cognitive processing of the scene. Thus, it is helpful to transform them into a structured, visually grounded format which captures the visual nature of the data, ideally able to be aligned with the corresponding gaze data. To achieve this, the present paper conceptually proposes the annotation framework for verbal elicitation paths as a data type and presents the first release of the associated newly developed *CoordiMap* annotation tool. The tool enables structured mapping of verbal elicitation data from experimental studies onto the corresponding visual stimuli. Independent of specific paradigms, the tool supports the annotation of verbal utterances in a linearized form based on coordinates directly marked on the image of the stimulus. The format is conceptually inspired by eye-tracking data formats, in which gaze behavior is represented as temporally linearized paths overlaid on the stimulus. The paper motivates the development of the tool and its annotation methodology by theoretical and experimental considerations regarding the relationship between visual perception and language production. As this a work in progress, the functionality of the annotation tool is demonstrated through an exemplary use case.

Keywords: Multi-modality, gaze data, elicitation data

1. Introduction

The collection and analysis of linguistic elicitation data in multi-modal experiments poses a challenge for researchers, particularly in terms of consistently linking diverse data modalities—such as verbal utterances, visual stimuli, and gaze data. This is especially relevant in experiments using static images as stimulus material, for example in scene or spatial description tasks, where there is a need for scientific tools that enable precise anchoring of linguistic expressions to the respective stimulus.

A typical use case involves linguistic elicitation experiments in which participants describe images containing specific visual cues or spatial configurations. These verbal data often include implicit cues about the participant's visual perception or cognitive processing of the scene (Griffin and Bock, 2000)—for example, by means of the sequence of the description or the choice of specific referential anchors (Klein, 2015). To systematically analyze such data, it is helpful to transform them into a structured, visually grounded format. The tool introduced here was developed with this goal in mind: it enables the annotation of linguistic elicitation paths directly on the stimulus image. Users can upload an image, mark relevant points according to the chosen linguistic paradigm by clicking on them, and assign a label to each point. These points are then connected in the chronological order of annotation, forming the elicitation path—a graphical representation of linear language production anchored to the visual stimulus.

The aim is to render verbal data in a form that is visually interpretable and analyzable, and that can be directly compared to other modalities—especially eye-tracking data. Just as eye-tracking operationalizes visual perception as temporally sequential paths over the stimulus, language similarly linearizes cognitive processes. The tool therefore simulates fixations, fixation durations, and saccades through the position, repetition, and connection of annotated points, offering a novel approach to analyzing linguistic and visual data within a shared coordinate framework.

The application was intentionally implemented as a framework-agnostic tool, allowing flexible integration into a wide range of theoretical and methodological research contexts and paradigms. Users can determine which linguistic units to annotate—from simple coreference expressions to complex information-structural constructions. The exported data (label, X, and Y coordinates) in `.csv` format allow for straightforward post-processing and integration into the analysis of related datasets. This tool thus provides a specialized and user-friendly platform that supports the anchoring of verbal elicitation paths on visual stimuli within a unified workflow—whether for the analysis of elicitation-only experiments or for combination with

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 35–40

eye-tracking data based on shared stimulus materials. To promote open-access resources, *CoordiMap* will be made available under a CC BY 4.0 license[1].

2. Previous work

In empirical language research—especially within multi-modal experimental setups—a key challenge lies in linking different modalities such as language with visual attention, similarly to pragmatic metrics like gesture (Lücking et al., 2015; Pfeiffer et al., 2006) or semantic metrics such as spatial description in spatial cognition research (Delucchi Danhier, 2019). This is particularly true for experiments employing static visual stimuli designed to elicit verbal responses, which require methods for precisely mapping verbal data onto the corresponding perceived regions of an image.

A well-established method for capturing visual perception is eye-tracking, which records eye movements and yields a linearized representation of the perceptual trajectory across a stimulus. This form of data—consisting of sequences of fixations, saccades, and fixation durations—captures a temporally ordered perception path (Blake, 2013), which conceptionally aligns with linguistic production data that themselves constitute a linear representation of multidimensional cognitive processes (Ferreira and Henderson, 1998; Delucchi Danhier, 2019). Not only in spatial cognition but in all forms of language production, there is a need to abstract multidimensional information into a sequential format—a process likewise inherent to the temporally ordered nature of linguistic signals in human language (Ferreira and Henderson, 1998).

The planning and structuring of language production—known as conceptualization (Levelt, 1989)—is influenced, among other factors, by the experimental task or "Quaestio" (Delucchi Danhier, 2019). This process involves, for instance, the selection and linearization of information as well as the contextually appropriate choice of granularity in arranging that information (von Stutterheim and Carroll, 2007). Assuming that such factors are reflected not only in the composition of verbal elicitation but also in the associated gaze behavior, classic studies such as Tanenhaus et al. (1995) and Griffin and Bock (2000)—which integrated visual and linguistic data based on shared stimuli—have already demonstrated the conceptual and methodological viability of combining eye-tracking with linguistic elicitation. However, the encoding of combined data in such studies has varied, often mapping verbal data as temporal markers onto gaze paths rather than treating both as independent, structurally comparable trajectories (Griffin and Bock, 2000)—an approach that highlights the potential added value of path-based comparison proposed in this paper. This becomes particularly relevant in the case where the experiments—eye-tracking and elicitation—are performed by two separate groups of participants and thus not produce temporally matching gaze and elicitation data.

Related approaches such as 'Meaning Maps' (Henderson and Hayes, 2017, 2018) further establish visual annotation types that map image stimuli based on their semantic relevance. These mappings not only provide insights into task-driven visual salience but also serve as annotation formats for modeling attention in experimental contexts. Such methods motivate the joint investigation of gaze and language behavior in relation to shared visual stimuli—and underscore the need for an annotation tool that facilitates this type of multi-modal analysis.

2.1. Motivation for the Elicitation Path as a Data Type

Within such experimental designs, there is a need to transform verbal elicitation data into a visually grounded and formally structured data type. Verbal descriptions of image content—such as in spatial or scene descriptions—often follow implicit cognitive paths that can be traced back to the stimulus itself. Explicitly modeling these sequences as elicitation paths allows not only for the visualization of linguistic production processes but also creates a basis for direct comparison with eye-tracking data: both modalities sequentially represent perceptual and constructive processes anchored on the same stimulus. This format is particularly well-suited for experimental tasks that investigate the relationship between perception and linguistic structure—for instance, with respect to reference strategies, information structure, or spatial description patterns (Delucchi Danhier, 2019; Griffin and Bock, 2000). In contrast to individual markers that denote isolated referential points, the elicitation path enables the annotation of more complex structures that can be visualized as paths across the image—potentially aligning closely with the conceptual structure of gaze paths.

2.2. Motivation for the Annotation Tool

Existing annotation tools such as *LabelImg* (Tzutalin, 2015) or the *VGG Image Annotator (VIA)* (Dutta and Zisserman, 2019) offer robust functionality for visual image annotation (e.g., through bounding boxes), but they are not designed for the annotation of linguistic elicitation paths. However, specialized approaches like the 'Meaning Map' by Henderson and Hayes (2017), and their contribution

[1] https://osf.io/8ke4c/overview.

to linguistic research, highlight the need for task-specific tools that support linguistic annotation of visual stimuli—thereby expanding the methodological repertoire of empirical linguistics. In this context the *CoordiMap* tool was developed: a framework-agnostic, user-friendly annotation tool that makes verbal elicitation paths visually accessible on static stimuli and transforms anchors into concrete locations on the image for analysis and evaluation of the path. Users can upload a stimulus image, define relevant anchor points through simple clicks, and have these points automatically connected into paths. Each path represents a verbal utterance or cognitive sequence of linguistic production, based on the user's underlying theoretical model and can be exported as structured `.csv` files containing labels and X/Y-coordinates. The resulting data format is explicitly implemented to support comparability with eye-tracking data, as both are spatially grounded on the coordinate level and conceptually simulate fixations and saccades. This enables, for example, the investigation of whether the sequence of verbal descriptions of a scene corresponds to its visual perception—a line of inquiry particularly relevant for combined eye-tracking and elicitation studies like Griffin and Bock (2000). Importantly, the tool allows for flexible integration into a wide range of theoretical annotation frameworks—from simple coreference annotations and semantic descriptions of space (Kababgi et al., 2024; Sitter et al., 2025) to more complex information-structural annotations in the context of spatial cognition (Delucchi Danhier, 2015, 2019; Delucchi Danhier et al., 2025).

3. Functionality of *CoordiMap*

CoordiMap is a lightweight, locally run tool that is straight-forward to use and does not require an extensive amount of training. It was implemented in Python (Van Rossum and Drake, 2009) and comes with a `README` file for set-up. The tool incorporates the libraries tkinter (Lundh, 1999), pillow (Clark, 2015), numpy (Harris et al., 2020), and matplotlib (Hunter, 2007). To showcase the functionality of the new tool, the individual functions are demonstrated in a simple exemplary annotation. For demonstration purposes assume the simple annotation task of the annotation of entities represented as plain nominal mentions. A very simple exemplary elicitation regarding the example image might look like this:

(1) *'There is a tree. A bird is sitting next to a branch.'*

This elicitation thus contains the three nominal mentions of *tree*, *bird*, and *branch*, which will have to be annotated in this example. The following paragraphs will demonstrate the workflow for this annotation.

The Graphical User Interface (GUI). To run the software, the script has to be executed within the current working directory. It will then launch a GUI, where all further tasks can be carried out. Alternatively, it can be launched via the accompanying `.exe` file. The tool will be used by means of the buttons on the left side (see left side of Figure 1) and mouse clicks on the image, which can be uploaded from the user's file architecture. When a file is selected, the tool will upload it into the canvas-space in the GUI (see right side of Figure 1).

Figure 1: Exemplary annotation of the elicitation path. Reuse functionality (red) and renaming functionality (blue).

Annotation. To start annotation, the user can click the position on the image that is supposed to be annotated and thus anchor the linguistic unit to the respective position on the stimulus. The image on the canvas will then display a dot at the selected position. Every subsequent anchor is then connected to the previous one with a line creating a verbal path across the image (see Figure 1). This format is designed to mirror the format of fixations and saccades forming a view path in the data of eye-tracking experiments. To simulate prolonged fixation times typically found in eye-tracking data, the tool is able to detect multiple consecutive annotations of the same position and increase the size of the dot at this position. It will differentiate between consecutive anchors and non-consecutive anchors separated by other anchors. All individual anchors are listed in the 'point'-drop-down menu for further use. The labels are displayed next to the points on the canvas, if the user hovers the cursor over the respective point (see 'bird'-anchor in Figure 1).

Polygon Mode and Regions. In case larger entities or collective descriptions, which span extended areas of the image, need to be annotated, the tool offers a polygon mode, to create regions of adaptable sizes according to the chosen paradigm. In

the demonstration example this could apply for the mention of ‘tree’ or ‘branch’, as those span larger chunks of the image compared to ‘bird’. Those regions enable the calculation of the entities centroid–the coordinate at the center of the region–to represent the region in the form of a single anchor. To annotate regions, the mode needs to be changed from path mode to polygon mode by toggling the ‘Switch to Polygon Mode’-button. Once the mode has been changed, a region of variable size can be created by outlining the respective part of the image with clicks. The tool will automatically calculate the centroid and use it as a representative position for the region.

Figure 2: Exemplary annotation of the Polygon Mode Interface (dotted orange outline) and the Polygon Naming Interface.

Selecting and Reusing Regions and Points. To ensure a consistent annotation across participants on the same stimulus image, points and regions may be reused once annotated. This guarantees exact coordinate positions when the same unit has to be annotated. To reuse a point or a region it can simply be selected in the respective drop-down menu (see Figure 1) and used by clicking ‘Use selected point/region’.

Renaming Regions and Points. To rename a point or a region for further use, the point or region has to be selected in the respective drop-down menu. Then the user can click ‘Rename point/region’, which will open an interface, where a new label can be entered. The new label will replace all instances of the previous name of the label; even in the export file. This functionality will also update the hover-labels already displayed on the canvas (see Figure 1).

Export Coordinates of the Path. To export a finished path, a filename has to be entered into the according export interface. The export function only accepts `.csv` file-format. A `.csv`-file is then exported into the current working directory containing the annotated coordinates aligned to the stimulus in pixels as well as the label. Table 1 shows the output for the demonstration example of tree, bird and branch. The left column contains all x-coordinates of the annotations, the right column contains all the y-coordinates. After exporting the annotations, the user can either clear the current path or the entire set-up.

Label	X-Coordinate	Y-Coordinate
Tree	615	554
Bird	532	372
Branch	831	639

Table 1: Exemplary output of the export function in pixels.

4. Discussion

The presented tool constitutes a first version of a specialized annotation platform for elicitation data, aimed at precisely mapping verbal elicitation paths onto corresponding visual stimuli. Its conceptual foundation is based on the assumption that both gaze behavior and verbal description can be understood as linearizing processes of cognitive perception (Delucchi Danhier, 2019; Ferreira and Henderson, 1998). By visually anchoring linguistic units to a shared stimulus, the tool enables the identification of structural parallels between language production and perception—particularly in experimental setups that combine eye-tracking with verbal elicitation (Griffin and Bock, 2000). The tool represents an important first step toward not only conceptually but also practically linking two well-established data modalities in empirical language research: visually anchored speech and visual perception. In the long term, the framework and tool may prove useful not only in experimental linguistic contexts but also in interdisciplinary fields such as cognitive science or human-computer interaction, where the integration of multimodal data plays an increasingly central role.

In particular, the approach may prove relevant for the development and training of multimodal LLMs. The framework introduced here could support the manual data annotation processes that typically precede computational modeling, especially in contexts where aligned visual–linguistic data are required. By explicitly encoding the relationship between verbal production and visual reference points, the tool contributes to the creation of structured datasets that may facilitate the learning of grounded language representations. At the same time, it should be noted that the toy example utilized to demonstrate the functionality of the tool represents a deliberately simple application scenario, serving to illustrate the core mechanics

of the reasoning behind the scheme and the annotation process. It does, however, not exhaust the methodological potential of the approach. With the development of more complex and task-specific annotation guidelines the elicitation data could be encoded even more fully. For instance, future extensions of the framework could incorporate the temporal dimension of speech into the annotation process by integrating temporal information such as onset times, durations, or pauses. The elicitation paths as a data type could thus be enriched to reflect not only the sequential order of linguistic units but also their temporal unfolding. This would allow for an even closer comparison with eye-tracking data, in which the temporal dimension plays a fundamental role, thereby further aligning the two data types. Furthermore, the annotation framework could be expanded to explicitly incorporate the linguistic instructions of the elicitation task, since task design (the Quaestio) has a substantial impact on both linguistic production and perceptual strategies, as mentioned previously and thereby providing additional explanatory power for observed patterns in the data.

More generally, the annotation process itself could be supported by (semi-)automatic preprocessing of the elicitation data. For example, syntactic parsing, morphological tagging, or automatic coreference resolution could guide annotation decisions depending on the chosen theoretical framework. Such integrations would not only increase annotation efficiency but also form a interdisciplinary bridge between psycholinguistic research and NLP. Regarding these considerations, the current implementation should be understood as a foundational step. Its primary contribution lies in the proposition of an extensible framework that can be further adapted to complex research questions across disciplines.

5. Conclusion

The framework and tool introduced here conceptually address a methodological gap in experimental linguistics by visually anchoring verbal elicitation data and transforming it into a format that parallels the structure of eye-tracking data. The tool was developed with the goal of providing a lightweight, framework-agnostic, and locally executable instrument that can be used across a variety of experimental settings—especially for analyzing the relationship between visual perception and language production. In a follow-up step of this ongoing project, both the framework of verbal elicitation paths and the *CoordiMap* tool will be tested and evaluated empirically in a pilot annotation study comparing gaze data with elicitation paths. Future versions of the tool are intended to expand on the current functionality and adapt to the empirical demands of linguistic and cognitive research.

6. Limitations

Despite the potential of the application, the current version has several limitations:

No Automatic Alignment Functionality with Eye-tracking Data At present, the annotated elicitation paths are exported using pixel-based fixed-resolution coordinates, which are not automatically aligned with typical metrics of eye-tracking data (e.g., normalized stimulus regions, fixation durations, Areas of Interest, or the specific layout of the respective eye-tracking system). To enable direct comparability, manual post-processing—such as coordinate transformation—is currently required. Future versions may include automated coordinate alignment features to further facilitate the integration of verbal and visual paths. Additional functionality could include automatic linking of annotated points to pre-parsed linguistic features depending on the chosen paradigm—for example, syntactic or morphological information annotated in advance for each anchor point. Other annotation types, such as coreference, could also be added to mark different realizations of referential expressions.

No Empirical Evaluation with Annotated Elicitation Data As this is a work in progress, the current version of *CoordiMap* primarily serves to introduce the annotation concept and showcase its technical feasibility. A systematic application to real-world data—such as a pilot annotation study comparing elicitation and eye-tracking data—has still to be conducted. In such a future pilot study it is especially important to analyze inter-annotator consistency and conduct a usability evaluation. Only such empirical use cases will allow for a comprehensive evaluation of the tool's added methodological value for experimental linguistic research.

Theoretical Dependence of Annotation Interpretability Because the tool is designed to be framework-agnostic, the interpretability and informativeness of the annotated paths largely depend on the chosen theoretical model. The quality and granularity of the annotations may vary according to the underlying linguistic framework (e.g., coreference, information structure, semantics) and must be supported by clearly defined annotation guidelines.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments. This research is funded by

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

7. Bibliographical References

C. Blake. 2013. Eye-tracking: Grundlagen und anwendungsfelder. In W. Möhring and D. Schlütz, editors, *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*. Wiesbaden, Springer VS. Doi:10.1007/978-3-531-18776-1_20.

A. Clark. 2015. Pillow (pil fork) documentation.

R. Delucchi Danhier. 2015. *Sprachspezifische Aspekte der Informationsverteilung in Wegan-weisungen*. Schneider Verlag Hohengehren, Baltmannsweiler.

R. Delucchi Danhier. 2019. Linearisierungsstrategien und ihr einfluss auf die informationsstruktur und die syntaktische komplexität von zimmerbeschreibungen. In Tübingen:, editor, *Raumrelationen im Deutschen*, pages 69–89. Stauffenburg.

R. Delucchi Danhier, B. Mertins, H. Mertins, and G. Schneider. 2025. Entropy as a lens: Exploring visual behavior patterns in architects. *Journal of Eye Movement Research*, 18(5).

A. Dutta and A. Zisserman. 2019. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA. ACM.

F. Ferreira and J. M. Henderson. 1998. Linearization strategies during language production. *Mem. Cognit.*, 26(1):88–96.

Z. M. Griffin and K. Bock. 2000. What the eyes say about speaking. *Psychological Science*, 11(4):274–279. PMID: 11273384.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R.Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernandez del Rio, M. Wiebe, P. Peterson, P. Gerard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

J. M. Henderson and T. R. Hayes. 2017. Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, 1:743–747. Doi:10.1038/s441562-017-0208-0.

