Orthographic Complexity Moderates Eye Movements While Reading in Hindi, Along with Length & Frequency

Sidak Singh Kalra, Niket Agarwal, Ark Verma
Department of Cognitive Science, Indian Institute of Technology Kanpur, Kanpur, U.P., India.

Abstract

Past research on eye-movements during reading and comprehension has primarily focused on alphabetic scripts, such as the Roman script used to write European languages like English, Dutch, German, and Spanish, where classical measures like word length can be easily calculated by counting characters. However, this approach may not generalize to alphasyllabic languages like Hindi and other Indian languages written using the *Devanagari* script, where many characters depend on diacritic markers for proper pronunciation. This poses challenges in studying these languages in eye-tracking research, discourages eye-tracking studies with these languages. To address this gap, we asked 61 native Hindi speakers (L1) read Hindi text, while their eye-movements were being tracked. Results revealed that a complexity metric for the script predicts variables such as first fixation duration, gaze duration, single fixation duration, total reading time, and number of fixations. These results were also correlated with variables such as word frequency (van Heuven et al., 2014) for all eye-tracking measures.

Keywords: Reading; Eye-tracking; Psycholinguistics; Hindi;

Introduction

Research on eye movements and eye movement corpora is crucial for understanding word processing, vocabulary acquisition, language learning, and reading (Brysbaert & Drieghe, 2024). Eye tracking provides a robust way to study eye-movements during reading characterized by fixation counts, gaze duration, refixation probability etc. It is also insightful for studying bilingual activations. In recent years, multiple studies and corpora have been published showing eye-tracking measures in first (L1) and second (L2) languages (Siegelman et al., 2022). While much of the research has focused on European languages, such as Dutch (GECO by Cop et al., 2016), German (Postdam corpus by Kliegl et al., 2004) and English (Dundee corpus by Kennedy et al., 2013), studies on Asian languages, aside from Chinese (Sui et al., 2022) remain limited.

There are however, few previous studies such as one on the Potsdam-Allahabad Hindi Eye-Tracking Corpus by Husain et al. (2014) collected eye-tracking data from 30 participants reading 74 Hindi sentences, where the authors found effects of syllable length and bigram frequency on first-pass reading times and outgoing saccades. Although the authors accounted for factors such as word complexity, syllable length, unigram and bigram frequency, distance, and storage cost for sentence dependency. No study has yet examined the effects of word frequency (Zipf values) and word length measures (*akshara, matra*) on eye movement variables in Hindi.

In the current study, we tested the effects of word frequency and several variations of word length, unique to Indian languages, on first fixation duration, single fixation duration, gaze duration, total reading time, and refixation probability. The main goal was to replicate and extend well-established effects of word length and frequency on eye movement variables. Additionally, we aimed to test multiple versions of word length and complexity metrics specific to Indian languages.

Challenges specific to Indian languages

Research on Indian languages, such as Hindi, presents unique challenges due to fundamental script-level differences from the alphabetic scripts. These challenges include ill-defined parameters for word length, word complexity, and syllable structure, which complicate cross-linguistic comparisons.

Hindi is written in the *Devanagari* script, which consists of 52 letters (*varnas*) divided in: 13 vowels (*swars*), 33 consonants (*vyanjans*), 4 combined consonants (e.g., क्ष, त्र, ज्ञ, श्र) and 2 additional consonants (ड़, ढ़). Hindi uses alpha-syllabic characters i.e., *aksharas* rather than individual letters like English. In this system, consonants (*vyanjanas*) are the base units, and vowel signs (*matras*) are added to modify their pronunciation. These diacritics (*matras*) can appear to the right, left, above, or below these consonants (Fig. 1). Additionally, multiple consonants can be combined to form conjunct consonants (*sanyuktakshars*).

