applied sciences

MDPI

Article

Cognitively Driven Arabic Text Readability Assessment Using Eye-Tracking

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

[1] College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; hendk@ksu.edu.sa (H.A.-K.); salman@ksu.edu.sa (A.A.-S.)
[2] The National Center for Data Analytics and Artificial Intelligence, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
* Correspondence: ibaazeem@kacst.edu.sa

Abstract: Using physiological data helps to identify the cognitive processing in the human brain. One method of obtaining these behavioral signals is by using eye-tracking technology. Previous cognitive psychology literature shows that readable and difficult-to-read texts are associated with certain eye movement patterns, which has recently encouraged researchers to use these patterns for readability assessment tasks. However, although it seems promising, this research direction has not been explored adequately, particularly for Arabic. The Arabic language is defined by its own rules and has its own characteristics and challenges. There is still a clear gap in determining the potential of using eye-tracking measures to improve Arabic text. Motivated by this, we present a pilot study to explore the extent to which eye-tracking measures enhance Arabic text readability. We collected the eye movements of 41 participants while reading Arabic texts to provide real-time processing of the text; these data were further analyzed and used to build several readability prediction models using different regression algorithms. The findings show an improvement in the readability prediction task, which requires further investigation. To the best of our knowledge, this work is the first study to explore the relationship between Arabic readability and eye movement patterns.

Keywords: Arabic language; eye-tracking; eye movements; human processing; machine learning; natural language processing; readability assessment; text difficulty

Citation: Baazeem, I.; Al-Khalifa, H.; Al-Salman, A. Cognitively Driven Arabic Text Readability Assessment Using Eye-Tracking. *Appl. Sci.* **2021**, *11*, 8607. https://doi.org/10.3390/app11188607

Academic Editors: José Ignacio Abreu Salas and Yoan Gutiérrez Vázquez

Received: 14 August 2021
Accepted: 13 September 2021
Published: 16 September 2021

1. Introduction

Written language is one of the primary sources for knowledge acquisition and communication [1,2], and reading is a complex series of actions that allows access to this knowledge. "Reading is a coordinated execution of a series of processes which involve word encoding, lexical access, assigning semantic roles, and relating the information contained in a sentence to earlier sentences in the same text and the reader's prior knowledge" [2]. While one is reading a text, readability is a crucial factor for successful comprehension [1]. According to Dale and Chall [3], text readability is defined as "the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at optimal speed, and find it interesting." These elements include typographical aspects, such as supporting graphics and tables, and readers' interest, which is affected by education level, for example, and writing style (such as text syntax). Another definition of readability by Klare, as reported by [4], is "the ease of understanding or comprehension because of the writing style."

Text readability assessment, which involves developing effective methods for determining readability level, has been a challenging research field since the 1920s [5]. This concern has become more significant in the information age, with the huge amount of written text available in different domains [1].

Although there are suggestions for analyzing a document's comprehension and clarity by observing a sample of its audience or by having professional editors review written documents, these operations are often impractical due to their labor expense, subjectivity, and time consumption [6]. Consequently, the demand has risen for technological tools that help create and edit documents to make them clear and understandable according to how readers comprehend the text [6]. However, because readability depends on many factors that enable readers to process a text, cognitive linguists have pointed out two main features that might influence readers' successful comprehension [2,6,7]: readers' background knowledge and the contents of the text itself.

Therefore, in addition to the text characteristics, the problem of readability is mainly rooted in the characteristics of the reader, as indicated by Dale and Chall [3], and readability is not just determined by literacy skills, but also by the readers' personal characteristics and backgrounds [5]. Given this, text being considered readable or not and the features used should be built upon the cognitive characteristics and skills of a population. However, although the primary goal of text readability assessment models is to guarantee the predictions that reflect readers' difficulties, the majority of these models are annotated using experts' judgments and not built using the actual audiences' reading performance [8,9].

Recently, this approach has been criticized in education research. One way to target this limitation is building corpora using humans' physiological data, such as brain electroencephalography (EEG) signals and eye-tracking data to anticipate human processing effort during reading. Collecting this psychometric information has improved many natural language processing (NLP) tasks such as named-entity recognition, sentiment annotation complexity, and sarcasm detection. It allows for a better understanding of a human's cognitive behavior [10–12]. As indicated by [2], real-time readability assessments from behavioral signals could be a significant sign of text readability. It is one of the potential research directions for the readability assessment task.