J. M. Henderson and T. R. Hayes. 2018. Rehrig, g., ferreira. *F. Meaning guides attention during real-world scene description. Scientific Reports*, 8:13504.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

D. Kababgi, G. Grisot, F. Pennino, and B. Herrmann. 2024. Recognising non-named spatial entities in literary texts: a novel spatial entities classifier. In *Proceedings of the Computational Humanities Research Conference*, volume 3834 of *CEUR Workshop Proceedings*, pages 472–481.

W. Klein. 2015. *Überall und nirgendwo. Subjektive und objektive Momente in der Raumreferenz (1990)*, pages 177–207. J.B. Metzler, Stuttgart.

W. J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. The MIT Press.

F. Lundh. 1999. An introduction to tkinter. *URL: www. pythonware. com/library/tkinter/introduction/index. htm*.

A. Lücking, T. Pfeiffer, and H. Rieser. 2015. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79. Doi:10.1016/j.pragma.2014.12.013.

T. Pfeiffer, A. Kranstedt, and A. Lücking. 2006. Sprach-gestik experimente mit iade, dem interactive augmented data explorer. In *Dritter Workshop Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR, Koblenz*, pages 61–72.

E. Sitter, O. Momen, F. Steig, J. B. Herrmann, and S. Zarriess. 2025. Annotating spatial descriptions in literary and non-literary text. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 308–325, Vienna, Austria. Association for Computational Linguistics.

M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N. Y.)*, 268(5217):1632–1634. Doi:10.1126/science.7777863.

Tzutalin. 2015. Labelimg. Free Software: MIT License.

G. Van Rossum and F. L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

C. von Stutterheim and M. Carroll. 2007. Durch die grammatik fokussiert. *Zeitschrift für Literaturwissenschaft und Linguistik*, 145:35–60.

A Comparative Study Between Mouse and Eye Tracking Signals for Long Romanian Texts

Alexandru Bogdan Gheorghe Sergiu Nisioi

Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest

ghrghbogdan@gmail.com, sergiu.nisioi@unibuc.ro

Abstract

Understanding human language processing via eye-tracking (ET) is precise but limited by scalability. Mouse-Tracking (MoTR) offers a cost-effective alternative, yet its viability for long-form reading in languages like Romanian remains underexplored. The primary challenge lies in the motor-induced noise and biomechanical discrepancies between hand and eye movements. Here we show that combining targeted technical enhancements with a Hertz-based velocity transformation suggests that MoTR can capture significant cognitive signals comparable to ET. We evaluate this by training a BERT-enhanced Fusion Model that integrates semantic context to bridge the mechanical gap, achieving an internal consistency of $\rho \approx 0.58$ and a cross-modal correlation of $\rho \approx 0.22$ in the velocity domain. These results indicate that when properly normalized, manual tracking captures similar cognitive constraints as gaze, with predictive accuracy approaching the empirical bounds of human behavioral variance.

Keywords: Mouse-tracking, Eye-tracking, Cognitive Modeling, BERT Fusion, Gaze Prediction

1. Introduction

Although eye-tracking (ET) remains the gold standard for capturing the 'Eye-Mind' connection, the prohibitive cost of hardware and the constraints of laboratory environments pose significant barriers to data collection, particularly for low-resource languages. While initiatives like the MultiplEYE project[1] aim to provide standardized multilingual corpora, the data available for the Romanian language remain scarce compared to English (Kennedy et al., 2003; Luke and Christianson, 2018). On the one hand, Romanian shares certain structural similarities with other Romance languages, but on the other hand it presents particular challenges: the widespread use of diacritics increases visual crowding and decoding difficulty, a relatively flexible word order that complicates incremental syntactic integration and increases word-level surprisal.

Mouse-tracking while reading (MoTR) has emerged as a promising proxy, contributing to large-scale browser-based experiments Wilcox et al. (2024). Despite its potential, the transition from ocular to manual tracking introduces significant motor noise. We identify two critical bottlenecks: first, the "noise" generated during return sweeps, where the cursor inadvertently records data while transitioning between lines; and second, the variability in participants' manual dexterity. We observed that maintaining the cursor precisely on a text line is a demanding task that varies significantly across individuals, often leading to vertical "drifting" that corrupts the signal .

To address these challenges, we propose an Enhanced MoTR Framework (illustrated in Figure 1) featuring structural interface modifications: a Row-Level Gatekeeper to isolate line-specific reading and a Vertical Axis Constraint (vertical lock) to compensate for varying dexterity. Due to the restrictions of the gatekeeper system, we implement a click-to-release mechanism that enables intentional regressions, allowing users to manually bypass line constraints to revisit previous segments without compromising the system's automated tracking integrity.

Beyond structural interface modifications, we consider the mathematical representation of reading behavior. In our observations, reading durations in MoTR typically follow a more pronounced right-skewed, long-tail distribution, which can be suboptimal for linear analyses such as Pearson's correlation. This is problematic because it may distort the perceived alignment between modalities by over-emphasizing high-duration motor outliers. To mitigate this, we explore projecting these durations into the velocity domain (Hertz). Such transformation provides a natural normalization of the data, potentially stabilizing the signal and leading to a more consistent comparison between the different mechanical scales of ocular and manual movements.

Finally, we show how surface-level linguistic features (word length or frequency) can be augmented

All authors are corresponding authors.

[1] https://multipleye.eu

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 41–49
May 12, 2026.

to better capture the cognitive effort. Thus, we integrate contextual embeddings into our predictive engine. Observing how latent semantic difficulty influences reading behavior provides a more nuanced lens through which to evaluate the alignment between manual and ocular signals, rather than relying solely on surface-level statistics.

2. Related Work

The foundation of eye-tracking research in psycholinguistics rests on the Eye-Mind Hypothesis (Just and Carpenter, 1980), which posits that gaze duration is a direct proxy for cognitive processing time. Landmark studies by Rayner (1998); Rayner et al. (2003) established how lexical features, such as word length and frequency, systematically influence fixation patterns. More recently, the MultiplEYE initiative (Jäger et al., 2026) has expanded these by developing standardized multilingual corpora. However, collecting high-fidelity ocular data for the Romanian language remains a significant challenge, with the MultiplEYE project standing as the sole major initiative currently addressing this lack of standardized resources.

As a scalable alternative to expensive eye-tracking hardware, Wilcox et al. (2024) introduced the MoTR paradigm, demonstrating that cursor trajectories can successfully capture fundamental psycholinguistic effects. Despite its potential, existing literature (Wilcox et al., 2024; Huang et al., 2012) highlights that manual signals are inherently "noisier" than ocular ones. Challenges such as varying manual dexterity and the noise generated during "return sweeps" (transitions between lines) remain significant hurdles for signal precision.

Most existing studies rely on single-sentence stimuli or short fragments to minimize motor variance (Oğuz et al., 2025; Popescu and Nisioi, 2025). While it is effective for isolating lexical effects, these approaches do not capture the complexity of long-form naturalistic reading. As the text length increases, the cumulative effects of fatigue become more pronounced, leading to errors. Recent task-oriented datasets like PreferRead (de Langis et al., 2025) have begun using MoTR for longer texts to monitor LLM annotators, but they focus on evaluation rather than naturalistic reading.

Initially tested on English passages from the Provo Corpus (Wilcox et al., 2024), the method has recently been expanded to other languages. For instance, Oğuz et al. (2025) utilized MoTR to probe gender agreement in Russian, while Haveriku et al. (2025) explored its feasibility for Albanian. In the Romanian context, Popescu and Nisioi (2025) provided the first application of MoTR using isolated sentences from the MLSP dataset.

Traditional predictive models of reading behavior have relied on surface-level features, such as word length or Zipf frequency (Kliegl et al., 2004a). The integration of pre-trained language models like BERT (Devlin et al., 2019) provides a means to represent semantic difficulty and "spillover effects". By utilizing PCA for dimensionality reduction (Pedregosa et al., 2011), we aim to focus the model on the most significant semantic patterns while facilitating a more stable training process.

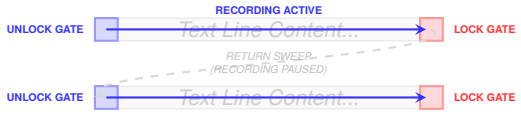
Figure 1: The enhanced MoTR interface featuring the Row-Level Gatekeeper system and the spotlight mechanism.

3. Methodology

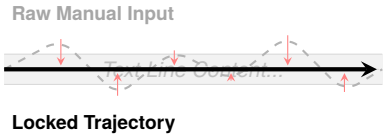
We propose a multi-stage pipeline that integrates structural modifications to the Mouse-Tracking while Reading (MoTR) paradigm with a specialized data-processing workflow. These enhancements are specifically made to minimize motor-induced noise and enforce a tighter coupling between manual displacement and cognitive processing. The process is organized into two primary stages: an experimental setup involving interface constraints and dual-modality data collection, followed by a computational pipeline that integrates speed-based normalization and contextual modeling.

3.1. Experimental Setup and Technical Enhancements

The study involved 18 participants, divided into two equal groups: 9 subjects for the Mouse-Tracking (MoTR) tests and 9 subjects for the Eye-Tracking (ET) baseline. The MoTR group consisted of university-level students aged between 20 and 29, all of whom are native Romanian speakers. This specific age range was selected to ensure that all participants had similar computer skills and reading habits, reducing the risk of motor-skill differences affecting the data.



(a) Row-Level Gatekeeper mechanism.



(b) Vertical Axis Constraint (OY lock).

Figure 2: Functional enhancements to the MoTR interface for noise reduction.

The reading materials were selected from the Romanian sub-corpus of the MultiplEYE project (Jakobi et al., 2026; Nisioi et al., 2026; Kasperė et al., 2026). We used these texts because they are standardized and provide a consistent mix of different genres and difficulty levels. To ensure that the data reflects genuine cognitive effort, each reading session was followed by six comprehension questions with four response options. These questions served as a control mechanism.

Furthermore, the questions were designed by the MultiplEYE team (Hollenstein et al., 2026) to assess the active reading of the text. Instead of focusing on simple surface-level facts, they required participants to integrate information and process the deeper meaning of the content, ensuring that the tracking data reflects high-level cognitive engagement.

To minimize motor-induced noise and improve data quality, we introduced three critical functional improvements to the original MoTR paradigm. These modifications were specifically designed to ensure that the mouse movements reflect the user's mental focus as accurately as possible.

Row-Level Gatekeeper System: We implemented this system (see Figure 2a) to separate actual reading time from the "noise" created during line transitions. In standard reading, the time spent moving the cursor from the end of one line to the start of the next (the "return sweep") does not represent text processing and implies a significant amount of noise. By using neutral-colored trigger boxes to "unlock" and "lock" each row, we ensure that we only record the time when the user is actively engaged with the words on that specific line.

Vertical Axis Constraint (OY lock): This feature (visualized in Figure 2b) was added to prevent the cursor from drifting vertically between rows. During our initial observations, we noticed that participants have different levels of manual dexterity; without a lock, the cursor would often slip off the line, making the data messy and adding cognitive effort to the participant. By locking the vertical axis, we force the "spotlight" to stay perfectly centered on the text, ensuring that the recorded movement reflects the user's cognitive progress rather than accidental hand instability.

Nonlinear Navigation and Snapping: To allow for natural reading habits, such as going back to re-read a word, we developed a click-and-drag mechanism. When a user wants to re-read a part of the text, they simply hold the click and drag to the word that they want to read again. When the mouse button is released on the specific word, the cursor automatically "snaps" to the row and reactivates the vertical lock. This modification helps the user to move freely through the text while keeping the data structured and perfectly aligned with the rows.

Metric	Raw	Z-score	Log	Hertz
Total Duration	0.13	0.13	0.07	**0.23**
First Duration	0.08	0.08	0.07	**0.17**

Table 1: Pearson correlation coefficients across different normalization techniques on MoTR and ET means.

3.2. Data Pipeline and Predictive Modeling

Both MoTR and ET recordings were processed using the same pipeline to ensure they could be compared directly. We used unique word identifiers to align the tracking signals with each specific token in the text. To model genuine reading behavior, we filtered out data points that were too short to represent cognitive effort—specifically durations under 160 ms for MoTR and 80 ms for ET. These thresholds account for the natural physical limits of the hand and the eye (Rayner, 1998; Card et al., 1983), removing "noise" from the dataset.

Statistical analyses of reading times often encounter 'long-tail' distributions in raw millisecond data, which can distort linear correlations (Kliegl et al., 2004b). While log-transformations are a standard approach for variance stabilization and Z-score normalization is frequently used for scale alignment, the latter remains a linear transformation that fails to address the underlying non-normality and skewness of the data, we explore as an alternative the projection of these durations into the velocity domain (Hertz). This transition is achieved by inverting the temporal duration (d) recorded for each word using the formula $Hz = 1000/d$, effectively converting 'time spent' into 'processing speed'. As demonstrated in Table 1, this approach aims to provide a more robust basis for cross-modal com-

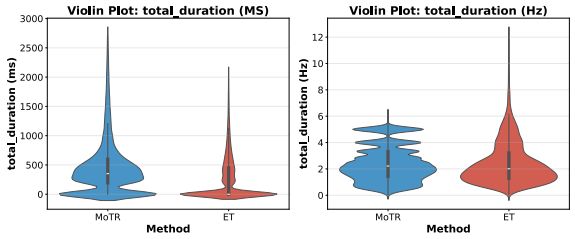


Figure 3: MoTR and ET duration distributions in ms (left) and Hz (right). The Hz projection aligns core densities and stabilizes MoTR variance, while ET's extended tail reflects superior ocular velocity.

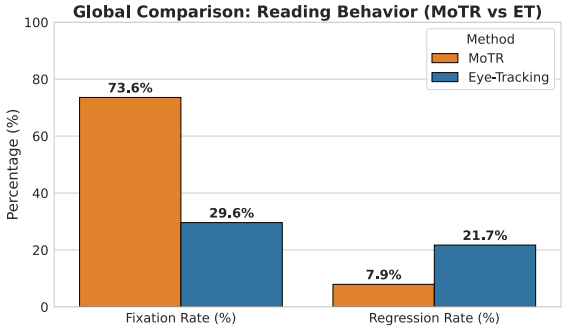


Figure 4: Global reading metrics comparison between MoTR and ET. The manual modality (MoTR) exhibits a significantly higher fixation rate (73.6%), while spontaneous regressions are more frequent in eye-tracking (21.7%), reflecting the mechanical differences between cursor-based and ocular reading behaviors.

parison—referring here to the statistical alignment between the ocular (ET) and manual (MoTR) tracking signals. By doing so, we seek to normalize the mechanical lag between hand and eye movements into a unified, speed-based metric that may better reflect the underlying cognitive pace.

Converting to Hertz (speed) balances these values, making the data more consistent and easier for our models to process. Also, projecting the data into a velocity space (speed) makes the mouse and eye signals look more alike (Figure 3). This transformation helps the model to focus on the overall "pace" of reading, which we found to be a more stable signal for predicting eye movements from mouse data.

The predictive **Baseline** employs a Random Forest Regressor with 100 estimators and a minimum leaf size of 10 selected for its robustness in handling heterogeneous features and its capacity to capture non-linear interactions between lexical attributes without assuming a specific data distribution. We compared two distinct configurations to evaluate how different types of information affect the prediction of reading speed. This model uses basic word features, such as Zipf frequency (Speer and Chin, 2021), which provides a logarithmic measure of word commonality on a scale (typically 0-8) rather than raw occurrence counts, word length, and syllable count (Kozea, 2025) that utilizes **Hunspell** hyphenation rules to account for the phonological complexity of each token.

To capture "spillover effects", where the difficulty of a previous word affects the current one, the model considers a window of two words before and one word after the target token. While the semantic embeddings are decontextualized, the situational context is explicitly modeled through this sliding window of linguistic features. Specifically, the regressor is provided with the left context (frequency and length of t-1,t-2) and right context (the immediate succeeding word t+1), alongside the relative position of the word within the paragraph. To enhance the baseline by adding semantic context we decided to make a **Fusion Model** with our Random Forest Regression and BERT (`bert-base-multilingual-cased`). To handle tokenization, each word was processed as an independent sequence from which we extracted the hidden state of the [CLS] token from the final hidden layer. This captures the highest level of semantic abstraction and serves as a pre-computed aggregate representation for all sub-word units generated by the tokenizer for that specific word. To maintain model efficiency and robustness, we utilized Principal Component Analysis (PCA) to condense the 768-dimensional BERT embeddings into 32 principal components, retaining 95% of the semantic variance while significantly reducing the feature space (Takeshita et al., 2025). This dimensionality reduction prevents the Random Forest Regressor from overfitting to the high-dimensional noise often found in large embeddings, thereby ensuring the model generalizes better to unseen texts rather than simply memorizing the training data. The models were trained and tested using an 80/20 split. Performance is reported using Spearman's rank correlation (ρ) and Pearson's (r) to measure the alignment between predicted and real reading speeds, along with Mean Absolute Error (MAE) to track the average prediction gap.

4. Results and Discussion

4.1. Behavioral Analysis and Signal Validations

The fundamental behavioral discrepancies between the two tracking modalities are illustrated in Figure 4. In this context, we define the fixation rate as the percentage of unique tokens receiving at least one tracking event, while the regression rate represents the proportion of words revisited

through backward movements. MoTR and ET exhibit divergent reading signatures: MoTR shows a significantly higher fixation rate (73.6%) compared to ET (29.6%), whereas ET reveals a much higher regression rate (21.7%) than MoTR (7.9%). The data indicates that the MoTR paradigm enforces a linear reading trajectory, which directly accounts for the higher fixation rate observed. While ocular reading is rapid and almost effortless for text scanning, the manual interface requires a deliberate line-by-line progression. Consequently, the regression rate is significantly suppressed in MoTR because manual cursor repositioning imposes a substantial physical and cognitive load on the participant. Because the overall time spent on each token is inherently higher in the mouse-tracking modality, users may achieve deeper initial processing, resorting to backward movements as a last resort only when comprehension is critically challenged. In essence, the mechanical effort of the hand serves as a filter that occurs only in the most cognitively necessary regressions.

To quantify the initial alignment between the two modalities, we calculated Pearson correlations based on the means of all participants. As detailed in Table 1, projecting reading metrics into the velocity domain (Hz) almost doubles the alignment for total duration, rising from $r = 0.13$ to $r = 0.23$. This improvement suggests that processing speed provides a stable basis for cross-modal comparison than raw temporal data, as it successfully compensates for the mechanical inertia and the inherent gap between hand and eye movements. This alignment is further reinforced by analyzing the Word Length Effect, a standard psycholinguistic benchmark (Rayner, 1998). As illustrated in Figure 6, the correlation in the velocity domain reaches $r = 0.95$ for the total duration and $r = 0.89$ for the first duration, providing strong empirical evidence that our enhanced MoTR framework reflects the same fundamental linguistic processing speeds as high-fidelity eye-tracking.