Figure 1: *Example of Devanagari script (used for Hindi) with akshara (in black, center) and possible matras (on top, bottom, left, right - superimposed)*

Most Indian languages follow a similar structure, with the *akshara* (character unit) as the fundamental building block. An *akshara* typically consists of a consonant (C) and a vowel (V), forming a [CV] unit, with words being constructed from one or more *aksharas*.

Hindi differs from other Western languages in terms of how syllables and vowels are defined in the language. In Hindi, each consonant (*vyanjana*) inherently carries a hidden vowel

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

(*schwa*). For instance, क (/*k*/) is both a consonant and an *akshara*, representing the [CV] unit /*ka*/. Diacritic vowel markers (*matras*) are added to consonants to modify the pronunciation, as in क + ा = का (/*kaa*/) and क + ि = कि (/*ki*/). These *matras* can be written in a linear (का) or non-linear (कि) fashion. Some of these markers (*matras*) are written to the top or bottom of the consonant (*akshara*), each modifying the pronunciation in a different way. (Fig. 1).

These script-specific features pose challenges in calculating fundamental parameters, such as word length, which are typically designed for alphabetic languages. In this study, we also explored several methods for calculating a length complexity metric that accounts for the unique characteristics of the Hindi script.

This study

The data for this study were collected as part of the multi-lab collaborative resource, the Multilingual Eye-Movements Corpus (MECO) (Siegelman et al., 2022; Kuperman et al., 2024). The corpus includes text reading data in English as a second language (L2) and 12 different first languages (L1s) in wave 1, and 13 distinct (L1s) in wave 2. Using this resource not only allowed us to explore these effects in Hindi (L1), but also enables us to compare the same effects in English (L2) for bilingual text comprehension. Additionally, this resource provides a unique opportunity to study eye movement patterns in other languages, as the texts in all languages were matched for difficulty and topics.

Complexity score

In this study, we present a novel method for calculating word length in Hindi and other Indian languages. For simplicity, we have kept this metric as straightforward as possible, calculating the weighted sum of *aksharas* and *matras* (and other diacritics), with a weight of 1 for each *akshara* and 0.5 for each *matra*.

Similar metrics have been used in previous studies, though they can be more complex to explain. For example, a study by Shallam & Vaidya (2019) developed a complexity metric for Malayalam words, that incorporated token frequency, morphology, and orthography. They calculated the morpheme count by summing the number of stems and suffixes, then normalized the value to produce a score between 0 and 1. For the orthographic component, they assigned a penalty of 1 for non-linear matras, and 0.5 for all other *matras*. They also assigned a penalty of 1 for each conjunct consonant (composite glyph).

Similarly, a study by Husain et al. (2014) developed similar metric for Hindi words. They assigned penalty of 1 for non-linear *matras*, 1 for each conjunct, and 0.5 for all other *matras*.

We drew inspiration from previously presented metrics and simplified them to make the explanation easier. As observed in earlier studies conducted in our lab, we found that *aksharas* have a greater effect on word recognition performance. Therefore, we assigned a weight of 1 to all *aksharas*. For all other diacritics, we used a weight of 0.5. For conjunct consonants, we also assigned a weight of 1.

Materials and Methods

Participants

Sixty-one Hindi (L1) – English (L2) bilingual participants (40 males, 21 females) participated in the study for course credit. Participants were matched for age and educational level, with a mean age of 21.11 years (SD = 1.90). All participants were recruited from undergraduate and postgraduate courses at a university. The mean subjective Hindi proficiency rating was 8.33 (SD = 1.06), and participants also completed an objective Hindi lexical proficiency test (HiLex; developed in our lab) using a lexical decision task paradigm, with a mean proficiency of 75.67% (SD = 6.41). All participants had normal or corrected-to-normal vision and did not report any reading or language-related disabilities. The study tasks were approved by the institute's research board.

Materials

All participants read 12 short texts in both Hindi and English. The English passages were sourced from the training materials for the ACCUPLACER Reading test. Text lengths in English ranged from 98 to 185 words, and the materials were appropriate for our intermediate-to-advanced sample of English L2 readers (for details, refer to Kuperman et al., 2023). The Hindi texts were matched in length, difficulty, and topic to their English counterparts. A total of 112 sentences were presented to participants, comprising 2,444 tokens. Detailed information about the texts in Hindi is provided in Table 1.