According to Mathias et al. [10], there is a significant correlation between reader eye movement and the cognitive processing of texts. This is linked to Just and Carpenter's [13] eye–mind hypothesis, which proposes that there is no discernible delay between eye fixation and cognitive processing. Eye-tracking provides objective measurements of reader behavior, and it is effective in detecting processing problems during naturalistic reading [14]. Recent advances in the development of eye-tracking technology have led to research into the indirect behavior of readers and text difficulties [12,15,16].

However, substantial scholarship has focused on the quantification and automatic calculation of textual complexity in accordance with different linguistic features [11]. However, few studies have used eye-tracking to improve text readability. Eye-tracking experimentation has only been used for evaluation. Thus, while using gaze data to explore reading processes is not a new idea [17], it remains underexploited as a research tool in readability assessment [11]. Eye-tracking experimentation may well be time and resource intense. Still, it represents a more naturalistic approach to exploring reading processes and permits text re-reading to be examined [8,9].

Motivated by the potential benefit of using physiological eye-tracking data and the insufficiency of the investigation of these data in text readability prediction, this research proposes incorporating eye-movement data for automatic readability assessment, particularly for the Arabic language. Offering human reading data while reading Arabic text could serve different applications in addition to readability assessment, which would be a great contribution for Arabic. The study is concerned primarily with eye-tracking and linguistic text features associated with text difficulty. Thus, a document's graphical components that may affect readability have not been considered, such as font size, images, or tables [2].

The main contribution of this study is to discover how incorporating eye-tracking data for Arabic readability assessment has the potential to improve this task, particularly given that research on the readability of Arabic texts has used simple features that do not consider the reader in its prediction. This prevents these simple measures from reflecting the correct level of difficulty. To the best of our knowledge, this is the first study to incorporate eye

movement data into the Arabic text readability task. Hopefully, this study's results will help evaluate and adjust current and future Arabic texts found in educational, medical, industrial, and other contexts. Eventually, effective readability assessment will contribute to increasing the prevalence of Arabic written information [18]. Additionally, assessing reading difficulties will benefit not only Arabic natives but also Arabic learners, based on the results of [12,17,19,20], which indicate that measuring text difficulty for a language's learners can be efficiently estimated by the native speakers of this language.

The rest of the paper is organized as follows. Section 2 describes some required background knowledge related to text readability assessments and eye-tracking. Section 3 surveys existing studies targeting readability assessment with and without using eye-tracking data. Section 4 illustrates the primary research objectives and presents a conducted pilot study and all experimental procedure details. Section 5 illustrates the results and a discussion of them. Finally, we conclude in Section 6 with future work.

2. Background

This section illustrates some required background knowledge related to automatic text readability assessment in general and for Arabic in particular. It also explains the concept of eye-tracking technology and reading. Readability tests predict the difficulty of texts for potential readers [21]. There is extensive research in this area [6], and many studies have attempted to investigate how to measure text readability to produce clearer documents. Readability measurement tools have two main categories. The first category relies on traditional readability formulas in its readability estimation, while the second category utilizes machine learning approaches [18]. The following subsections shed light on some traditional (or classical) and data-driven, machine-learning-based text readability assessment approaches.

2.1. Overview of Classic Text Readability Assessments

Classic or traditional readability assessments consider semantic (familiarity of words, phrases) and syntactic (complexity) features to be the main features of their measurements [1,2]. Generally, classic readability measures work better with a combination of semantic and syntactic features [2]. We conducted a review of well-known English readability formulas, followed by a review of available readability formulas initiated for Arabic.

2.1.1. English Classic Readability Measures

Assessing English readability has long been studied in the literature. Since the beginning of the last century, hundreds of mathematical formulas have been developed and used for assigning English text to its correct readability level [22]. These formulas have been used widely in the educational context.