4.2. Predictive Performance and Behavioral Consistency

To establish a framework for evaluating our results, we first defined an empirical upper bound through an inter-participant consistency analysis. We opted for this pairwise consistency measure over traditional split-half reliability to more effectively assess the robustness of the MoTR signal across the participant pool. Given our sample size, the pairwise approach ensures that the consistency estimate remains independent of arbitrary data partitioning, whereas split-half reliability would have been susceptible to selection bias. As illustrated in Figure 5, the MoTR signal maintains an average inter-reader

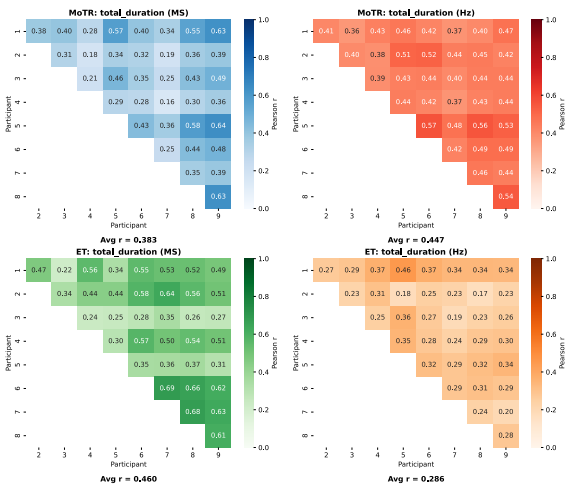


Figure 5: Intra-method Consistency: Inter-participant correlation matrices showing the behavioral upper bound for both modalities. The correlation has been measured between shifted subgroups (Subjects 1-8 vs. Subjects 2-9) to avoid self-correlation.

correlation of $r \approx 0.46$. Since the correlation between two human readers within the same modality represents the natural ceiling of behavioral similarity, it is statistically unrealistic to expect a cross-modal predictive model to significantly exceed this threshold. When viewed against this benchmark, our model's performance—reaching a Spearman $\rho = 0.34$ for the ET $\rightarrow$ MoTR direction and $\rho = 0.22$ for the MoTR $\rightarrow$ ET direction demonstrates that our approach is capturing a meaningful portion of the shared cognitive signal. The predictive results, summarized in Table 2, further reveal that the Fusion Model consistently outperforms the linguistic baseline across all scenarios and normalization strategies. While the standard log-transformation stabilizes variance and improves alignment over raw millisecond data (e.g., increasing the MoTR $\rightarrow$ ET correlation from 0.17 to 0.20), the best results are observed in the velocity domain (Hz). This superiority confirms that semantic context, provided by BERT embeddings, is important for bridging the inherent gap between motor and ocular behaviors. While the baseline model relies on surface-level features like word length, the Fusion Model uses contextual information to account for the "cognitive lag" between the eye's near-instantaneous movement and the hand's more deliberate cursor control. This integration is particularly evident in the velocity domain (Hz), where the alignment reaches its peak. By projecting both signals into a unified velocity space, we enable a more effective alignment between BERT-derived semantic representations and the observed reading pace, successfully bridging the mechanical gap between hand and eye movements. These results support the use of MoTR as

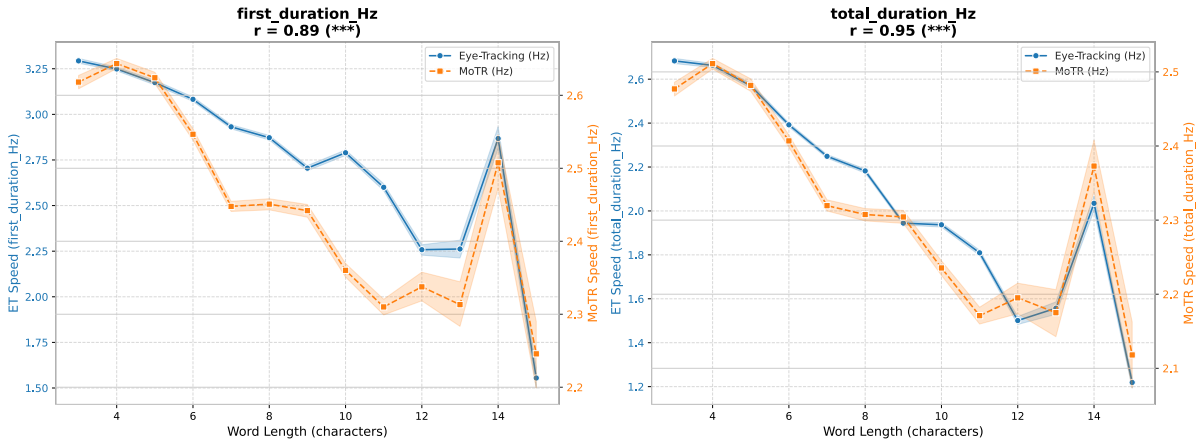


Figure 6: Psycholinguistic Validation: Word Length Effect comparison between ET and MoTR in the Velocity Domain (Hz), showing a near-perfect alignment ($r = 0.95$).

	Raw Temporal (ms)		Log-Transformed		Velocity Domain (Hz)	
Validation Scenario	**Base (ρ)**	**Fusion (ρ)**	**Base (ρ)**	**Fusion (ρ)**	**Base (ρ)**	**Fusion (ρ)**
MoTR $\rightarrow$ MoTR (Internal)	0.54	0.53	0.57	**0.58**	0.56	**0.58**
MoTR $\rightarrow$ ET (Cross-Modal)	0.16	0.17	0.19	0.20	0.20	**0.22**
ET $\rightarrow$ ET (Internal)	0.40	0.41	0.42	**0.44**	0.41	0.43
ET $\rightarrow$ MoTR (Cross-Modal)	0.26	0.34	0.27	**0.35**	0.27	0.34

Table 2: Performance comparison (Spearman ρ) across raw temporal (ms), log-transformed, and velocity (Hz) domains. Bold values indicate the best performance for each scenario.

a viable proxy for eye-tracking, providing a scalable alternative for capturing cognitive signals in NLP research.

5. Conclusion

In this study, our findings suggest that Mouse-Tracking (MoTR), when augmented with specific technical and methodological enhancements, shows potential as a scalable alternative to Eye-Tracking (ET) in cognitive NLP tasks. The implementation of the Row-Level Gatekeeper system and the vertical axis constraint were designed to isolate cognitive signals from motor-induced noise. Participants enjoyed these enhancements, who reported an intuitive and natural reading experience. These architectural modifications ensured that the recorded cursor trajectories reflect the reader's mental state rather than accidental hand instability or mechanical artifacts from row transitions. Our findings highlight that Hertz transformation can attenuate the skewness of the data and improve the linear relationship between reading times. By projecting raw durations into a unified velocity domain, we achieved an alignment regarding the Word Length Effect ($r = 0.95$), providing additional proof that MoTR captures similar underlying linguistic processing speeds as professional-grade eye-tracking hardware. Furthermore, the integration of semantic context through the BERT-enhanced Fusion Model was intended to capture additional linguistic nuances and incorporate contextual information that surface-level features alone may not fully represent. The proximity of our model's performance ($\rho = 0.34$) to the empirical ceiling of inter-participant consistency ($r \approx 0.46$) suggests that the predictive results are nearing the upper bound defined by inherent human behavioral variance. Ultimately, this research adds towards to Mouse Tracking framework for democratizing access to high-resolution cognitive signals.

By removing the financial and logistical barriers associated with traditional eye-tracking, our proposed changes to the MoTR paradigm can further aid the collection of large-scale, crowdsourced cognitive datasets. Future work will focus on expanding this validation to cross-linguistic settings and investigating the utility of these signals in low-resource language modeling.

6. Limitations

Several limitations must be acknowledged. First, the participant pool consisted of 9 subjects per modality; a larger and more diverse demographic sample would further strengthen the generalizability of these findings. Second, mechanical hand inertia remains a significant factor. The hand is fun-

damentally slower than the eye, as evidenced by higher fixation rates (73.6%) and lower regression rates (7.9%) in MoTR compared to ET. Manual regressions are mechanically unnatural and require more physical effort, while the necessity of synchronizing hand and eye movement slows the reading pace, potentially facilitating better initial comprehension and reducing the need for backward movements.

Furthermore, it should be noted that this study did not include a direct performance comparison between the original, unmodified MoTR framework and our enhanced version. Consequently, while our results are promising, the specific contribution of each individual refinement—such as the vertical axis constraint or the gatekeeper system—remains to be empirically quantified in future ablation studies. Finally, this study focused exclusively on Romanian texts from the MultiplEYE corpus. Expanding this methodology to different writing systems, such as right-to-left (RTL) languages, would necessitate structural adaptations to the Row-Level Gatekeeper system to accommodate reversed reading directions and navigation patterns.

7. Ethical Considerations

Data collection followed the Declaration of Helsinki and GDPR standards. Participants provided written informed consent and were briefed on their right to withdraw. Raw data were pseudonymized at the point of collection, and no personally identifiable information was utilized during the modeling. Stimuli consisted of standardized MultiplEYE texts, appropriate for research and free from harmful material.

8. Acknowledgements

This work was supported by the European Cooperation in Science and Technology under COST Action CA21131 (MultiplEYE), by the project InstRead: Research Instruments for the Text Complexity, Simplification, and Readability Assessment CNCS - UEFISCDI project number PN-IV-P2-2.1-TE-2023-2007.

9. Bibliographical References

Stuart K. Card, Thomas P. Moran, and Allen Newell. 1983. The psychology of human-computer interaction.

Karin de Langis, William Walker, Khanh Chi Le, and Dongyeop Kang. 2025. Tracing how annotators think: Augmenting preference judgments with reading processes.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alba Haveriku, Sara Bedulla, Nelda Kote, and Elinda Kajo Meçe. 2025. *Understanding Reading Patterns of Albanian Native Readers Through Mouse Tracking Analysis*, pages 433–443. Springer, Cham.

Nora Hollenstein, Marie-Luise Müller, Deborah N. Jakobi, Cui Ding, Maja Stegenwallner-Schütz, Ana Matić, Eva Pavlinušić Vilus, Ramunė Kasperė, Anna Bondar, Maroš Filip, Stefan Frank, Jana Hofmann, Thyra Krosness, Kaidi Lõo, Johanne Nedergaard, Chiara Tschirner, and Lena A. Jäger. 2026. *MultiplEYE Data Collection Guidelines*.

Jeff Huang, Ryen White, and Georg Buscher. 2012. User see, user point: gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1341–1350, New York, NY, USA. Association for Computing Machinery.

Deborah N. Jakobi, Andreas Säuberli, Jana M. Hofmann, Carlson Büth, Anastassia Shaitarova, Daniel Krakowczyk, and Lena A. Jäger. 2026. The multipleye preprocessing pipeline. `https://github.com/MultiplEYE-COST/multipleye-preprocessing`. GitHub repository.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Ramunė Kasperė, Anna Bondar, Sergiu Nisioi, Maja Stegenwallner-Schütz, Hanne B. Søndergaard Knudsen, Ana Matić, Eva Pavlinušić Vilus, Dorota Klimek-Jankowska, Chiara Tschirner, Not Battesta Soliva, Deborah N. Jakobi, Cui Ding, Dima Abu Romi, Cengiz Acarturk, Matilda Agdler, Anton Marius Alexandru, Mohd Faizan Ansari, Annalisa Arcidiacono, Elizabete Ausma Velta Barisa, Ana Bautista, Lisa Beinborn, Yevgeni Berzak, Nedeljka Bjelanović, Anna Isabelle Bothmann, Jan Brasser, Caterina Cacioli, Anila Çepani, Ilze Ceple, Adelina Çerpja, Dalí Chirino, Jan Chromý, Alessandro Corona Mendozza, Iria de Dios-Flores, Nazik Dinçtopal Deniz, Ana Došen, Kristian Elersic,

Inmaculada Fajardo, Zigmunds Freibergs, Angelina Ganebnaya, Shan Gao, Jéssica Gomes, Annjo Klungervik Greenall, Alba Haveriku, Miao He, Anamaria Hodivoianu, Yu-Yin Hsu, Amanda Isaksen, Andreia Janeiro, Kristine Jensen de López, Aleksandar Jevremovic, Vojislav Jovanović, Hanna Kędzierska, Nik Kharlamov, Sara Košutar, Nelda Kote, Vanja Kovic, Izabela Krejtz, Thyra Krosness, Oleksandra Kuvshynova, Eilam Lavy, Ella Lion, Marta Łockiewicz, Kaidi Lõo, Paula Luegi, Mircea Mihai Marin, Clara Martin, Svitlana Matvieieva, Diane C. Mézière, Xavier Mínguez-López, Valeriia Modina, Jurgita Motiejunienė, Marie-Luise Müller, Tolgonai Nasipbek kyzy, Jamal Abdul Nasir, Johanne S. K. Nedergaard, Ayşegül Özkan, Patrizia Paggio, Marijan Palmović, Maria Christina Panagiotopoulou, Alberto Parola, Helena Pérez, Klaudia Petersen, Anja Podlesek, Eva Pospíšilová, Marta Prauliņa, Mikuláš Preininger, Loredana Pungă, Diego Rossini, Špela Rot, Habib Sani Yahaya, Irina A. Sekerina, Anne Gabija Skadiņă, Jordi Solé-Casals, Lonneke van der Plas, Saara M. Varjopuro, Spyridoula Varlokosta, João Veríssimo, Oskari Juhapekka Virtanen, Nemanja Vračar, Mila Vulchanova, Ahmad Mustapha Wali, Peizheng Wu, Nilgün Yücel, Stefan Frank, Nora Hollenstein, and Lena A. Jäger. 2026. The multipleye text corpus: Towards a diverse and ever-expanding multilingual text corpus. In *Proceedings of the 2026 International Conference on Language Resources and Evaluation (LREC 2026)*, Rabat, Morocco. European Language Resources Association and International Committee on Computational Linguistics.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004a. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004b. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology - EUR J COGN PSYCHOL*, 16:262–284.

Kozea. 2025. Pyphen: A pure python module to hyphenate text.

Sergiu Nisioi, Anna Bondar, Ramuné Kasperé, and Maja Stegenwallner-Schütz. 2026. The multipleye text corpus data and materials.

Metehan Oğuz, Cui Ding, Ethan Gotlieb Wilcox, and Zuzanna Fuchs. 2025. Using motr to probe agreement processing in russian. *Open Mind*, 9:1682–1710.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Cristina Maria Popescu and Sergiu Nisioi. 2025. Exploring mouse tracking for reading on Romanian data. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 44–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124 3:372–422.

Keith Rayner, Alexander Pollatsek, and Erik Reichle. 2003. Eye movements in reading: Models and data. *Behavioral and Brain Sciences*, 26.

Robyn Speer and Joshua Chin. 2021. rspeer/wordfreq: v3.2.

Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. 2025. Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27705–27726, Suzhou, China. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Cui Ding, Mrinmaya Sachan, and Lena Ann Jäger. 2024. Mouse tracking for reading (motr): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138:104534.

10. Language Resource References

Lena A. Jäger, Nora Hollenstein, Ana Matić Škorić, Deborah N. Jakobi, Maja Stegenwallner-Schütz, Cui Ding, Eva Pavlinušić Vilus, Ramunė Kasperė, and Marie-Luise Müller. 2026. Multipleye: Enabling multilingual eye-tracking data collection for human and machine language processing research.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European Conference on Eye Movements (ECEM)*.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833. Available at https://osf.io/sjefs.

Eye-Contact and Facial Expression Tracking for Assertiveness Training in VR-Based Anti-Bullying Education

Lubomir Ivanov
Iona University
715 North Avenue
New Rochele, NY 10801
livanov@iona.edu

Mary Vrahimis
Iona University
715 North Avenue
New Rochele, NY 10801
mvrahimis@iona.edu

Anabel Nolasco
Iona University
715 North Avenue
New Rochele, NY 10801
anolasco@iona.edu

Abstract

This paper described the use of eye-contact and facial expression tracking as part of a comprehensive approach to assertiveness training in a VR-based anti-bullying simulation environment. We briefly discuss the psychological foundations of assertiveness and then focus on our approach to tracking the facial expressions and eye-contact that a user maintains while communicating with the virtual bully in the simulation. We also outline additional non-verbal indicators tracked by the software and discuss the dialog system, which drives the simulation. Finally, we outline some ethical considerations, discuss the limitations of our current software prototype, and list future directions for enhancing assertiveness training in anti-bullying education.

Keywords: eye-contact, emotion tracking, personal space, voice level, AI dialog, assertiveness, anti-bullying, VR

1. Introduction

Bullying is a pervasive societal problem impacting people of all ages, genders, and ethnicities. It is characterized by a pattern of repeated intentional physical or psychological assault or abuse of another individual or a group of individuals. At its core, bullying is based on a real or perceived imbalance of power between the bully and the victim(s). Bullies “desire power and dominance over their peers” (Olweus, 1997). Victims tend to be sensitive, introverted, and lack assertiveness, which “leaves them lonely and increases their insecurities“ (Olweus, 1997). Many studies have demonstrated that one of the most effective ways of addressing bullying is at a young age, before the harmful traits of bullies (cruelty, aggressiveness, domineering, lack of empathy, etc.) and victims (insecurity, anxiety, low self-esteem, etc.) become engrained into children’s personalities (Bijttebier & Vertommen, 2017; Cornell & Limber, 2015; Cornell & Mehta, 2011; Eslea & Smith, 1998; Gaffney, Farrington, & Ttofi, 2019; Gaffney, Ttofi, Farrington, 2019; Griffin & Gross, 2004; Olweus, 1997; Olweus & Limber, 2010). Unfortunately, decades of traditional psychological approaches have yielded very limited results. In recent years, software-based approaches to anti-bullying prevention and education have proliferated (BRIM; GoSpeakUp; HIBster; Stavroulia et al, 2016). Among those, virtual reality (VR) based anti-bullying projects have taken a lead role by providing an immersive educational experience (Barreda-Ángeles, 2021; Ingram et al, 2019; Ivanov & Ramos, 2020; Ivanov 2022; Stavroulia et al, 2016; ClassVR; VR Action Lab; Upstander). VR-based anti-bullying education has been demonstrated to be at least as effective as traditional methods and much more engaging for the participating students.

The most important aspect of anti-bullying education is training users to maintain confidence and assertiveness during a bullying confrontation. Assertive behavior is exhibited through a combination of behavioral traits such as keeping a firm posture at an acceptable social distance, using appropriate responses, delivered in a level tone of voice, and maitaining proper eye-contact.

This paper describes our approach to VR-based assertiveness training for anti-bullying education focusing on eye-contact- and facial expression tracking and their interaction with other assertiveness traits. We begin by briefly discussing the psychological underpinnings of assertiveness and explaining the architecture of our anti-bullying environment. We then describe the implementation of user eye-contact- and facial expression tracking and their relation to the NLP dialog system, which drives the simulation scenarios. The integration of eye-contact and facial expression tracking with other non-verbal traits tracking is briefly discussed. Finally, we outline directions for further research.

2. Psychological Aspects of Assertiveness

Assertiveness is an essential element of interpersonal interaction, the foundation of healthy relationships, successful careers, and the ability to stand up for oneself in the face of adversity. In the context of bullying, projecting self-confidence and assertiveness is a vital skill for resolving a bullying incident (Boket et al, 2016)

Assertiveness is a communication style, which allows a person to express their opinions or feelings or to stand up for their rights while respecting the rights, feelings, and opinions of others. In contrast, an aggressive person puts their own feelings, opinions, and rights exclusively above those of others, while a passive/submissive

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 50–57
May 12, 2026.

person puts the rights, opinions, and feelings of others above their own.

Assertive communication has several traits:

- Maintaining proper eye-contact with the other person while communicating
- Preserving a calm, neutral facial expression
- Respecting the personal space of the other communicator
- Speaking in a calm, level tone of voice
- Using appropriate vocabulary and avoiding the use of words and phrases that can be construed as aggressive or passive

A well-structure assertiveness training program should emphasizes the combined use of these traits in interpersonal communication to help individuals learn to express their thoughts and ideas with a high degree of self-confidence (Eslami et all, 2014, Gündoğdu, 2012)

3. Software Achitecture

Our project implements a complex, first-person, VR environment for anti-bullying education of pre-teens (ages 8 through 12). The software is implemented using the (Unity) game engine and consists of several modules, which present the user with different types of bullying challenges involving one- or multiple bullies (Figures 1 & 2).

Figure 1: One-on-one bullying scenario.

In some scenarios, bystanders may be present, who can be either sympathetic to the victim or side with the bully. The dynamics of the interaction and the strategies for a successful resolution depend on the number of bullies and the presence/absence and type of bystanders. Additional factors that the user must consider are the presence/absence of adults, who the user can ask for help, and the environment in which the bullying incident occurs. For example, the optimal strategies are different depending on whether the user is backed into a corner or standing in the open, where escape is a viable option.

Figure 2: Multiple bullies bystander scenario.

In addition to teaching children strategies for resolving bullying incidents, our software provides modules for training bystanders: These modules provide the user with the opportunity to experience bullying from a different perspective and learn essential techniques to assist a victim in diffusing a bullying situation.

The initial prototype of our software is specifically designed for the training of pre-teen girls, who are subject primarily to verbal bullying. Physical bullying, which is much more common among young boys, is not modeled, though the user has the option to “punch” the bully. Doing so, however, leads to an unsatisfactory resolution of the bullying incident with both the victim and the bully being reprimanded by a virtual adult.

4. Tracking Assertiveness

Tracking the user’s assertiveness during a bullying incident requires the integrated, continual monitoring of all five assertiveness traits – proper eye-contact, calm facial expression, respect for personal space, use of positive, non-aggressive language, and an even tone of voice.

4.1 Eye-contact

Eye-contact is one of the most important aspects of assertive communication. It is a key indicator of the power dynamic between the individuals in real-life communication. Psychological studies have demonstrated that normal interpersonal communication is commonly bound by the “50-70 rule”: A communicator should maintain eye-contact 50% of the time while speaking and 70% of the time while listening. Moreover, normal eye-contact should be maintained for 3 to 5 seconds at a time, and then the communicator should look away before resuming eye-contact. Looking away too quickly and not maintain steady eye-contact while communicating can be perceived as a lack

of self-confidence or even dishonesty, while maintaining uninterrupted eye-contact for more than 5 seconds at a time while speaking can be perceived as aggression. In the context of bullying, the rules of normal communication need to be altered in order to project strength and assertiveness in the face of aggression. Thus, it is necessary to maintain eye-contact more than 50% of the time while speaking: 3 to 5 seconds of eye-contact followed by only a brief (0.5s to 2s) look-away period before eye-contact is resumed.

Our implementation of the tracking of the user's eye-contact with the virtual bully uses VR raycasting: An invisible ray (i.e., an unrestricted-length Vector3 object) is projected forward from the center point between the user's virtual eyes. The ray moves relative to the position and rotation of the parent object (i.e., the user's VR headset). A trigger collider component is added around the eyes of the bully character. When the ray intersects the collider, a 5 second timer is initiated. If the user's eyes do not shift away before the timer fires, a small yellow eye icon (Figure 3) is displayed to remind the user to look away. If the user's gaze does not shift away from the bully for another second, a red icon is displayed as a final warning. If the user's gaze still does not shift away from the bully's "eyes", the simulation scenario is terminated.

Figure 3: Yellow eye-contact reminder icon.

If the user's gaze shifts away from the bully's "eyes" before the 5 seconds timer expires, then the current timer value is compared to a predefined minimum (e.g., 3 seconds). If the timer value is smaller than the minimum, the yellow eye icon is displayed and blinks to remind the user he/she has not maintained their gaze on the bully long enough. If the timer value is within the specified range (3s to 5s), then the timer is reset with a 2 second value - the user must return their gaze back to the bully's "eyes" before the timer runs out or risk being perceived as passive/submissive by the bully. If the user does not look up, then after 2 seconds, the yellow icon reminder is displayed again, followed by the red icon (Figure 4) reminder one second later. If the user still does not look up, the simulation is terminated with an appropriate explanation.

Figure 4: Eye-contact icons.

When looking away, the user must not look straight down, which can be perceived by the bully as a sign of weekness or submissiveness. To track this feature of eye-contact, a different trigger collider is placed directly in front and below the eye level of the user. If the user looks straight down, the raycast intersects this collider, triggering the display of a diffent attention icon – one with two arrow pointing sideways (Figure 5). This is a reminder to the user to look to either side instead of straight down.

Figure 5: "Look to the side" icon.

The eye-contact interaction timers are active only while the user or the bully are speaking. If no verbal communication is occurring, the user is free to look around.

4.2 Emotive Facial Expressions

Facial expressions are a form of nonverbal communication using the movement of facial muscles. Along with eye-contact, facial expressions constitute a crucial aspect of communication since they often accurately reflect an individual's emotional state (American Psychological Association**)**. We focused on seven universal emotions - anger, happiness, sadness, fear, contentment, disgust, and surprise (Stichter et al, 2011). Each of those emotions is expressed by facial expressions, which result from the movement of specific facial muscle groups – lowering or raising of the eyebrows, squinting or opening of the eyelids, curving the mouth corners up or down, etc.

Conveying emotion through facial expressions not only serves the purpose of expressing feelings but also influences the behaviors of others (Frontiers). For instance, if the facial expression of one communicator displays anger, then the

other communicator may react with fear. Moreover, different facial expressions and gaze patterns can alter the interpretation of a person's speech. For example, if an individual verbally expresses something positive but their facial expression indicates disgust or anger, then the overall message may be perceived as sarcasm. Speakers often produce visual cues that demonstrate their confidence level. Speakers with lower confidence are more prone to exhibit a distressed facial expression often accompanied by an averted gaze. Understanding the interplay between verbal and nonverbal cues is essential to the accurate modeling of natural communication.

Children are especially sensitive to non-verbal emotional communication. In (Fox et al, 2017), the correlation between facial emotion recognition skills and the behavior of young adolescents in bullying situations was investigated. It was revealed that fear was highly recognized and could potentially aid aggressive individuals in identifying vulnerable victims. In situations where no bystanders are present, fear, anger, sadness, and disgust might empower the bully by signaling weakness in a victim. Conversely, if bystanders are present, the detection of fear and sadness may elicit empathetic concern for the victim and increase the likelihood of bystander intervention.

Our VR anti-bullying environment employs the eye-tracking and facial expression tracking features of the Meta Quest Pro headset and Meta's (Movement SDK) to process in real-time the user's emotional reactions to the bullying conversation and notify the user if the displayed emotion is counterproductive to resolving the bullying incident.

For each of the universal emotions from (Stichter et al, 2011), we developed numerical range representations. To do this, we used the Meta's avatar "Aura" (Figures 6) to approximate the facial expressions corresponding to the basic emotions by adjusting the avatar's blendshapes including eyebrow raising/lowering, jaw dropping, and lip-, cheek- and eye movements. For each facial expression, we recorded the weight ranges of every adjustable blend-shape connected to the skin meshed renderer on the avatar's face. Establishing these weight ranges involved careful considerations: Each facial expression is a combination of facial movements occurring simultaneously. Thus, we had to ensure that in each case the weights were strong enough to avoid interference with other expressions sharing similar facial movements. Additionally, we grouped expressions in a manner that prevents activation through speech or subtle facial movements. Additional finetuning was required once we started testing against actual human expressions during the simulation.

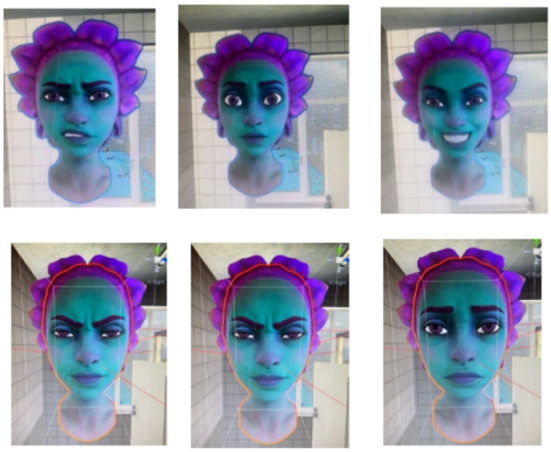


Figure 6: Facial expressions expressing different emotions

At run time, we use the Meta Quest Pro's five inward-facing infrared cameras and Meta's Movement SDK to track the actual facial expressions of the user, comparing the recorded values to predefined emotional range sets. When a facial expression, which is counterproductive to the anti-bullying effort is detected, a small icon is displayed to remind the user to control their emotions as best as possible (Figure 7).

Figure 7: Emotion icon displayed to the user

4.3 Other Non-Verbal Features

4.3.1 Personal Space

Maintaining an appropriate distance and respecting the personal space of others is an essential aspect of communication. In the 1960s, the anthropologist Edward Hall conducted extensive research on the relationship between communication distance and the type of communication (Hall, 1963). He defined four types of separation during communication:

- *Intimate distance*: from less than an inch to about 18 inches
- *Personal distance*: approximately 2 to 4 feet
- *Social distance*: approximately 4 to 12 feet
- *Public distance*: 12 to 25 feet or more

The distances between the bully and the victim during a bullying incident are usually dictated by the bully: Typically, the initial confrontation begins at a social distance of 8-12 feet, but if the confrontation escalates, the distance can quickly decrease to less than a foot.

It is crucial that the victim attempts to preserve a social- or at least a personal distance during the confrontation. This is a delicate balance: If the victim steps too far back, the move can be perceived by the bully as a sign of weakness. On the other hand, the victim's stepping forward or refusal to step back can be perceived as aggressive and lead to a further escalation of the conflict. The victim must attempt to preserve the distance established at the beginning of the conflict, taking a step back or sideways only if the distance shrinks rapidly and stepping forward only after the incident has been resolved.

Our implementation of personal space tracking involves the use of trigger colliders. A minimum distance collider is placed around the victim. The radius of the collider is based on the bullying scenario: In one-on-one bullying, the minimal distances are usually 2ft to 4ft. In many-on-one bullying, the minimal distance is somewhat larger (3ft to 6ft) and depends on the number of bullies and bystanders. Each bully is programmed to move based on the state of the bullying dialog: If the exchange escalates, the bully is programmed to take one or more steps towards the user. If the dialog moves towards an acceptable resolution, the bully is programmed to take a step back. The user is expected to attempt to keep the bully outside the minimum distance. If a bully crosses into the minimum distance collider, a small red icon is displayed (Figure 8) promtping the user take a step back until the bully is outside the minimum distance collider. A second, larger trigger collider around the player displays a yellow icon if the player backs off too far from the bully or attempts to flee (Figure 9). The radius of that collider is usually set to 8ft. The maximum distance collider is disabled in certain training scenarios when escape is a viable alternative. In general, however, the player must keep the bully between the larger (max-distance) collider and the smaller (min-distance) collider until the conflict has been resolved.

Figure 8: Red icon prompts the user to step back

Figure 9: Yellow icon prompts the user to stop backing up and take a step forward

4.3.2 Tone of Voice

Assertive communication requires an even, sufficiently loud tone of voice. A quiet voice can be construed as a sign of weakness, whereas too loud a voice can be perceived as aggressive. Additionally, excessive fluctuations in the vocal pitch and loudness can be a sign of emotional distress. A confident voice exhibits relatively small, smooth fluctuations in pitch and loudness.

We use a script to capture live feed from the headset's microphone and measure the loudness and frequency spectrum of the user's voice: When the user approaches the bully, the script begins capturing sets of microphone input samples. The microphone signal is passed through a low pass filter to cut out background noise and frequencies that do not correspond to human speech (80Hz to 255Hz for adults and 250Hz to 400Hz for children). We use the GetSpectrumData() and GetData() methods of Unity's Audiosource and AudioClip classes to obtain the loudness and frequency data from the microphone samples. The loudness data is used to control a small volume bar icon, which changes color from green

(appropriate volume) to yellow (slightly low/high) to red (very low/high). The scale of the sound bar shows if the user is talking too quietly (partially-filled red or yellow bar) or too loudly (completely full red or yellow bar) (Figure 10).

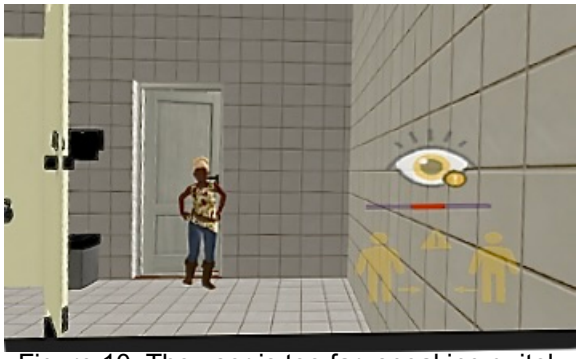


Figure 10: The user is too far, speaking quitely and looking at the bully too long.

4.4 The Verbal Communication System

The most important aspect of any anti-bullying training environment is the dialog system. While impactful, non-verbal cues usually cannot, by themselves, resolve a bullying situation. Therefore, it is imperative to implement a verbal communication system, which teaches the user how to succesfully deal with bullying incidents.

Most anti-bullying education has traditionally revolved around scripted dialogs, where the user is taught how to respond to specific bullying taunts. Our approach is to let the user try, and fail, and eventually succeed in resolving the situation on their own. After each session, we provide guidance as to how the situation could have been handled better. But the goal is to let the user determine for themselves what works.

To implement the verbal communication system we use the Microsoft (Speech SDK) to provide real-time text-to-speech (tts) and speech-to-text (stt) functionality and OpenAI's (ChatGPT) to act as a "bully" as well as to evaluate the user's responces. The conversation is initiated by the "bully" when the user crosses the bully character's proximity collider. We prompt ChatGPT to generate a child-like taunt, which is then passed to the tts system to play the taunt to the user. The system waits for the user to respond and, once the response is received, it is converted to text and passed to ChatGPT along with a prompt "In one word, is this sentence positive, negative, or neutral". Depending on the estimated polarity of the user's response, ChatGPT is instructed further to act either in a more adversarial- or a more friendly manner. The dialog is usually short – no more than 5-6 bully-user exchanges.

The user's success or failure is judged by a point system which combines the weighted scores from the verbal exchange as well as from the various non-verbal cues displayed during the simulation. The point systems takes into account the interaction between the verbal and non-verbal communcation: For example, if the user responds positively to a bully taunt while exhibiting a positive or neutral facial expression, an additional positive point is awarded. Conversely, if the positive response is accompanied by a negative facial expression, an extra point is subtracted.

5. Conclusion

This paper describes the methodologies used for tracking of user assertiveness and teaching assertive communication skills in a complex VR-based anti-bullying training environment. The software integrates verbal and non-verbal interaction tracking including eye-contact, facial expressions, personal space, tone of voice, as well as appropriate verbal commumication. Each of these topics is very complex and requires a significant amount of future adjustments.

Much work remains to be done on refining our tone-of-voice algorithms. Notably, frequency spectrum inconsistencies and loudness fluctuations can be due to overlapping background noise frequencies as well as vocalization artifacts such as exclamations and prosodic pauses.

Refinements will be added to the eye-contact tracking methodology as well: When facing multiple bullies, it is important that the victim keeps shifting their gaze from one bully to the next rather than focusing on a specific bully. Doing so is even more important in the presence of bystanders since engaging bystanders through eye-contact is likely to establish a more personal connection and help bring the bystander to the victim's side. Implementing an eye-contact tracking methodology which encourages gaze shifting is one of the next goals of our project.

Eye-contact and facial expressions tracking are not the only aspects of body-posture assertiveness. Gestures such as crossing one's arms, leaning forward, or slumping are implicit expressions of assertiveness, aggression, or submissiveness. The limited availability and high cost of full-body VR suits makes integrating the tracking of such gestures infeasible for the moment. However, one of our immediate next steps will be adding the tracking of hand-gestures, such as raising one's fists.

Finally, we hope to expand the use of our assertiveness tracking methodology beyond anti-bullying training. Assertiveness is a valuable trait in any human interaction and creating a stand-alone software environment to teach assertiveness in different social settings (e.g., negotiating for a salary raise, asking someone on a date, or debating a topic with friends) will be a beneficial endeavor for society as a whole.

6. Ethical Considerations and Limitations

Bullying is a highly complex and sensitive topic. Any anti-bullying methodology needs to be carefully vetted by one or more qualified institutional review boards (IRB) and undergo a thorough testing by qualified professionals. Large-scale deployment should only be attempted after a series of smallest test studies involving limited groups of testers, beginning with psychology experts and eventually moving towards target audience participants. Our software is still in its prototype phase and a review by our IRB is still pending.

There are many areas of concern: Foremore among them is the reliability of Generative AI acting as a simulated bully. While the ChatGPT prompts that drive the simulation are carefully vetted, there are no assurances as to what responces ChatGPT (or any other LLM) will generate. It is, therefore, virtually impossible to guarantee that a particular response will not be inappropriate or harmful, especially to a child participant.

Any bullying situation – even a simulated one – can be very stressful to some individuals and even more so to children. Part of the reason we opted for VR instead of an augmented/mixed reality simulation was because some of the stress can be mitigated by using characters and environments that are intentionally designed to look somewhat cartoonish and not hyper-realistic. Adopting this design approach provides the simulation with a more game-play atmosphere that aleviates some of the stress of participating in the training. Regardless, it is highly recommended that a qualified professional – either a psychologist or a school counselor – observe the training session and be ready to intervene if the participant exhibits signs of strong emotional distress.

It is well known that VR causes vertigo in some individuals due to vection - the (often subcontious) illusion of motion while seated. Our simulation software has been designed with the aim of minimising VR discomfort for the user. We provide smooth motion, avoiding sudden/jerky camera movements, and do not use teleportation. We do our best to avoid bringing up peripheral visual stimuli, though this is occasionally difficult or impossible to do given the complex nature of the simulation environement, involving numerous virtual characters and moving objects. Once again, the vigilance of a school psychologist or counselor is necessary to ensure that the participants do not suffer from the effects of VR-induced motion sickness.

7. Bibliographical References

American Psychological Association. "Facial Expression." APA Dictionary of Psychology, dictionary.apa.org/facial-expression.

Barreda-Ángeles M, Serra-Blasco M., Trepat E., Pereda-Baños A., Pàmias M., Palao D., Goldberg X., Cardoner N., 2021, Development and experimental validation of a dataset of 360°videos for facilitating school-based bullying prevention programs, Computers & Education, Volume 161, 104065, ISSN 03601315

Bijttebier, P., Vertommen, H. 2017. Coping with Peer Arguments in School Age Children with Bully Victim Problems. In British Journal of Educational Psychology 68 (1998): 387. ProQuest.

Boket EG, Bahrami M, Kolyaie L, Hosseini SA. 2016. The effect of assertiveness skills training on reduction of verbal victimization of high school students. Int J Hum Cultur Stud.2016;3:2356–5926

BRIM Anti-bullying Software (last visited: Feb.7, 2026): https://antibullyingsoftware.com/

ChatGPT (last visited: Feb.7, 2026): https://chatgpt.com/

Cornell, D., & Limber, S. P. 2015. Law and policy on the concept of bullying at school. American Psychologist, 70, 333. doi: /10.1037/a0038558

Cornell, D., & Mehta, S. B. 2011. Counselor confirmation of middle school student self-reports of bullying victimization. Professional School Counseling,14(4), 2156759X1101400402

Eslami AA, Rabiei L, Afzali SM, Hamidizadeh S, Masoudi R. 2016. The Effectiveness of Assertiveness Training on the Levels of Stress, Anxiety, and Depression of High School Students. Iran Red Crescent Med J. 18(1):e21096. doi: 10.5812/ircmj.21096. PMID: 26889390; PMCID: PMC4752719

Eslea M., Smith P. 1998. The long-term effectiveness of anti-bullying work in primary

schools, Educational Research, 40 (2), pp. 203218, 10.1080/0013188980400208
Fox, Nathan A., et al. 2017. "Emotional Expressions and Visual Awareness: A Comment on Yang et al. (2017)." National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/pmc/articles/PMC5683572
Frontiers, "The Interpersonal Effects of Emotions: The Influence of Facial Expressions on Social Interactions." Frontiers Editorial Office. www.frontiersin.org/research-topics/23358/the-interpersonal-effects-of-emotions-the-influence-of-facial-expressions-on-social-interactions
Gaffney H., Farrington D., Ttofi M. 2019. Examining the effectiveness of school bullying intervention programs globally: A meta-analysis, International Journal of Bullying Prevention, 1 (1), pp. 1431, 10.1007/s4238001900074
Gaffney H., Ttofi M., Farrington D. 2019. Evaluating the effectiveness of school bullying prevention programs: An updated meta-analytical review, Aggression and Violent Behavior, 45, pp. 111133, 10.1016/j.avb.2018.07.001
GoSpeakUp (last visited: Feb.7, 2026): https://www.gospeakup.com/pri_sec/
Gündoğdu, R. 2012. Effect of the Creative Drama-Based Assertiveness Program on the Assertiveness Skill of Psychological Counsellor Candidates. Educational Sciences Theory and Practice.
HIBster (last visited: Feb.7, 2026): https://hibstersafe.com
Hall, E. 1963. "A System for the Notation of ProxemicBehavior". *American Anthropologist* . **65** (5):1003–1026. doi:10.1525/aa.1963.65.5.02a00020.
Ingram K., Espelage D., Merrin G., Valido A., Heinhorst J., Joyce M., 2019, Evaluation of a virtual reality enhanced bullying prevention curriculum pilot trial, J. of Adolescence, 71, pp.7283,10.1016/j.adolescence.2018.12.006
Ivanov L., Ramos N., 2020, Bully: A Virtual Reality Environment for Anti-bullying Education, Proceedings of FLAIRS33 Conference, Miami Beach, FL
Ivanov L., 2022. Optimizing the User Experience in VR-based Anti-Bullying Education, UMAP'22 Conference, Barcelona, Spain
Meta Movement SDK (last visited: Feb.9, 2026): https://developers.meta.com/horizon/documentation/unity/move-overview/
Microsoft Speech SDK (last visited: Feb.9, 2026): https://www.microsoft.com/en-us/download/details.aspx?id=10121
Olweus, D. 1997. Bully/Victim Problems in School: Facts and Intervention. European J. of Psychology of Education 12.4: 495510.
Olweus D., Limber S., 2010, Bullying in school: Evaluation and dissemination of the Olweus bullying prevention program, American Journal of Orthopsychiatry, 80 (1), pp. 124134, 10.1111/j.19390025.2010.01015.x
Stavroulia et al, 2016. A 3D virtual environment for training teachers to identify bullying. In 18th Mediterranean Electrotechnical Conference (MELECON), Lemesos, Cyprus, DOI: 10.1109/MELCON.2016.7495417
Stichter, Janine, et al. 2011. Teaching Facial Expressions of Emotion. ResearchGate, www.researchgate.net/publication/243963294_Teaching_Facial_Expressions_of_Emotion.
Unity: https://unity.com
Upstander (last visited: Feb.7, 2026): https://www.meta.com/community/vr-for-good/upstander/
VR Action Lab by Harmony Labs (last visited: Feb.7, 2026): https://harmonylabs.org/

Predicting Gaze Location without Camera or Eye-Tracker

Saman Rezapoor,* **Sajad Shirali-Shahreza,*†** **Gerald Penn†**
*Amirkabir University of Technology
350 Hafez Ave., Tehran, IRAN
saman.rezapour1379.sr@gmail.com, shirali@aut.ac.ir

†University of Toronto
4283-40 St. George St., Toronto, CANADA
gpenn@cs.toronto.edu

Abstract

The task of identifying the location that a user looks at, commonly known as gaze estimation, has various HCI and NLP applications. Traditional gaze estimation methods use special hardware such as eye-trackers or ordinary cameras such as webcams to perform this. However, they are not applicable to the majority of web users either because the user does not have them or does not want to use them due to privacy reasons. In this paper, we propose the idea of using multimodal LLMs to analyze the content of the user's screen along with mouse location to estimate the gaze location. It primarily uses the results of studies that extract common reading patterns such as the F-pattern and Z-pattern. Our experimental results on The Eye Of The Typer (EOTT) dataset provide promising results for estimating gaze location.

Keywords: Gaze Estimation, LLM

1. Introduction

Reading online sources such as news, documentation, and long-form articles has become a dominant mode of everyday information consumption. Estimating *where the user is looking* on a webpage, which is usually referred to as gaze estimation or prediction, provides a direct signal of attention. This will enable a range of HCI and NLP applications, including saliency modeling (Buscher et al., 2009), adaptive interfaces, and improved placement (or suppression) of distracting elements(e.g., advertisements) (Owens et al., 2011).

Gaze prediction data also offer a valuable supervision signal for language understanding tasks, because reading behavior reflects how users allocate attention across words, headings, and images. Despite its value, gaze estimation remains difficult to deploy at scale: high-quality eye trackers are expensive, sensitive to setup conditions, and rarely available in naturalistic browsing settings.

To reduce cost and deployment friction, previous work has studied *proxies and constraints* for gaze estimation while the user is browsing. Reading often produces structured scan-paths shaped by page layout (commonly described by patterns such as *F-shaped* or *Z-shaped* scanning). These regularities can be leveraged to constrain gaze inference (Soegaard, 2021; Lorigo et al., 2008). Other studies have analyzed the relation between gaze location and interaction logs. They showed that mouse movements and clicks correlate with attention during navigation and reading (Huang et al., 2012; Navalpakkam et al., 2013). They used classical machine learning models to predict gaze from such signals.

Webcam-based systems such as WebGazer[1] further illustrate that gaze can be predicted without dedicated hardware, though accuracy can vary with calibration and user/environmental factors (Papoutsaki et al., 2018).

Motivated by these precedents, we tried to determine whether *multimodal large language models (LLMs)* can predict gaze in a naturalistic browsing setting *without* any special hardware such as eye trackers or even webcams. Our idea focuses on previous findings in regard to reading behaviour and temporal continuity. We report our preliminary results in this paper. Concretely, our contributions are:

1. Prompt-driven gaze prediction: encoding reading scan-pattern priors and temporal smoothing constraints into a system prompt and enforcing *structured JSON outputs* (coordinates, attention-pattern label, reasoning-mode label, and confidence).
2. Two evaluation conditions: *full-video* prediction from screenshot sequences, which does not need any actual gaze location data, and a *per-frame setting* that conditions each prediction on the actual gaze location in the previous frame. [2]

[1] webgazer.cs.brown.edu/

[2] The second condition aims to remove the effect of previous prediction errors (accumulated error) in predicting the gaze in the next frame, i.e., isolating the effect of temporal anchoring.

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 58–63
May 12, 2026.

3. Quantitative metrics and analysis: We report *pixel-level error* and *region-level accuracy*, and analyze how model-reported confidence relates to error; this includes failures in low-information cases such as empty and calibration pages.

2. Related Work

Our work lies at the intersection of (i) mouse-based proxies for gaze and attention, and (ii) learning-based models of gaze and reading behavior. We briefly review both areas and position our contribution in relation to them.

2.1. Mouse Proxy for Gaze/Attention

Human-computer interaction has long grappled with the challenge of whether the cursor trace, including mouse movements, hovers, and clicks, can be used to estimate where the user is looking and what they are attending to.

Huang et al.'s (2012) work on web search concerns the context dependency of gaze and cursor alignment and demonstrates how gaze prediction benefits from the inclusion of behavioral features rather than just the cursor position. Perhaps most interestingly, their results suggest that the lag between the two is consistent, with the cursor lagging behind the gaze by close to $700ms$ on average, which in turn supports the intuitive idea that users tend to look ahead of their mouse movements (Huang et al., 2012; Milisavljevic et al., 2021).

Related results have also been presented in search and browsing settings where eye and mouse movements are modeled explicitly over time and page structure. Navalpakkam et al. (2013) demonstrate that, in non-linear page structures, mouse movements can be used for predicting attention patterns, even though they also exhibit systematic discrepancies depending on the task and state of interaction. Guo and Agichtein (2010) also present results on gaze prediction based on mouse movements in web search settings. They emphasize the importance of inferring the dynamic coordination between gaze and mouse movements.

More recently, Popescu and Nisioi (2025) propose the Mouse Tracking for Reading (MoTR) method for predicting the reading time. They blurred the text except for under the mouse cursor, which allows them to estimate reading times based on mouse movements. The authors show that the reading times obtained with MoTR capture standard psycholinguistic effects and are predictable using lexical features and transformer models. Although MoTR provides evidence for the viability of mouse tracking for reading analysis, it does not directly validate mouse traces against simultaneous eye tracking, nor does it attempt gaze estimation. Therefore, it is not trying to predict gaze from mouse tracking data, which is our main idea in this paper.

2.2. Learning-based modeling of gaze and reading behavior

A second line of work involves the use of machine-learning methods for modeling eye-tracking outcomes from linguistic and interactive features. Alves (2025) benchmarks LLM-based methods for predicting eye-tracking reading-time measures (first-fixation duration, gaze duration, total fixation time). Their results show a high variance in predicting such values. That work focused on temporal effort signals (durations), however, rather than predicting the gaze (spatial gaze coordinates) on rendered pages.

Another proposed idea for performing scalable gaze estimation without the use of eye-tracking devices is to use webcams for gaze prediction. Papoutsaki et al. (2016) demonstrate the viability of the method using the WebGazer tool, which utilizes self-calibration based on user interactions for the estimation of the gaze location using commodity-grade webcams. They also released The Eye Of The Typer Dataset (EOTT)[3], which provides a collection of synchronized screen recordings, mouse movements, and eye-gaze locations for different tasks (Papoutsaki et al., 2018).

Recently, Ahmadzadeh (2024) used classical supervised methods to predict gaze location or areas-of-interest (AoI) using interaction traces (mouse and keyboard events) only. The results show that **predicting AoI** is much more accurate than predicting exact coordinates. The present paper is one example of such work (Ahmadzadeh, 2024).

2.3. Positioning of the present work

In comparison with related gaze prediction studies, our approach does not rely on a uniform mouse-gaze mapping but rather uses mouse traces as an additional cue. It emphasizes layout-driven reading priors and temporal continuity. Compared with ML/NLP methods that predict reading-time-related quantities (Michaelov and Levy, 2026), we predict spatial gaze location from screenshots. We use two metrics to measure our accuracy: pixel-wise error and region-wise accuracy. As to methodology, we also examine how the accumulated errors of multimodal LLMs during gaze prediction reduce accuracy over the duration of the task.

[3] https://webgazer.cs.brown.edu/data/

3. Our Proposed Method

We propose a prompt-driven gaze prediction framework that targets an article-reading scenario. The framework (i) specifies attention-pattern priors and temporal smoothing constraints through an LLM prompt, and (ii) evaluates the resulting model behavior on a sequence of screenshots sampled from long interaction traces. Because the prompt is tailored specifically article reading, we filter the available frames to article-like pages and then obtain model predictions for each frame in chronological order.

3.1. Dataset

We base our experiments on **The Eye of the Typer (EOTT)**[4] dataset, released with Papoutsaki et al. (2018). This dataset provides synchronized screen recordings, eye-tracing logs (i.e., actual gaze location), and interaction traces from 51 participants during multiple web-based tasks.

From the full dataset, we randomly selected 3 users and used the per-user metadata in `Participant_Characteristics.csv` to obtain session durations and native screen resolutions. The gaze coordinates are represented in pixel space in the dataset logs.

3.2. Extraction and synchronization pipeline

To create the input prompts for the multimodal LLM, we created a pre-processing pipeline based on the EOTT "dataset extractor"[5] utility. We adapted it to extract (i) screenshots, (ii) gaze samples aligned to each screenshot, and (iii) mouse events rendered directly onto the screenshot images.

3.2.1. Screenshot sampling

We extracted video frames as screenshots at a fixed 5-second interval. This rate was selected to avoid excessive oversampling while still preserving coarse reading progression. We selected 50 screenshots in total from each participant. The main emphasis was on segments where participants were reading article-like pages, plus a small number of sparse/empty pages (e.g., calibration or low-content transitions) to test robustness under low visual density.

3.2.2. Gaze log alignment

For each screenshot captured at timestamp t, we selected gaze samples from `[Dataset_Name].txt` whose timestamps fall within a temporal window of $\pm 500\text{ms}$ around t. The gaze stream in our extracted logs is sampled at 60 Hz, as stored in the dataset traces used by our pipeline. For each screenshot, we retained up to 5 gaze points closest to t (a fixed window size), yielding a compact set of candidate gaze locations per frame.

3.2.3. Mouse event extraction and rendering.

Mouse events were read from `[Dataset_Name].json`. We rendered interaction traces onto screenshots using OpenCV[6]: (i) a polyline for mouse movement and (ii) a filled red circle at click locations. This augmentation was motivated by earlier findings that mouse activity can correlate with visual attention in some interaction modes, while remaining an imperfect proxy overall (Huang et al., 2012; Navalpakkam et al., 2013).

3.3. Reference gaze per screenshot via geometric median

Because raw gaze samples may include transient noise, drift, or occasional rapid eye saccades, we aggregate the windowed gaze points into a single *reference gaze coordinate* per screenshot using a geometric median (i.e., Fermat–Weber point). Concretely, for gaze samples $\{p_i\}_{i=1}^{n}$ in a screenshot window, we estimate:

$$\hat{p} = \arg\min_{p} \sum_{i=1}^{n} \|p - p_i\|_2.$$

We compute $\hat{p}$ using Weiszfeld's algorithm, iterating until it converges (Beck and Sabach, 2015).

We selected the geometric median instead of an arithmetic mean because it is more robust to outliers: when a participant's gaze briefly "jumps" (e.g., due to a saccade or tracking noise), the median reduces the influence of those samples relative to the mean. In our implementation, this robustness serves as an implicit outlier-mitigation step rather than applying a separate rejection rule. [7]

3.4. Prompted gaze prediction with multimodal LLMs

We evaluate two different multimodal LLMs as **prompted gaze predictors** conditioned on screenshot content and interaction overlays. Our prompts

[4] https://webgazer.cs.brown.edu/data/

[5] https://github.com/brownhci/WebGazer/

[6] https://opencv.org/

[7] Indeed, another view of gaze, as pointed out by an anonymous reviewer, is that attention is established through time and is mediated through an explicit temporal window over which probability distributions of gaze and mouse location could both be calculated. Our smoothing constraint is a crude approximation of this more sophisticated model.

encode established web-reading heuristics, particularly layout-driven scanning patterns such as the F-pattern and Z-pattern, to constrain the model towards plausible reading behavior over article-style pages (Huang et al., 2012; Navalpakkam et al., 2013).

3.4.1. Models

We tested OpenAI's GPT-5.2 Instant and Google's Gemini 3 Flash (preview) as representative of current state-of-the-art multimodal LLMs capable of image understanding. Inputs were downsampled and compressed prior to submission to comply with API limitations in processing screenshots. The gaze predictions are mapped back to screenshot coordinates later on. The output predictions were also clipped to valid screen bounds.

3.4.2. Structured output constraint

Each inference produces a strictly valid JSON:

- (x, y) pixel coordinates
- `attention_pattern` $\in$ {`F-pattern`, `Z-pattern`, `Center-focus`, `Cursor-aligned`}
- `reasoning_mode` $\in$ {`Reading`, `Scanning`, `Targeting`, `Idle`}
- `confidence` $\in [0, 1]$

A schema-based constraint is used to reduce free-form text and enforce machine-readable outputs for downstream evaluation.

4. Experimental Results

4.1. Experimental Scenarios

We define two evaluation conditions or scenarios:

1. **Full-video prediction.** The model receives only the screenshot sequence (with mouse overlays) and produces a gaze estimate and metadata for each frame. No ground-truth gaze is provided during the run.

2. **Per-frame prediction.** After the model predicts the gaze for frame t, we immediately provide the actual reference gaze coordinate $\hat{p}_t$(geometric median) as ground truth. The next frame's prompt explicitly instructs the model to use this provided coordinate rather than its previous prediction. This setting provides an online correction signal, analogous to teacher forcing for sequence prediction, to prevent error accumulation as the model predict frames. This enables us to test the model's power in predicting the next gaze location. As a collateral benefit, it enforces **temporal continuity** in that it ensures smooth gaze trajectories across 5-second steps.

4.2. Numerical Results

We tested two multimodal LLMs under identical conditions: *Google's Gemini 3 Flash (preview)* and *Open-AI's GPT-5.2 Instant*. Tests were conducted on data from three randomly selected participants from the EOTT dataset (Papoutsaki et al., 2018) (P1, P2, and P7) and under two scenarios: full-video and and per-frame.

To calculate region-level estimation accuracy, we divided the screen into 300x300 pixel areas and checked whether the predicted gaze location is in the same region as the actual gaze location. This evaluation regimen is commonly used in previous work (e.g., (Ahmadzadeh, 2024; Papoutsaki et al., 2018)), because some applications only need the rough location of the gaze.

4.2.1. Model Comparison

GPT-5.2 Instant yields more stable predictions than Gemini 3 Flash. The results of these two models for one of the selected participants (P1) are shown in Table 1. The results are very close, both in terms of region accuracy and gaze coordination error, although GPT-5.2 Instant is slightly better. A key qualitative difference is that GPT-5.2 Instant tends to lower its confidence when page content provides insufficient evidence (e.g., sparse or ambiguous layouts), whereas Gemini 3 Flash often maintains moderate confidence in such cases.

Therefore, in the analysis below, we only show the results for GPT-5.2 Instant.

4.2.2. Overall results

Table 2 reports detailed results for the tested users, as well as macro-averaged results.

Our first observation is the relatively low accuracy in estimating the gaze region (around 17% for Full-Video and around 24% in the Per-Frame scenario). However, when the region is not accurately estimated, in more than half of the cases, the estimated region was one of the adjacent regions (up, down, left, or right). In other words, the estimated region is not very far from the actual location of the gaze in the majority of the cases.

As we expect, the estimation errors accumulate in the Full-Video scenario and therefore the performance is worse: around 50 pixels higher for estimated location error and around 10% lower region accuracy. This trend is consistent for all participants.

Model	Condition	Region Accuracy	Average Error (px)	Median Error (px) (All)	(CL>0.8)	(CL>0.9)
Gemini 3 Flash	Full-Video	15.7%	357	296	293	307
	Per-Frame	27.5%	290	229	228	224
GPT-5.2 Instant	Full-Video	17.7%	356	322	321	322
	Per-Frame	27.4%	286	210	196	210

Table 1: Comparison of different multimodal LLMs.

User	Condition	Average Error (px)	Median Error (px) (All)	(CL>0.8)	(CL>0.9)	Region Accuracy (%) (All)	(CL>0.8)	(CL>0.9)
P1	Full-Video	356	322	321	322	17.7	16.2	16.2
	Per-Frame	286	210	196	210	27.4	25.6	25.6
P2	Full-Video	357	341	341	341	17.6	17.6	17.6
	Per-Frame	295	253	253	262	27.5	30.4	30.4
P7	Full-Video	428	411	411	446	15.7	10.7	10.7
	Per-Frame	410	385	403	414	17.6	14.3	14.3
Average	Full-Video	380	358	358	370	17.0	14.8	13.5
	Per-Frame	330	283	284	295	24.2	23.4	23.4

Table 2: Gaze prediction results of our method for different users.

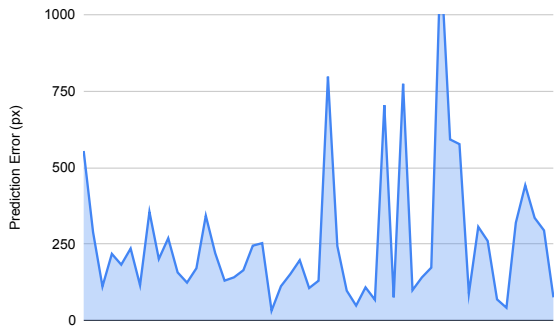


Figure 1: Per-frame error spikes of one user.

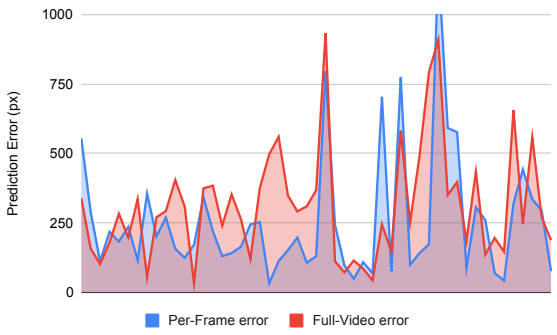


Figure 2: Comparison of per-frame and full-video error rates.

We also see that while the performance is similar for the first two users, it is significantly worse for the third user, especially in the per-frame scenario. This can be due to individual differences between human users.

4.2.3. Temporal error dynamics

Figure 1 plots the per-frame error over time for one of the users (P1). Error spikes align with context discontinuities, such as navigation events or transitions to empty/calibration-like pages. In such frames, the model lacks strong layout cues (text blocks, headings, images) needed to infer reading patterns, leading to unreliable gaze estimates. Moreover, empty-page segments can affect subsequent predictions by disrupting the inferred scanpath trajectory, especially in the full-frame scenario, in which errors accumulate.

4.2.4. Effect of per-frame prediction

Figure 2 contrasts the full-video and per-frame scenarios for one user (P1), illustrating that providing the previous ground-truth gaze substantially stabilizes trajectories after context shifts by preventing error accumulation. This suggests that temporal anchoring mitigates drift and reduces cascading errors when visual evidence is weak.

5. Conclusion

In this paper, we have proposed the idea of using multimodal LLMs to predict gaze using only a screenshot and the movement of the mouse pointer. Our preliminary results show that, while our results are not very high, they are promising. For example, our approach can correctly guess either the correct region or a region adjacent to it (among 12 300px x 300px regions on the page) in over half of the

cases we tested. We think with the advances in small multimodal LLMs (such as small Gemma 3 variants that can easily be run on user devices), our idea, with refinement, can serve as the basis for on-device gaze estimation without further need for any special hardware.

Limitations

The main motivation of our work was to remove the need for special hardware (such as eye-tracker or cameras) in gaze estimation. Such hardware typically requires continuously watching the user with a camera, which is a significant privacy concern. Our approach only needs to analyze the content of the screen. It is therefore less intrusive. On the other hand, it requires submitting a screenshot to an LLM.

A limitation of our current results is that it is based on a public dataset that was collected in a lab setting. We should follow up with more evaluations that run in a user's setting (e.g., in their home or office). We should also evaluate the accuracy of gaze prediction with techniques that are usually used during calibration.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Seyed Fatemeh Ahmadzadeh. 2024. User's gaze estimation based on the movement of the mouse and screen information. Bachelor's thesis, Amirkabir University of Technology, Tehran, IRAN, June.

Diego Alves. 2025. Benchmarking language model surprisal for eye-tracking predictions in Brazilian Portuguese. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 7–17, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Amir Beck and Shoham Sabach. 2015. Weiszfeld's method: Old and new results. *J. Optimization Theory and Applications*, 164(1):1–40.

Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. 2009. What do you see when you're surfing? using eye tracking to predict salient regions of web pages. In *Proceedings of CHI*, pages 21–30.

Qi Guo and Eugene Agichtein. 2010. Towards predicting web searcher gaze position from mouse movements. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 3601–3606. ACM.

Jeff Huang, Ryen White, and Georg Buscher. 2012. User see, user point: gaze and cursor alignment in web search. In *Proceedings of CHI*, pages 1341–1350.

Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *JASIST*, 59:1041–1052.

James A. Michaelov and Roger P. Levy. 2026. N-gram-like language models predict reading time best. Technical Report 2603.09872, arXiv.

Alexandre Milisavljevic, Fabrice Abate, Thomas Le Bras, Bernard Gosselin, Matei Mancas, and Karine Doré-Mazars. 2021. Similarities and differences between eye and mouse dynamics during web pages exploration. *Front. Psychol.*, 12:554595.

Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd Int. Conference on World Wide Web*, pages 953–964.

Justin W. Owens, Barbara S. Chaparro, and Evan M. Palmer. 2011. Text advertising blindness: The new banner blindness? *J. Usability Studies*, 6(3):172–197.

Alexandra Papoutsaki, Aaron Gokaslan, James Tompkin, Yuze He, and Jeff Huang. 2018. The eye of the typer: a benchmark and analysis of gaze behavior during typing. In *Proceedings of ACM ETRA*.

Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. Webgazer: scalable webcam eye tracking using user interactions. In *Proceedings of IJCAI*, pages 3839–3845. AAAI Press.

Cristina Maria Popescu and Sergiu Nisioi. 2025. Exploring mouse tracking for reading on Romanian data. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 44–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Mads Soegaard. 2021. Visual hierarchy: Organizing content to follow natural eye movement patterns. IxDF - Interaction Design Foundation (accessed 29th March, 2026).

A Survey of Incorporating Gaze Data into Natural Language Processing Models and Applications

Cengiz Acartürk[1], **Burcu Can**[2], **Melike Çağlayan**[1], **Jamal Abdul Nasir**[3], **Çağrı Çöltekin**[4]

[1]Centre for Cognitive Science, Jagiellonian University, [2]Department of Computing Science and Mathematics, University of Stirling, [3]School of Computer Science, University of Galway
[4]Department of Linguistics, University of Tübingen
cengiz.acarturk@uj.edu.pl, burcu.can@stir.ac.uk, melike.caglayan@uj.edu.pl, jamal.nasir@universityofgalway.ie, cagri.coeltekin@uni-tuebingen.de

Abstract

This study presents a survey of research integrating eye-tracking (gaze) data into Language Models (LMs) as a means of cognitively grounding NLP models and applications in human reading behavior. Although contemporary LMs excel at learning statistical patterns from text, they fundamentally lack human-like reading and comprehension capabilities. Incorporating gaze data may offer a window into cognitive processing, yet its impact on LMs remains underexplored. Addressing a persistent bottleneck, namely, the high cost and limited scale of laboratory eye-tracking, we propose a roadmap consisting of three streams of research for advancing this novel research domain: (1) developing cognitive multimodal corpora, (2) leveraging generative models for gaze synthesis to overcome the data bottleneck caused by the high costs of human eye-tracking, and (3) training LMs with gaze-guided attention mechanisms and input augmentation. Furthermore, we illustrate practical applications in readability assessment, educational analytics, and assistive communication, demonstrating how gaze-informed models can enable adaptive technologies. Finally, we critically examine ongoing challenges, including the lack of data standardization, the misalignment between human and machine language processing, and the urgent ethical imperative for privacy-preserving architectures to protect sensitive biometric gaze data, motivating privacy-aware data practices and model designs for scalable deployment.

Keywords: Gaze data, Language Models, Human Reading Behavior

1. Introduction

Natural Language Processing (NLP) has achieved remarkable success, driven largely by deep learning architectures that learn statistical patterns from massive text corpora (language models, henceforth LMs). However, LMs often lack the cognitive grounding that characterizes human language processing and comprehension. Eye-tracking data, specifically the measurement of fixations and saccades, provide a psychophysiological window into these cognitive processes. Common gaze features used in NLP include fixation duration (processing effort), first-pass reading time (early processing), total reading time (integration difficulty), regression rate (reanalysis), and skipping probability (predictability). These features provide measurable indicators of cognitive processing during reading and are widely used in computational models.

The integration of gaze data into NLP models and applications offers a pathway to bridge the gap between statistical probability and human-like reading and comprehension by informing models about which words might carry the most information, and how attention and cognitive load might change during reading from a human-reader perspective (cf., Cognitively Inspired NLP, Mishra et al. 2017; cognitive signals, Hollenstein et al. 2020a). This provides the development of a novel research domain that bridges NLP and cognitive sciences, attracting current research and proceeding towards establishing a community of researchers (Acartürk et al., 2025). To give a concrete example to the studies in this context, gaze data may bridge a text-specific property, such as sentence complexity, to its human-reader counterpart, namely, reading difficulty.

Recently, the research on developing computational models of eye-movement during reading faced a significant data bottleneck, as collecting gaze data involves high cost and time effort, also requiring specialized hardware and controlled laboratory environments. These challenges have limited the use of gaze data in NLP to relatively small-scale studies. However, a recent shift has been moving the eye-tracking field from small datasets to standardized multimodal and multilingual corpora by incorporating synthetic gaze generation.

This survey reviews the current role of gaze data in NLP research and outlines a structured research direction through a three-stream roadmap (see Figure 1): (1) Developing cognitive multimodal corpora, (2) enriching cognitive multimodal corpora by gaze synthesis, and (3) training LMs for gaze guidance embedded in enriched cognitive multimodal corpora. We also address methodological and ethical considerations for large-scale collection, synthesis, and deployment due to the personal and potentially sensitive nature of gaze data.

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 64–76
May 12, 2026.

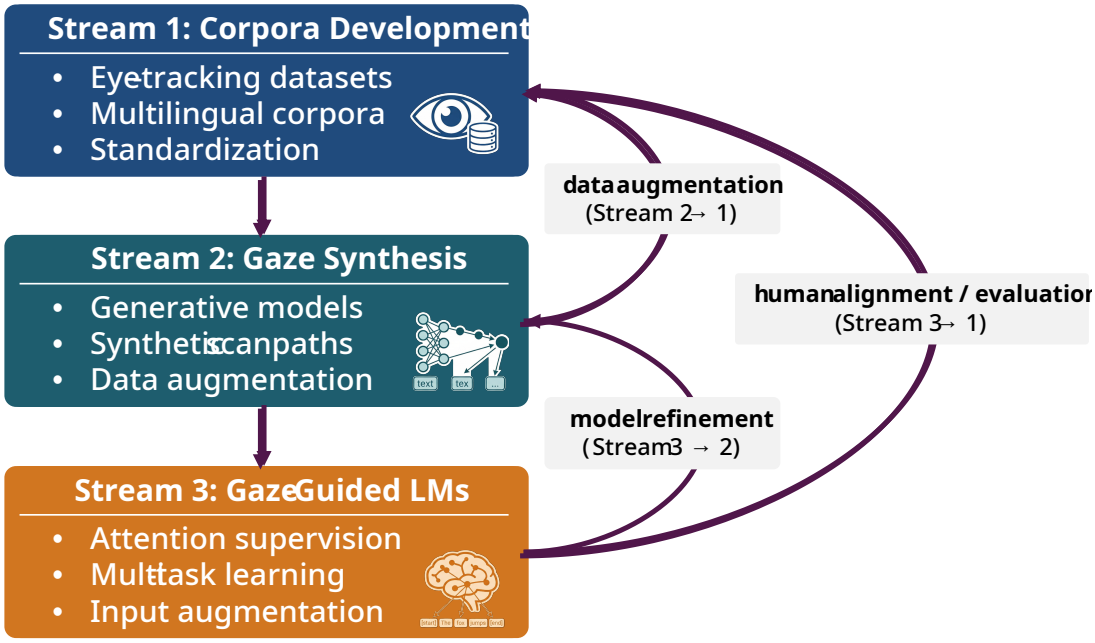


Figure 1: Roadmap for gaze-integrated NLP. The three streams for corpora, gaze synthesis, and gaze-guided models can be developed in parallel, with feedback loops that support continuous improvement and data augmentation.

2. A Roadmap for Gaze for NLP

This section introduces the roadmap in three streams of research that can proceed in parallel with minor dependencies. The first stream presents the studies targeting the development of cognitive multimodal corpora. Given their limited size, due to challenges in developing multimodal corpora, the next stream addresses synthesizing gaze data to enrich multimodal corpora and the studies conducted in this direction. The next stream focuses on training LMs on enriched cognitive multimodal corpora through gaze guidance.

2.1. Developing Cognitive Multimodal Corpora

Developing corpora has been the very first step for conducting quantitative linguistic research and developing practical NLP applications for many decades. The scope of corpus development has been expanded for the past decade in at least two directions: The development of text-only corpora and the development of multimodal cognitive corpora, both for training LMs. Due to historical reasons, several terminological ambiguities have emerged. One is the use of the term "multimodal", which has been used in NLP studies to refer to language resources including linguistic and nonlinguistic content, such as text and images. Another use of the term refers to the method of measurement and the resulting data, such as gaze and brain imaging data. To resolve the terminological ambiguity in the field, we propose a distinction between extrinsic multimodality (content-multimodal corpora), which concerns the diverse formats of information representation (e.g., text paired with images or video), and intrinsic multimodality (cognitive multimodal corpora), which concerns the diverse channels of human cognitive processing (e.g., simultaneous recording of gaze, EEG, and fMRI while reading text). This survey specifically addresses the latter, using the term cognitive multimodal corpora, by leveraging intrinsic cognitive signals to ground language representations in human processing patterns. While extrinsic multimodal systems learn to relate objects and concepts across different external media, intrinsic multimodal systems learn to model the observable signals obtained from the reader.

Several cognitive multimodal corpora have been released during the past decade. The datasets, such as ZuCo 2.0 Zurich Cognitive Language Processing Corpus (Hollenstein et al., 2020b) and GECO Ghent Eye-Tracking Corpus (Cop et al., 2017) provide standardized benchmarks. For instance, ZuCo 2.0, provides simultaneous eye-tracking and EEG recordings during both natural reading and specific annotation tasks of 739 English sentences read by 18 participants. Similarly, GECO consists of gaze data from monolingual English speakers and Dutch-English L2 learners reading an entire novel, providing a continuous narrative context often missing in sentence-level datasets.

While early work focused primarily on English, the field has rapidly expanded into diverse lan-

guages and specialized domains to create standardized benchmarks. Recent efforts include the Copenhagen Corpus (CopCo), which provides the first eye-tracking-while-reading corpus for the Danish language, consisting of 1,832 sentences and nearly 35,000 tokens (Hollenstein et al., 2022). Similarly, the PoTeC, Potsdam Textbook Corpus (Jakobi et al., 2025a), is a German naturalistic resource comprising data from 75 participants reading scientific texts. It employs a factorial design contrasting expert and novice readers to investigate how domain expertise and prior knowledge influence cognitive processing during reading. Recently, GAZE4HATE dataset expands the scope of reading stimuli from common text to instances of annotated hate-speech segments (Alacam et al., 2024).

To address the historical lack of uniformity across laboratories, the MultiplEYE initiative has been establishing a large-scale, multilingual corpus with consistent recording protocols (Jakobi et al., 2025b) Novel datasets like CoLAGaze (Bondar et al., 2025) provide the first broad-coverage eye-tracking corpus on grammatical and ungrammatical sentences, aligned to the CoLA benchmark, enabling direct comparison between model grammaticality judgments and human processing patterns. Furthermore, WebQAmGaze (Ribeiro et al., 2023) addresses accessibility, utilizing webcam-based eye tracking from 600 participants across four languages (English, German, Spanish, and Turkish), demonstrating that low-cost, scalable data collection is feasible without specialized hardware. InteRead (Zermiani et al., 2024), with 50 participants, annotates interruptions and resumption lags during reading, addressing the ecological validity of learning environments. Additional multilingual resources continue to expand the range of cognitive multimodal corpora. For example, the FETA corpus provides French eye-tracking data collected from 46 readers across general, medical, and clinical texts presented in both original and manually simplified versions, with multiple word-level gaze features available for analysis (Ivchenko and Grabar, 2025). To provide a clearer overview of existing resources, Table 1 summarizes representative cognitive multimodal corpora.

The available cognitive multimodal corpora, as exemplified in this section, provides valuable resources for developing LMs by incorporating gaze data. On the other hand, a major challenge in the field is the limited size and scope of the available datasets, mostly consisting of a few dozen participants and a few thousand words, due to the high cost and time effort required in gaze data collection. This scarcity has necessitated a shift in research focus for the past decade, seeking an answer to the following question: Is it possible to synthesize gaze data to predict human gaze patterns on unseen text, so as to scale up training sets for gaze-augmented models? The studies in this direction comprise the second stream of the research roadmap, presented below.

2.2. Gaze Synthesis to Enrich Cognitive Multimodal Corpora

The predictive models of eye movements during reading emerged in the beginning of the century, established on experimental findings obtained in previous research. The empirical research largely addressed the factors impacting eye-movement patterns on text, such as corpus frequency of lexical items, word length, and predictability of words in a sentential context. A set of computational cognitive models aimed to parameterize these factors to predict when and where to move the eyes during reading (E-Z Reader, Reichle et al. 1998; SWIFT Saccade-generation with inhibition by foveal targets, Engbert et al. 2002; Glenmore, Reilly and Radach 2003; OB1-reader, Snell et al. 2018; see Acartürk 2025 for a recent review). For the past several decades, statistical regularities on eye movements during reading have attracted research interest from a modeling point of view, still being an ongoing debate (Kimchi and Siegelman, 2026).

The scope of early modeling in the cognitive computational models has been limited to rule-based, algorithmic approaches while some employed connectionist architectures or parameter optimization. Recent models have involved explicit machine-learning approaches, also targeting synthesized gaze patterns. For instance, models like ScanTextGAN and the hybrid text saliency models proposed by Khurana et al. (2023) indicate that synthetic scanpaths can be used to supervise NLP models. Specifically, they were able to improve the accuracy of multiple NLP tasks, such as sentiment analysis, named entity recognition, relation extraction, and grammatical error detection, using synthetic scanpaths produced by the model, on a variety of datasets. ScanTextGAN presents a generative approach designed to synthesize human-like scanpaths (the sequence of fixations and saccades) over text, also aiming to replicate the temporal sequence of reading, including regressions (looking back) and skips. The model was trained on real eye-tracking data to learn the probability distribution of eye movements during reading. Hybrid and cognitive text saliency models aim to predict aggregate gaze metrics (like total fixation duration) rather than full scanpaths, yet combining traditional cognitive reading models with data-driven gaze supervision. This approach reduces the full reliance on black-box neural models, by integrating established psycholinguistic features (e.g., word length, frequency) with neural attention. Empirical findings

Dataset	Language	Participants	Size
ZuCo 2.0	English	18	739 Sentences
GECO	English, Dutch (L2)	30	5,000 Sentences
CopCo	Danish	22	1,832 Sentences
PoTeC	German	75	12 Scientific Texts
FETA	French	46	32 → 179 Sentences
WebQAm Gaze	English, German, Spanish, Turkish	600	XQuAD & MECO texts
Inte Read	English	50	28 Pages
CoLA Gaze	English	42	306 Sentences
Multipl EYE	Multilingual (30+ languages)	100+*	10 Natural Texts
GAZE4HATE	English	43	90 items

Table 1: Overview of representative cognitive multimodal corpora with eye-tracking data (*per lab, ongoing)

suggest that the field will increasingly pivot toward gaze synthesis in the near future, relying on the expectation that the accuracy of synthesized gaze will further and further approximate real gaze data.

Alongside ScanTextGAN, several architectures have emerged for gaze synthesis. Eyettention (Deng et al., 2023) is such a gaze-synthesis model, one that processes linguistic tokens and chronological fixation sequences simultaneously using cross-sequence attention. By modeling the complex temporal dynamics of reading, Eyettention performs scanpath prediction across multiple languages and datasets. Another model is ScanEZ (Sood et al., 2025), which integrates traditional cognitive models with self-supervised learning to produce spatiotemporally realistic scanpaths. This hybrid approach demonstrates that incorporating traditional, computational cognitive models that address psycholinguistic factors can improve generalization beyond purely data-driven methods.

These synthesis approaches differ among each other in both their representational assumptions and architectural design. Sequence-based models conceptualize scanpaths as ordered discrete events (i.e., fixations and saccades), typically modeled with autoregressive or recurrent frameworks. Complementary work has also investigated the linguistic determinants of saccade targeting, showing that, beyond word length and frequency, contextual meaning, prior fixation history, and saccade distance contribute to predicting forward and backward eye movements during naturalistic reading (Rego et al., 2025). In contrast, trajectory-based models treat gaze as a continuous spatiotemporal signal and often employ probabilistic or neural dynamical systems to capture smooth transitions in gaze coordinates. Conditioning strategies also vary: some models incorporate explicit linguistic features (e.g., readability indices, syntactic complexity, or lexical frequency), whereas others learn end-to-end mappings from token-level embeddings directly to gaze trajectories without handcrafted features.

While generating synthetic gaze data offers a theoretical workaround to the high cost of collecting human eye-tracking data, it is crucial to recognize significant limitations inherent in relying on synthesized data. Currently, it remains unclear how well synthetic gaze truly captures the complexity of human reading behavior. A primary shortcoming of synthetic gaze data is the reduction of natural variability. Because generative models are trained to optimize for widespread statistical regularities, they risk ignoring the rich, idiosyncratic variations present in human reading patterns. Another weakness related to the existing models is the lack of assessment methodologies in the recent work. In general, the evaluation of the relationship between machine behavior and human behavior is scarce in the context of incorporating gaze data into LMs, except for a few studies in this direction, such as Ikhwantri et al. (2023). Aggregate behavioral statistics (e.g., mean fixation duration, saccade length, regression rate) or predictive measures (e.g., next-fixation accuracy) are generally used to measure synthetic gaze. Although informative, these measures mainly capture superficial likeness and may not indicate whether the generated scanpaths reflect the finer-grained linguistic processes in human language processing (e.g., sensitivity to syntactic boundaries, semantic ambiguity, or surprisal). As a result, matching global distributions does not guarantee that synthetic gaze preserves the cognitively compatible signal required for modeling. Specifically, current synthetic models generally fail to model individual differences. Reader-specific traits, such as working memory capacity, domain expertise, second-language proficiency, or reading disorders, profoundly influence gaze behavior. By heavily relying on aggregated training sets, generative models run the risk of biasing NLP applications toward the "average gazer." This homogenization means that models might ignore outlier behaviors or non-standard reading strategies, limiting the ecological validity and inclusivity of applications like assistive communication or personalized education.

In summary, the methodological issue of trying to determine that the synthesized scanpaths do not

store task-relevant information content, but are just an approximation of statistical regularities, is still unresolved. It is not clear that existing generative models reproduce systematic variability with regard to individual variability, task requirements, and processing linguistically complex or ambiguous stimuli. Synthetic gaze augmentation effectiveness is determined by whether models are able to encode both the statistical patterns of eye movements and the structured modulation of gaze based on textual properties, reader characteristics, and contextual constraints.

Consequently, while gaze synthesis provides an alternative avenue for data augmentation, its utility must be critically weighed against these conceptual and methodological shortcomings. Overreliance on synthetic datasets without addressing the loss of individual variability and nuanced cognitive processing poses a major risk. Acknowledging these limitations, the next challenge lies in architectural integration: How can both empirical and (cautiously applied) synthetic cognitive signals be effectively injected into neural NLP models? The studies that aim to find answers to these questions comprise the third and final stream of the research roadmap, presented below.

2.3. Gaze Guidance for Training LMs with Enriched Cognitive Multimodal Corpora

The previous studies on incorporating gaze data into LMs have mainly converged on two strategies: augmenting the input representation (concatenation) and supervising the internal computation (attention mechanisms). The relationship between gaze and visual attention is basically an operational assumption rather than being an observational fact. The eye movements during reading indicate visual attention while the two do not necessarily couple momentarily (Posner et al., 1980). Moreover, aligning a model's attention mechanism with human visual attention is also a technically sophisticated integration approach. In standard Transformer or RNN-based architectures, "attention" is a learned weight distribution that determines which input words the model should focus on. A feasible approach is to modify the loss function of the neural network to minimize the distance between the model's attention weights and human fixation distributions (or synthetic equivalents) instead of letting the model learn attention solely from the target task (e.g., translation). This approach assumes that human gaze is a proxy for attention. Recent research shows that integrating gaze features into the attention mechanisms of neural networks can improve performance in tasks like paraphrase generation and sentence compression (Khurana et al., 2023). Such supervised learning of attention mechanisms combined with gaze data provide a practical benefit, as once training is complete, it does not require gaze input during inference. It allows building cognitively enriched attention mechanisms, which have signals from both raw textual data and cognitive data.

Another approach for informing NLP systems with gaze data relies on joint, multi-task models that learn gaze behaviour alongside the target NLP task. In this approach, the model is trained to perform two tasks simultaneously: a primary task consisting of a specific NLP goal (e.g., sentence compression, sentiment analysis), and an auxiliary task aiming to predict gaze metrics (e.g., total fixation duration) for the input tokens. By sharing the underlying encoder layers between these two tasks, the model learns a generalized text representation that encodes both semantic meaning and cognitive processing effort. This joint-modeling approach ensures that gaze data regularizes the model, preventing it from overfitting to superficial statistical patterns in the text. Sood et al.'s (2020b) joint modeling approach— training the NLP model to solve the task and predict human gaze simultaneously— outperformed state-of-the-art baselines. Similarly, Mathias et al. (2020) introduce a model where gaze behaviour is learned along with essay grading, and the results show that modeling gaze behaviour along with essay grading provides statistically significant performance gains for the task.

A similar approach is to concatenate gaze vectors with word embeddings to enrich the input representation. Gaze measures can also be normalized and concatenated directly with word embeddings (e.g., BERT or GloVe vectors) to form the input to the network. However, this method increases the dimensionality of the input and requires the gaze data to be available at test time (unless replaced by synthetic features), making it less flexible than the attention supervision methods. On the contrary to the previous approach, where the attention mechanisms are trained along with the gaze features, this approach requires gaze features during inference time to be available, which makes the approach less practical compared to the previous approach.

Empirical evaluations of these approaches have yielded mixed but instructive results. Barrett et al. (2018) showed that regularizing recurrent models with human attention from eye-tracking corpora improved performance across sentiment analysis, grammatical error detection, and abusive language detection, suggesting that human gaze provides inductive biases that help models generalize. Similarly, Wang B. et al. (2024) investigated whether integrating gaze signals into BERT during pretraining or fine-tuning enhances performance, reporting gains on several tasks. Hollenstein and Zhang (2019) demonstrated consistent improve-

ments from gaze and EEG augmentation for named entity recognition, relation classification, and sentiment analysis across multiple datasets. However, Sood et al. (2020a) found that while model attention can be shaped to resemble human attention through supervision, the correlation between attention similarity and task performance is architecture-dependent: LSTM and CNN attention aligned better with human fixations and correlated with performance, whereas XLNet showed weaker alignment despite high task accuracy. This suggests that the concept of attention supervision can be best implemented when the internal process of the model can be compatible with the human-like sequential processing. The major outstanding problem is to enable type-level generalization, i.e. learners to use the gaze patterns trained on training texts without having to provide eye-tracking information at testing time (Hollenstein and Zhang, 2019).

In summary, training LMs with enriched cognitive multimodal corpora seems to be a promising research direction for effectively incorporating gaze data. Below, we present research domains that might extend the frontiers of the field through applications.

3. Applications

This section exemplifies several frontiers that the research on incorporating gaze data into LMs has been expanding

3.1. Text Complexity and Reading Difficulty

Text complexity and reading difficulty are two concepts that constitute the two piers of the bridge between NLP and cognitive sciences. The link between gaze behavior and text difficulty has been well established in psycholinguistics, but translating this link into computational models is not straightforward. Text difficulty is not a single construct; it depends on lexical properties, syntactic structures, discourse relations, and reader-dependent variables such as reading skill and prior knowledge. These factors interact and produce variable eye-movement patterns across readers and tasks.

Traditional readability formulas mainly use surface features such as word length and sentence length. These features provide an incomplete proxy for processing effort and do not directly reflect cognitive operations during reading, such as lexical access, syntactic parsing, and semantic and discourse integration. Eye-tracking studies show that fixation duration, regression rate, and skipping probability are influenced by linguistic predictors like predictability, and syntactic ambiguity. As a result, surface readability scores often fail to capture the variation observed in human gaze data. For computational modeling, this creates a mismatch between input representations and gaze signals. Gaze data are temporally structured and respond to local linguistic properties at the word and sentence level, while many traditional NLP models use global or static text features to represent difficulty. A more suitable approach is to incorporate psycholinguistically motivated predictors, such as surprisal, entropy, and syntactic dependency measures, alongside contextual embeddings. Recent work further shows that surprisal estimates from LLMs predict several eye-tracking measures, including first fixation duration and gaze duration, reproducing well-established reading-time effects across languages and datasets (Alves, 2025). This alignment can improve the mapping between text features and gaze patterns, supporting more accurate modeling of reading behavior. Overall, gaze-augmented methods aim to assess text difficulty by measuring actual processing effort during reading. Regression count indexes reanalysis caused by comprehension problems or structural ambiguity. First-pass reading time reflects early lexical access and syntactic integration, whereas total reading time captures later reanalysis and cumulative processing. These measures can diverge: a frequent word may show short first-pass time but longer total time when it appears in syntactically demanding contexts, indicating interaction between lexical and syntactic processing that surface features miss.

Nevertheless, applying these findings to readability systems raises practical issues. Gaze data shows high variability across readers due to differences in skill, working memory, domain knowledge, and reading strategy. Aggregating at the type level can mask this variability, while token-level modeling requires large datasets for stable estimates. In addition, the link between gaze and comprehension depends on task demands (e.g., skimming vs. careful reading) and genre, which limits its model generalization across corpora. Recent work has started to operationalize these findings. Studies that combine gaze features with linguistic and psycholinguistic variables in multi-task learning models report improved readability prediction in English (González-Garduño and Søgaard, 2017) and in morphologically rich languages such as Arabic (Baazeem et al., 2025). Extending this line of work to lower-resource languages, Hodivoianu et al. (2025) introduce the first Romanian eye-tracking dataset for reading and show that both feature-based models and fine-tuned Romanian BERT architectures can predict total reading time at the word level, with applications to lexical simplification and interactive readability support. Transformer-based language models have been shown to better account for human reading effort measures than RNNs, including self-paced reading times and neural activity during sentence processing (Merkx and Frank, 2021). These findings suggest that attention-

based architectures can capture aspects of human language processing, although their fit to behavioral data depends on the type of measures and the experimental setting.

3.2. Educational Analytics

A promising field of research that benefits from incorporating gaze data into LMs has been educational analytics. For instance, foreign language learning and its evaluation by automated essay scoring are the two domains that can be bridged by incorporating gaze data into LMs. In educational settings, gaze data has the potential to serve as a measure of learning Sharma et al. (2020); John et al. (2025). Although gaze is an indirect measure of learner engagement and comprehension difficulty (Hutt et al., 2024; Turčáni et al., 2024), specific gaze measures, such as regressive saccades, may correlate with various levels of text complexity, such as lexical- and syntactic-level complexity (Turčáni et al., 2024), also indicating challenges in language learning. An emerging application leverages gaze data to infer open-ended reading goals and information-seeking intentions from eye movements, without explicit user input (Hadar et al., 2025; Shubi et al., 2025). Multimodal LLMs trained on large-scale eye-tracking data have achieved promising success in predicting whether a reader is engaged in comprehension versus targeted information-seeking, and in some cases, reconstructing specific questions or goals driving the reading behavior. Shubi et al. (2025)) found that ensemble transformer-based models combining scanpath representations with LMs can predict reading goals in real-time, with performance modulated by textual properties and reader characteristics. Furthermore, Zhang et al. (2025a) demonstrated that converting raw eye-movement data into visual representations (line-graph images) and encoding them with vision transformers, temporally aligned to text via reading order, achieves good performance across multiple tasks.

These applications indicated that in practical contexts, such as automated essay scoring (AES), gaze-augmented models may improve traditional AES systems, which rely solely on text-level features (grammar, vocabulary size). Similarly, for second-language (L2) learners, gaze-augmented models may help identify which words impede comprehension. Gaze-augmented models trained on data can personalize reading materials by highlighting or simplifying difficult vocabulary in real-time. Intelligent tutoring systems can dynamically infer when a student is struggling with comprehension versus quick skimming, adjust content difficulty in response to detected cognitive load, and provide just-in-time scaffolding. In massive open online courses (MOOCs) and digital learning environments, gaze-based analytics could offer insights into learner engagement patterns, identifying moments of confusion or disengagement that might predict dropout.

3.3. Assistive Communication

Eye tracking technology has recently been increasingly used in assistive communication systems, particularly for individuals with severe speech and physical impairments. Eye gaze-based text entry (eye typing) is one of them, where the user with motor impairments points or looks at the desired letter within the user interface, where a screen keyboard and eye-tracking device are needed (Majaranta and Räihä, 2007; Panwar et al., 2012). In such assistive systems, dwell time is often used as the gaze input. However, it is likely that an incorrect input is activated when the user simply scans the interface, namely the Midas Touch problem (Jacob, 1991). To accompany a gaze-based text entry, current LMs are substantially used to enable autocompletion, context-aware predictions, and error correction during eye typing. The intended phrases can be inferred from partial input using neural language models, which also mitigates issues such as Midas Touch problem. Cai et al. (2024) introduce an LLM-assisted gaze-based text entry that saves 57% more motor actions than traditional predictive keyboards and has been tested on users with amyotrophic lateral sclerosis.

To assist users with speech disabilities and motor impairments, gaze-enabled communication platforms are designed in a way that gaze-based selection is enabled to generate speech. These Augmentation and Alternative Communication (AAC) platforms are designed to support or replace spoken and written language for individuals with speech and communication impairments. These platforms integrate gaze interfaces with symbolic communication boards, text-to-speech synthesis, and adaptive language modeling. Such systems can learn user-specific linguistic preferences, automatically generate grammatically well-formed utterances from selected semantic units, and provide context-aware topic prediction based on conversational history (Cai et al., 2024). By leveraging NLP techniques, gaze-enabled AAC platforms move beyond letter-by-letter input toward concept-based and intent-driven communication, showing the potential and future research directions for adaptive, context-aware communication technologies for non-verbal individuals (Beck Wells, 2025).