In addition to the reading tasks, participants completed a battery of individual differences tests and questionnaires for both languages. They also took a nonverbal IQ test (CFT20) and an abridged version of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007). For objective proficiency testing, additional non-eye-tracking measures were used, including the TOWRE (Test of Word Reading Efficiency), the HiLex (Hindi Lexical Proficiency test), and a spelling test.

The lexical characteristics for Hindi were obtained from the Shabd corpus (Verma et al., 2021). We used standardized Zipf values for frequency instead of the usual frequency per million words (fpmw), as Zipf values are independent of corpus size and are more commonly used for cross-linguistic comparisons (van Heuven et al., 2014). These values are on a logarithmic scale, where values of 1–3 represent low-frequency words and 4–7 represent high-frequency words. Word length in Hindi was defined in terms of the number of *aksharas*, *matras*, and *akshara+matras*. Additionally, a novel word complexity score was also calculated.

Table 1: *Stimuli used for Hindi reading*

Topic	ID	sentences (n)	words (n)
baseball (practice)	0	13	194
janus	1	10	212
shaka sign	2	8	187
doping	3	9	222
thylacine	4	9	175
environment day	5	8	162
monocle	6	7	148
wine tasting	7	8	238
orange juice	8	7	154
beekeeping	9	8	158
national flag	10	9	183
nature conservation	11	8	203
vehicle registration	12	8	208
Total (sum)		**112**	**2444**

Apparatus

The eye-tracking data for right eye was recorded by SR Eyelink 1000+ with a sampling rate of 1000 Hz. Calibration was performed by nine fixed points followed with validation. The experiment was designed in SR Research experiment builder. The experiment was performed in a quiet room and lights were kept dim during the main experiment. Participants were instructed to silently read the text on screen with their chins rested over the chinrest and a suitable finger placed over the desired key of the keyboard for navigating forward to the next part of the session. The drift check took place at the beginning of each trial, and calibration was monitored by the experimenter throughout the task and was redone if necessary. During the session, it was ensured that participants have minimal head movements while reading. The font used for Hindi was Kokila (size 28), which featured clear diacritic markings and single spacing. Participants were seated 70 cm away from a 24-inch screen (BenQ-XL2430T) with a resolution of 1920 x 1080 and a refresh rate of 60 Hz.

Procedure

The experimental session began with participants signing a consent form and filling out the LEAP-Q questionnaire (Marian et al., 2007). They then proceeded to an L1 reading task, in which they read 12 texts in Hindi silently for comprehension while their eye movements were recorded. After each text, they answered four yes/no questions. Following the L1 reading task, participants completed an individual difference battery in L1, which included the spelling test, HiLex, TOWRE and CFT-20. After a short break, participants completed the L2 reading task (in English). Although we have not analyzed the English eye-tracking data in this paper, it would be valuable to examine the relationship between L1 and L2 reading behaviors in future studies. The data can be freely accessed on OSF.

Analysis

We analyzed eye movement data at word level. Participants read overall 2444 words in Hindi. The key dependent eye-movement variables were: FFD, GD, SFD, TRT and RF. We calculated eye movement variables for all participants and correlated with word characteristics such as word-frequency (zipf), and multiple word-length measures for Hindi. (see Table 2).