For readability measures that rely on words and sentence lengths for their calculations, such as [23,24], the computation is based on the assumption that a text's readability becomes more challenging if it contains longer terms and sentences [6]. On the other hand, vocabulary-based measures, such as the Chall and Dale [25] formula, represent a significant variation on classic readability measures. They measure the semantic difficulty of a text's words based on their presence or absence in a predefined vocabulary resource. Thus, word difficulty depends on familiarity: low-frequency words indicate higher difficulty. Some of the most commonly used readability measures for English are the Flesch [23] reading ease test, the Gunning [24] fog index, smog factor grading [26], the Coleman and Liau [27] index (CLI), Kincaid et al. [28] grade level scoring, and the Chall and Dale [25] readability formula.

2.1.2. Arabic Classic Readability Measures

There has been a growing interest in Arabic language processing and translation. Still, despite this increasing interest, only a few researchers have tackled the problem of finding a proper Arabic readability index [29]. Similar to English and other languages, research

on Arabic readability of texts for first-language readers has started with the development of readability formulas and progressed to the use of statistical analysis approaches and machine learning algorithms for classification [30]. A summary of these formulas is illustrated in Table 1.

Table 1. Summary of readability formulas for Arabic.

Formula Name	Year	Formula Mathematical Calculation
Dawood Readability Score [31]	1977	Dawood Readability Score = $-0.0533 \times W - 0.2066 \times S + 5.5543 \times p - 1.0801$ where "W" is the average number of characters per word, "S" is the average number of words per sentence, and "*p*" is the average frequency.
Al-Heeti [32] Grade Level	1984	Al-Heeti Grade Level = (AWL × 4.414) − 13.468 where "AWL" is the average number of characters per word.
A Corpus-Based Readability Formula [33]	2013	Readability score of a sentence = Total reversed ranking of each word in a sentence/No. of words per sentence.
Automatic Arabic Readability Index (AARI) [34]	2014	AARIBase = (3.28 × NOC) + (1.43 × ACW) + (1.24 × AWS) where "NOC" is the character count, "ACW" is the average number of characters per word, and "AWS" is the average number of words per sentence. Afterwards, the AARI base formula was mapped to grade levels as follows: Grade Level = (AARI + 472.42)/1046.3
Open Source Metric for Measuring Arabic Narratives (OSMAN) [29]	2016	OSMAN Score = 200.791 − 1.015 × (A/B) − 24.181 × (C/A + D/A + G/A + H/A) "where 'A' is the total number of words counted using [the] Stanford Arabic word tokenizer, 'B' is the total number of sentences counted automatically using common delimiters to split text into sentences, 'C' is the number of hard words (words with more than 5 letters—Long words); the word length was counted with the absence of diacritics in order to avoid counting the diacritics as letters, 'D' is the number of syllables in a word, 'E' is the total number of characters ignoring digits, 'G' is the number of complex words in Arabic (words with more than four syllables), 'H' is the number of 'Faseeh' words (complex word with any of the following Arabic letters ('ء', 'ئ', 'وء', 'ذ', 'ظ') or ending with ('وا', 'ون')" [29].
Computational Formula for Arabic Reading Materials Among Non-Native Students [35]	2021	Y= 31.830 − 0.298 X1 − 0.178 X2 + 0.043 X3 + 2.444 where Y = readability level, X1 = common and frequent words, X2 = conjunction and punctuation, X3 = average number of words per sentence; Constant = 31.830 and Error = 2.444

2.2. Machine-Based Readability Measures

There are many limitations associated with the traditional readability formulas, despite their simple assumptions, such as the ignorance of text noise. Hence, they assume a text is composed of only well-formed sentences [2]. Additionally, they are non-effective with non-traditional documents, such as Web pages [2]. Furthermore, although readability research began a century ago [36], readability formulas are built on superficial text features. They do not consider in their calculation all components of texts and languages that are theoretically anticipated to contribute to the comprehension and difficulty of texts [18,33]. Many readability-related features, such as discourse coherence and syntactic ambiguity, are ignored [1,2,37]. Consequently, readability formulas might be misused by writers who wish to achieve higher readability scores by shortening their sentences. Conversely, these

texts tend to have lower cohesion and coherence, as sentences become very short and difficult to comprehend [6] given the previous limitations.