4. Discussion and Recent Challenges

A central concept that provides the appropriate framework for bridging NLP and cognitive science is the Human-in-the-Loop (HitL) framework. Although LMs effectively mimic human communication, their

internal logic remains fundamentally distinct from the unobservable mental processes of the brain. By incorporating gaze data to LMs, researchers aim to move beyond simple imitation toward models that reflect human-like attention and contextual understanding. Practical applications of this paradigm are promising for personalized reading interfaces. HitL systems can adaptively simplify or enrich text to meet a reader's specific needs—a significant breakthrough for language learners or those tackling complex technical material. Gaze data may also be useful for training LMs under low-resource scenarios. In the case of training a language model with limited data, gaze information may potentially allow better generalizations about language.

However, bridging the gap between artificial attention mechanisms and human skill remains a complex frontier, presenting significant theoretical and technical challenges that must be addressed to realize these adaptive technologies fully. Consequently, numerous research questions and technical challenges have remained unsolved recently.

A main challenge in incorporating gaze data into LMs is the lack of standardization of gaze data. The lack of standardized gaze data formats limits its reproducibility and data integration. In contrast to text data, which uses common schemas such as CoNLL-U or JSON, gaze datasets do not follow a shared representation. Datasets like ZuCo, GECO, CoLAGaze, and WebQAmGaze differ in sampling rates (e.g., 60Hz–1000Hz), coordinate systems (pixels, normalized space, visual angle), and preprocessing steps (fixation detection methods, saccade thresholds, outlier filtering). This variability makes cross-dataset comparison, transfer learning, and the development of general gaze-augmented models more difficult. To accelerate the adoption of gaze-augmented NLP, the community must establish a standardized format that decouples the raw eye-tracking data from the linguistic tokens.

Another challenge is the gap between human language processing, as indirectly reflected in eye movement patterns, and machine processing of language data. This gap is observable in the methodological assumptions related to using gaze data as input to LMs. Most existing studies do not directly feed raw gaze sequences into LLM prompts. Instead, LLMs are commonly used to generate linguistic labels or representations that are later compared with eye-tracking and neural measures. For example, work combining LLM-generated relevance labels with eye-tracking and EEG data shows that words marked as more relevant receive more and longer fixations, and fixation-related features can support above-chance classification of reading-related neural states (Zhang et al., 2024a,b). Recent research also reports measurable associations between LLM-derived representations and human reading behavior. Eye-tracking datasets designed for LLM evaluation show distinct fixation patterns for preferred versus rejected model responses, along with correlations between reading measures and transformer attention signals (Lopez-Cardona et al., 2025). Probing studies further indicate that internal LLM activations and attention patterns correlate with eye-tracking indicators of reading dynamics, suggesting partial alignment between model prediction processes and human behavioral signals (Wang et al., 2024).

However, empirical comparisons between human gaze patterns and transformer attention representations report only moderate alignment, with encoder models showing stronger correlations with eye-movement data than decoder models across reading tasks (Mouratidi and Poesio, 2025), suggesting a mismatch between behavioral evidence and model architecture. In addition, computational reading models that use LLM-based predictability estimates show better fits to human eye-movement data than traditional cloze-based predictability measures (Lopes Rego et al., 2024). Graph-based text structures generated by LLMs also correspond to fixation distributions, where nodes identified as more important attract higher fixation counts during reading (Zhang et al., 2025b). Alongside these approaches, an alternative direction considers providing LLMs with serialized gaze inputs, such as sequences of fixation coordinates and durations, as structured data. However, current evidence suggests that simple concatenation of numerical gaze sequences with text prompts is unlikely to produce strong performance gains without architectural mechanisms that account for the spatiotemporal nature of eye movements. This limitation arises because gaze signals are continuous, time-dependent, and spatially structured, whereas standard LLM tokenization is primarily optimized for discrete linguistic inputs. With the rapid development of multimodal LLMs and structured data tokenization methods, future systems may more effectively integrate gaze recordings through architectures that explicitly model spatiotemporal eye-movement signals alongside textual representations. A central open question is whether LLM pattern-learning and reasoning abilities extend to spatiotemporal gaze sequences when presented as structured inputs, or whether effective use of gaze data will continue to require dedicated multimodal architectures that explicitly model eye-movement dynamics.

Beyond the technical challenges and the opportunities provided by incorporating gaze data into NLP models and applications, an important aspect is ethics and privacy concerns, a responsibility unfortunately overlooked by many scholars. Eye-tracking data is increasingly treated as a sen-

sitive biometric signal, as patterns of pupil dynamics, gaze trajectories, and eye-movement behavior can support reliable user identification and contain detailed personal information (Kröger et al., 2020; David-John et al., 2021). Beyond identification, gaze recordings may reveal a wide range of attributes without conscious user control, including neurological and behavioral disorders, cognitive load, emotional states, and psychological traits. Eye-movement features have been examined as biomarkers in research on neurodegenerative and mental health conditions, and have been linked to measurable behavioral and psychological patterns (Przybyszewski et al., 2023; Singh and Sharma, 2024; Wang et al., 2025). Machine learning methods further increase the risk of such inferences by enabling the prediction of cognitive and behavioral characteristics from high-dimensional physiological and behavioral signals (Bhatt et al., 2023). A relevant issue is the potential risk of introducing and amplifying biases. By incorporating gaze data from "majority" (cf. the "average gazer", Section 2.2), models' tendency to adopt the majority positions may be reinforced further, leading to negative social consequences.

These concerns are becoming increasingly urgent as eye-tracking technologies are rapidly integrated into consumer and mixed-reality devices, including VR/AR headsets, smartphones, webcams, and camera-based interaction systems, allowing large-scale and often passive collection of gaze data in everyday settings (Zhu et al., 2025; Bozkir et al., 2025; Liebling and Preibusch, 2014). Prior research has shown that even natural gaze behavior in virtual environments can allow user identification, highlighting the need for privacy-preserving data architectures and controlled data access mechanisms (David-John et al., 2021). Because gaze data may reveal identity, fatigue, health-related states, and cognitive or affective processes, privacy-preserving processing pipelines are necessary, particularly as intelligent systems become capable of extracting sensitive biometric and psychological information from subtle behavioral signals (Kröger et al., 2020; Bozkir et al., 2025). As eye-free applications become more capable of predicting these signals from text interaction alone, researchers must develop privacy-preserving architectures that obscure sensitive biometric signatures while retaining utility for NLP tasks.

5. Conclusions

This survey examined the role of gaze data in NLP as a cognitive signal that can support more human-aligned language modeling, tracing the development from small experimental datasets to multilingual corpora and recent synthetic gaze generation approaches. In response to the long-standing data bottleneck caused by the cost and technical constraints of eye-tracking collection, the field has gradually moved toward scalable multimodal resources and gaze synthesis methods that enable broader model training. To structure this progression, we proposed a roadmap consisting of three streams of research: constructing cognitive multimodal corpora, enriching these corpora through synthetic gaze, and training language models with gaze-informed guidance. In addition, recent approaches that incorporate gaze signals into training objectives and alignment frameworks indicate that cognitive supervision can guide model representations toward patterns observed in human reading.

Several directions follow directly from this roadmap. The reviewed studies also show practical applications in readability assessment, educational analytics, and assistive communication, where gaze-informed models can support adaptive and cognitively grounded language technologies. First, standardized evaluation protocols for synthetic gaze are needed to verify whether generated signals preserve linguistic sensitivity and reading-related patterns rather than only surface-level statistical similarity. Second, methods that retain the functional benefits of gaze data while reducing exposure of sensitive biometric patterns should be further developed, given the personal nature of cognitive signals. Third, further expansion and standardization of multilingual eye-tracking corpora is necessary, as recent resources already include datasets in languages such as Danish, German, French, and Romanian, yet cross-linguistic comparability and shared annotation standards remain limited. Finally, future work should examine whether large language models can effectively incorporate spatiotemporal gaze signals and whether dedicated multimodal architectures that explicitly model eye-movement dynamics provide more reliable performance than text-only integration strategies.

Integrating gaze data into NLP supports cognitively grounded language modeling by linking textual representations with observable reading behavior. Across corpora development, gaze synthesis, and gaze-guided training, the reviewed studies show that eye-movement signals provide measurable indicators of processing effort, attention allocation, and reading dynamics that are not captured by text-only models. While current results demonstrate the feasibility of gaze-informed modeling, future progress depends on larger multilingual corpora, standardized evaluation protocols for synthetic gaze, and architectures designed to process spatiotemporal cognitive signals in a scalable and privacy-aware manner.

6. Bibliographical References

Cengiz Acartürk. 2025. *Eyes on Text: Eye movements in reading and language processing*. John Benjamins.

Cengiz Acartürk, Jamal Nasir, Burcu Can, and Cağrı Çöltekin, editors. 2025. *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*. INCOMA Ltd., Shoumen, BULGARIA, Varna, Bulgaria.

Özge Alacam, Sanne Hoeken, and Sina Zarrieß. 2024. Eyes don't lie: Subjective hate annotation and detection with gaze. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 187–205, Miami, Florida, USA. Association for Computational Linguistics.

Diego Alves. 2025. Benchmarking language model surprisal for eye-tracking predictions in Brazilian Portuguese. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 7–17, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2025. Integrating linguistic and eye movements features for arabic text readability assessment using ml and dl models. *Computation*, 13(11):258.

Melissa Beck Wells. 2025. Empowering non-verbal individuals through AI-driven symbolic text prediction: a metaliteracy approach to communication and inclusion. *Discover Education*, 4(1):360.

Priya Bhatt, Amanrose Sethi, Vaibhav Tasgaonkar, Jugal Shroff, Isha Pendharkar, Aditya Desai, Pratyush Sinha, Aditya Deshpande, Gargi Joshi, Anil Rahate, et al. 2023. Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions. *Brain informatics*, 10(1):18.

Anna Bondar, David Robert Reich, and Lena Ann Jäger. 2025. CoLAGaze: A corpus of eye movements for linguistic acceptability. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pages 1–9.

Efe Bozkir, Babette Bühler, Xiaoyuan Wu, Enkelejda Kasneci, Lujo Bauer, and Lorrie Faith Cranor. 2025. The impact of device type, data practices, and use case scenarios on privacy concerns about eye-tracked augmented reality in the United States and Germany. *Journal of Cybersecurity*, 11(1):tyaf036.

Shanqing Cai, Subhashini Venugopalan, Katie Seaver, Xiang Xiao, Katrin Tomanek, Sri Jalasutram, Meredith Ringel Morris, Shaun Kane, Ajit Narayanan, Robert L MacDonald, et al. 2024. Using large language models to accelerate communication for eye gaze typing users with als. *Nature Communications*, 15(1):9449.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

Brendan David-John, Diane Hosfelt, Kevin Butler, and Eakta Jain. 2021. A privacy-preserving approach to streaming eye-tracking data. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2555–2565.

Shuwen Deng, David R Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A Jäger. 2023. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–24.

Ralf Engbert, André Longtin, and Reinhold Kliegl. 2002. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision research*, 42(5):621–636.

Ana Valeria González-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

Cfir Avraham Hadar, Omer Shubi, Yoav Meiri, Amit Heshes, and Yevgeni Berzak. 2025. Decoding open-ended information seeking goals from eye movements in reading. *arXiv preprint arXiv:2505.02872*.

Anamaria Hodivoianu, Oleksandra Kuvshynova, Filip Popovici, Adrian Luca, and Sergiu Nisioi. 2025. Predicting total reading time using Romanian eye-tracking data. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 71–75, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020a. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The copenhagen corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020b. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephen Hutt, Aaron Wong, Alexandra Papoutsaki, Ryan S Baker, Joshua I Gold, and Caitlin Mills. 2024. Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior Research Methods*, 56(1):1–17.

Fariz Ikhwantri, Jan Wira Gotama Putra, Hiroaki Yamada, and Takenobu Tokunaga. 2023. Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour. *Information Processing & Management*, 60(2):103195.

Oksana Ivchenko and Natalia Grabar. 2025. A French eye-tracking corpus of original and simplified medical, clinical, and general texts - FETA. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 37–43, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Robert JK Jacob. 1991. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)*, 9(2):152–169.

Deborah N Jakobi, Thomas Kern, David R Reich, Patrick Haller, and Lena A Jäger. 2025a. PoTeC: A German naturalistic eye-tracking-while-reading corpus. *Behavior Research Methods*, 57(8):211.

Deborah Noemie Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusic Vilus, Stefan Frank, Marie-Luise Müller, Kristine M Jensen de López, et al. 2025b. MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pages 1–11.

Nathalie John, Sebastian P Korinth, Mareike Kunter, and Franziska Baier-Mosch. 2025. Gaze cluster analysis reveals heterogeneity in attention allocation and predicts learning outcomes. *Scientific Reports*, 15(1):20291.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Inbal Kimchi and Noam Siegelman. 2026. All together now: Random forests analysis reveals the joint impact of multiple statistical regularities on eye-movements during reading. *Psychonomic Bulletin & Review*, 33(1):17.

Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. 2020. What does your gaze reveal about you? on the privacy implications of eye tracking. In *IFIP International Summer School on Privacy and Identity Management*, pages 226–241. Springer.

Daniel J Liebling and Sören Preibusch. 2014. Privacy considerations for a pervasive eye tracking world. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1169–1177.

Adrielli Tina Lopes Rego, Joshua Snell, and Martijn Meeter. 2024. Language models outperform cloze predictability in a cognitive model of reading. *PLOS Computational Biology*, 20(9):e1012117.

Angela Lopez-Cardona, Sebastian Idesis, Miguel Barreda-Ángeles, Sergi Abadal, and Ioannis Arapakis. 2025. OASST-ETC dataset: alignment signals from eye-tracking analysis of LLM responses. *Proceedings of the ACM on Human-Computer Interaction*, 9(3):1–29.

Päivi Majaranta and Kari-Jouko Räihä. 2007. Text entry by gaze: Utilizing eye-tracking. In I. Scott MacKenzie and Kumiko Tanaka-Ishii, editors, *Text entry systems: Mobility, accessibility, universality*, 2007, pages 175–187. San Francisco: Morgan Kaufmann.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Maria Mouratidi and Massimo Poesio. 2025. Comparing eye-gaze and transformer attention mechanisms in reading tasks. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 26–36, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Prateek Panwar, Sayan Sarcar, and Debasis Samanta. 2012. EyeBoard: A fast and accurate eye gaze-based text entry system. In *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, pages 1–8. IEEE.

Michael I Posner, Charles R Snyder, and Brian J Davidson. 1980. Attention and the detection of signals. *Journal of experimental psychology: General*, 109(2):160.

Andrzej W Przybyszewski, Albert Śledzianowski, Artur Chudzik, Stanisław Szlufik, and Dariusz Koziorowski. 2023. Machine learning and eye movements give insights into neurodegenerative disease mechanisms. *Sensors*, 23(4):2145.

Adrielli Tina Lopes Rego, Joshua Snell, and Martijn Meeter. 2025. What determines where readers fixate next? leveraging NLP to investigate human cognition. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 1–6, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.

Ronan G Reilly and Ralph Radach. 2003. Foundations of an interactive activation model of eye movement control in reading. In *The Mind's Eye*, pages 429–455. Elsevier.

Tiago Ribeiro, Stephanie Brandl, Anders Søgaard, and Nora Hollenstein. 2023. WebQAmGaze: A multilingual webcam eye-tracking-while-reading dataset. *arXiv preprint arXiv:2303.17876*.

Kshitij Sharma, Michail Giannakos, and Pierre Dillenbourg. 2020. Eye-tracking and artificial intelligence to enhance motivation and learning. *Smart Learning Environments*, 7(1):13.

Omer Shubi, Cfir Avraham Hadar, and Yevgeni Berzak. 2025. Decoding reading goals from eye movements. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5616–5637, Vienna, Austria. Association for Computational Linguistics.

Jaiteg Singh and Deepika Sharma. 2024. Automated detection of mental disorders using physiological signals and machine learning: A systematic review and scientometric analysis. *Multimedia Tools and Applications*, 83(29):73329–73361.

Joshua Snell, Sam van Leipsig, Jonathan Grainger, and Martijn Meeter. 2018. OB1-reader: A model of word recognition and eye movements in text reading. *Psychological review*, 125(6):969.

Ekta Sood, Prajit Dhar, Enrica Troiano, Rosy Southwell, and Sidney K. D'Mello. 2025. ScanEZ: Integrating cognitive models with self-supervised learning for spatiotemporal scanpath prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1132–1142, Vienna, Austria. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341.

Milan Turčáni, Zoltan Balogh, and Michal Kohútek. 2024. Evaluating computer science students reading comprehension of educational multimedia-enhanced text using scalable eye-tracking methodology. *Smart Learning Environments*, 11(1):29.

Bingbing Wang, Bin Liang, Lanjun Zhou, and Ruifeng Xu. 2024. Gaze-infused BERT: Do

human gaze signals help pre-trained language models? *Neural Computing and Applications*, 36(20):12461–12482.

Min Wang, Ao Xu, Chenxiao Fan, and Xiao Sun. 2025. Machine learning for predicting personality and psychological symptoms from behavioral dynamics. *Electronics*, 14(3):583.

Francesca Zermiani, Prajit Dhar, Ekta Sood, Fabian Kögel, Andreas Bulling, and Maria Wirzberger. 2024. InteRead: An eye tracking dataset of interrupted reading. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9154–9169, Torino, Italia. ELRA and ICCL.

Dongsen Zhang, Peipei Li, Zekun Li, Yiwei Ru, Huijia Wu, and Zhaofeng He. 2025a. Eye movements as images: A multimodal framework for eye movements representation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuhong Zhang, Jialu Li, Shilai Yang, Yuchen Xu, Gert Cauwenberghs, and Tzyy-Ping Jung. 2025b. Graph representations for reading comprehension analysis using large language model and eye-tracking biomarker. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7. IEEE.

Yuhong Zhang, Qin Li, Sujal Nahata, Tasnia Jamal, Shih-Kuen Cheng, Gert Cauwenberghs, and Tzyy-Ping Jung. 2024a. Integrating large language model, EEG, and eye-tracking for word-level neural state classification in reading comprehension. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3465–3475.

Yuhong Zhang, Shilai Yang, Gert Cauwenberghs, and Tzyy-Ping Jung. 2024b. From word embedding to reading embedding using large language model, EEG and eye-tracking. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE.

Gancheng Zhu, Zehao Huang, Xiaoting Duan, Shuai Zhang, Rong Wang, Yongkai Li, and Zhiguo Wang. 2025. Smartphone eye-tracking with deep learning: Data quality and field testing. *Behavior Research Methods*, 57(7):202.

Author Index