Table 2: *Correlation values of eye-movement variables (FFD, GD, SFD, TRT and RF) with word features (zipf and word-length measures) in Hindi; highest correlation is shown in bold for each word-length*

Variable		***Pearson's (r)***	***p-value***
First fixation duration	*freq. (zipf)*	-0.252	< 0.001
	n_akshara	**0.229**	< 0.001
	n_matra	0.083	< 0.001
	AksMatra	0.193	< 0.001
	complexity	**0.222**	< 0.001
Gaze duration	*freq. (zipf)*	-0.520	< 0.001
	n_akshara	0.624	< 0.001
	n_matra	0.34	< 0.001
	AksMatra	0.590	< 0.001
	complexity	**0.633**	< 0.001
Single fixation duration	*freq. (zipf)*	-0.228	< 0.001
	n_akshara	**0.202**	< 0.001
	n_matra	0.051	< 0.001
	AksMatra	0.158	< 0.001
	complexity	0.188	< 0.001
Total reading time	*freq. (zipf)*	-0.564	< 0.001
	n_akshara	**0.653**	< 0.001
	n_matra	0.380	< 0.001
	AksMatra	0.628	< 0.001
	complexity	**0.668**	< 0.001
Refixations	*freq. (zipf)*	-0.572	< 0.001
	n_akshara	**0.747**	< 0.001
	n_matra	0.442	< 0.001
	AksMatra	0.442	< 0.001
	complexity	**0.763**	< 0.001

Word length measures were calculated as follows:

- ***n_akshara***: number of *aksharas* in a word
- ***n_matra***: number of *matras* in a word
- ***AksMatra***: *n_akshara* + *n_matra*
- ***complexity_score***: The complexity score assigns a weight of 1 to each *akshara* and 0.5 to each *matra* (diacritic).

Results

First fixation duration (FFD)

The first fixation duration for each word is defined as the time spent on fixating on a word for the first time during the normal reading, without any refixations. It is mainly associated with early visual processing and low-level visual information, hence importance for word identification.

We correlated the mean first fixation durations (FFD) for all participants across various word characteristics (Fig. 2). A significant negative correlation was found between FFD and word frequency (r = -0.25, p < 0.001), suggesting that shorter FFDs were associated with higher-frequency words. Additionally, FFD correlated positively with word length measures. The correlation was highest for number of aksharas and our word-complexity metric, (r = 0.22, p < 0.01), indicating that longer words required longer FFDs.

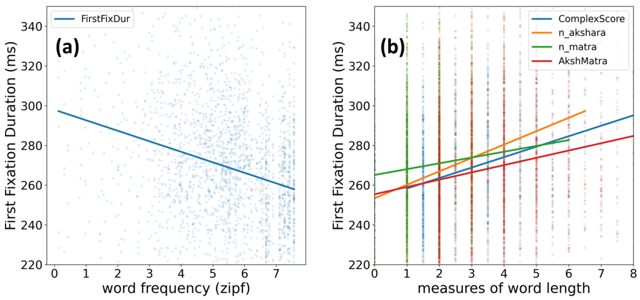

Figure 2: *First fixation duration with (a) word frequency (zipf); (b) measures of word length*

Gaze duration (GD)

The gaze duration for each word is defined as the summation of all fixation durations in the first run on the current word. In other words, it is the total time spent fixating on a word before moving to the next word (either through a forward saccade or regression) i.e. all fixations with run = 1. This is also an important metric that correlates with effortful reading.

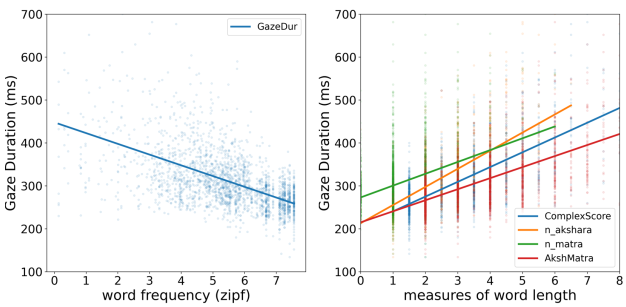

Figure 3: *Gaze duration with (a) word frequency (zipf); (b) measures of word length*

We correlated the mean gaze durations (GD) for all participants across various word characteristics (Fig. 3). A significant negative correlation was found between GD and word frequency (r = -0.52, p < 0.001), suggesting that shorter GDs were associated with higher-frequency words. Additionally, GD correlated positively with word length measures. The correlation was highest for word-complexity metric, (r = 0.63, p < 0.01), indicating that longer words required longer GDs. Word complexity metric was calculated as the weighed sum of aksharas and other orthographic features.

Single fixation duration (SFD)

The single fixation duration can only be calculated for a word, which is only fixated once. It is defined as the duration of fixation on a word, conditioned on whether the word is fixated exactly once.

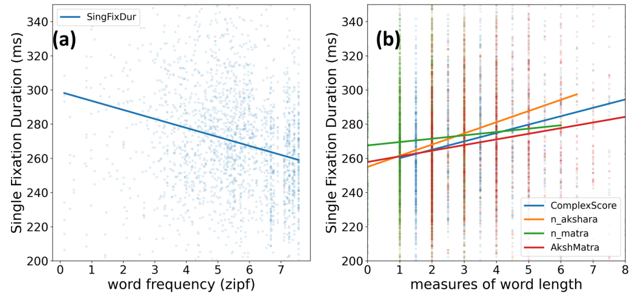

Figure 4: *Single fixation duration with (a) word frequency (zipf); (b) measures of word length*

We correlated the average single fixation durations (SFD) for all words, that are fixated only once, across various word characteristics (Fig. 4). A significant negative correlation was found between SFD and word frequency (r = -0.22, p < 0.001), suggesting that shorter SFDs were associated with higher-frequency words. Additionally, SFD correlated positively with word length measures. The correlation was highest for number of aksharas, (r = 0.20, p < 0.01), indicating that longer words required longer FFDs. It was not highest for word complexity metric.

Total reading time (TRT)

The total reading time or total fixation duration for each word is defined as the cumulative time spent in fixating on a word, regardless of whether it is first encountered or revisited after regression. It is an indicator of overall processing effort.

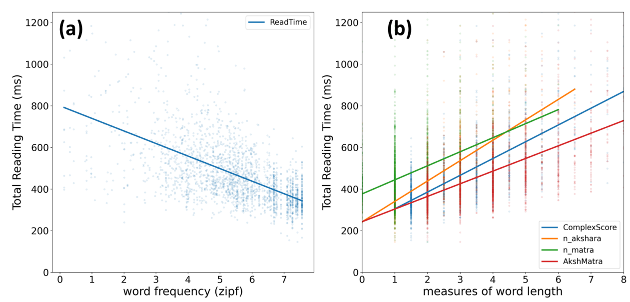

Figure 5: *Total reading time with (a) word frequency (zipf); (b) measures of word length*

We correlated the mean total single fixation durations (TRT) for each word across various word characteristics (Fig. 5). A significant negative correlation was found between TRT and word frequency (r = -0.56, p < 0.001), suggesting that shorter reading times were associated with higher-frequency words. Additionally, TRT correlated positively with word length measures. The correlation was strongest for word-complexity metric, (r = 0.66, p < 0.01), indicating that longer words required longer reading times.

Number of refixations (RF)

It is defined as the number of times a word has been fixated again (in the first run).

We correlated the average number of refixations for all words, across various word characteristics (Fig. 6). A significant negative correlation was found between RF and word frequency ($r = -0.57$, $p < 0.001$), suggesting that less fixations were associated with higher-frequency words. Additionally, number of refixations correlated positively with word length measures. The correlation was strongest for word-complexity metric, ($r = 0.76$, $p < 0.01$), indicating that longer words required more fixations.