Several technological tools have been proposed for analyzing and unraveling the semantic characteristics of documents [6]. The early 2000s, with the increasing amount of textual data and the development of more complex prediction models, saw the rise of what was called "the 'AI' (Artificial Intelligence) approach to readability" [2]. NLP technologies and machine learning (ML) methods have been adopted in much text readability assessment research. Compared to traditional readability measures, this integration has shown significant advancement in the accuracy and reliability of this assessment [1,2,36,37].

2.3. Eye-Tracking and Reading

Eye-tracking is a technique performed by employing several concealed near-infrared illuminators that generate reflection configurations on the cornea. Image sensors can then detect the presence of the reader's eyes and acquire data at a rapid sampling speed. A three-dimensional model of the reader's eyes can be configured by processing the data, pinpointing the pupil's location, and recognizing apposite reflections from the illuminators, along with their accurate coordinates [38].

Eye-tracking technology can be utilized to investigate a range of events relating to language processing. A number of the linguistic elements explored using eye-tracking include auditory, visual, and combined auditory and visual processing [39]. Since eye-tracking offers real-time information, it is the subject of expanding research, especially in applied linguistics and learning a second language. Eye-tracking has been deployed as an alternative to conventional assessments (e.g., cloze tests, think-aloud protocols, and interviews) [39].

In studies relating to subjects' reading [8,39,40], eye-tracking data are often used to perceive and to interpret the following reading measures:

1. Saccades: the participant's swift eye motions between areas of text. Longer saccades imply text that is straightforward and easily comprehended; short saccades suggest a greater challenge and time required to digest the language [40].
2. Fixations: points between saccades where the reader's eyes halt. Brief and infrequent fixations suggest text is readable; longer and numerous fixations are related to text difficulty and imply that the reader may be struggling with interpretation [12,40].
3. Regressions: backward movements to return to earlier areas of text. Short regressions may imply local challenges, whereas more prolonged regressions may indicate the text's general level of difficulty and lack of clarity [12,39].

In common with other reaction time tests, eye-tracking employs two empirical theories during subjects' gaze data acquisition [39]. First, the time duration of gaze fixation represents the cerebral effort necessary for processing; that is, prolonged or more frequent fixations imply a higher degree of cognitive focus and vice versa. Second, the object of the subjects' gaze is the subject of concentration. In view of these assumptions and appreciating that the spectrum of fixations and saccades can offer significant information about both the reader and the words on the page [41], eye motion can be deployed to elucidate the cognitive processes related to reading [12].

From a different perspective, the user should be aware of some limitations related to eye-tracking techniques that may influence their usefulness in studies [40]. Founded on the empirical assumptions relating to eye-tracking deployment, the gaze is linked to subjects' focus. Yet, realistically, focus can additionally be placed on items within the gaze periphery. This arises owing to the complexity of the human ophthalmic system. This restriction necessitates that the two options are embraced (i.e., the gaze does or does not correspond to the subject's precise focus). Additionally, information loss or inaccuracies may arise during eye-tracking experiments. Causes include physiological features of certain subjects (e.g., elongated eyelashes, make-up, or heavy eyelids), and external factors (e.g., reflective glasses or contact lenses). Moreover, since eye-tracking cameras operate with a set frequency, some eye motions may fail to be recorded. Furthermore, since the

recorded signal is smoothed out to subtract anomalies such as blinking or peripheral eye deviations, some data could be erroneously deleted.

3. Related Work

It has long been a challenge to discern satisfactory methods for measuring readability levels in order to find appropriate texts [1,2,4,6,42]. Besides the readability formulas mentioned in Section 2, this section reviews readability studies that tackled automatic text readability assessment for Arabic and other languages using machine learning approaches. Because "some researchers have shown that the factors affecting English readability can be useful for readability research in other languages, including Modern Standard Arabic (MSA)" [43], we have focused mainly on readability research for English as a starting point in our review, with some research targeting other languages [44–47].

3.1. General Text Readability Assessment Studies

Collins-Thompson [2] published a survey in 2014 that investigated studies in computational methods for readability assessment for different target populations. