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.