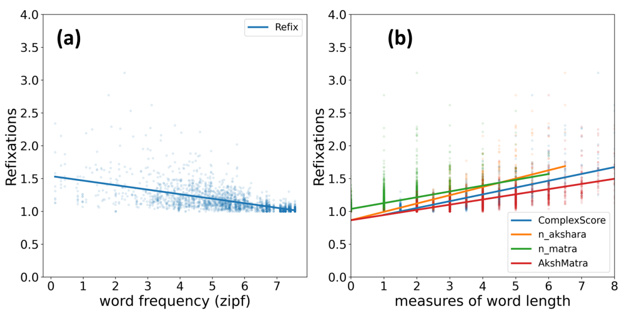

Figure 6: *Number of refixations with (a) word frequency (zipf); (b) measures of word length*

Discussion

We collected reading comprehension data from native Hindi speakers with English as a second language in naturalistic reading paradigm. The goal of the study was: (a) to replicate classic eye-tracking psycholinguistics findings for Hindi words, (b) to test novel measures of word-complexity for Indian languages, since classic word length can only be defined for alphabetic scripts.

We found strong effects of all lexical and orthographic features on the eye-tracking variables in Hindi, particularly word frequency (zipf values). However, for nearly all eye-tracking variables, the number of matras showed the weakest correlation, while the number of aksharas and our complexity score exhibited the strongest correlations. These results suggest that future studies should consider using the complexity score instead of word length when analyzing Indian languages. These results are also in line with previous studies such as Kumar et al., (2023; Rimzhim et al., 2021, Verma et al., 2021) where *aksharas* emerged as main predictors of word-length effects on RTs and accuracy.

In a follow-up study, it would be interesting to explore multiple versions of complexity metrics for Hindi. Additionally, comparing reading data from Hindi and English would be valuable, especially since all participants read both languages in a single session. It would also be worthwhile to examine whether participants' proficiency scores moderated the results. Also, given the MECO dataset is already public with more than 20 languages, it would be interesting for cross-linguistic comparison. However, more research on Indian languages is needed to verify the complexity metrics. This study paves the way for further investigation for robust eye-tracking reading experiments during lexical studies of abugida style (alpha-syllabic) languages.

References

- Brysbaert, M., & Drieghe, D. (2024). The use of eye movement corpora in vocabulary research. *Research Methods in Applied Linguistics*, *3*(1), 100093. https://doi.org/10.1016/j.rmal.2023.100093
- Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2016). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*(2), 602–615. https://doi.org/10.3758/s13428-016-0734-0
- Husain, S., Shravan Vasishth, & Srinivasan, N. (2014). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. Journal of Eye Movement Research, 8(2). https://doi.org/10.16910/jemr.8.2.3
- Kennedy, A., Pynte, J., Murray, W. S., & Paul, S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. Quarterly Journal of Experimental Psychology, 66(3), 601-618. https://doi.org/10.1080/17470218.2012.676054
- Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. European Journal of Cognitive Psychology, 16(1–2), 262–284. https://doi.org/10.1080/09541440340000213
- Kuperman, V., Schroeder, S., Acartürk, C., Agrawal, N., Bollinger, L. … Siegelman, N.. (2024). New data on text reading in English as a second language: the wave 2 expansion of the multilingual eye movements corpus (MECO). in Studies in Second Language Acquisition, Cambridge University Press.. https://hdl.handle.net/21.15107/rcub_reff_6934
- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech, Language, and Hearing Research, 50(4), 940–967.
- Shallam, R., & Vaidya, A. (2019). Towards measuring lexical complexity in Malayalam. c
- Siegelman, N., Schroeder, S., Cengiz Acartürk, Ahn, H.-D., Alexeeva, S., Amenta, S., … Marelli, M. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, *54*(6), 2843–2863. https://doi.org/10.3758/s13428-021-01772-6
- Sui, L., Dirix, N., Evy Woumans, & Duyck, W. (2022). GECO-CN: Ghent Eye-tracking COrpus of sentence reading for Chinese-English bilinguals. *Behavior Research Methods*, *55*(6), 2743–2763. https://doi.org/10.3758/s13428-022-01931-3
- van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

- Verma, A., Sikarwar, V., Yadav, H., Jaganathan, R., & Kumar, P. (2021). Shabd: A psycholinguistic database for Hindi. *Behavior Research Methods*, *54*(2), 830–844. https://doi.org/10.3758/s13428-021-01625-2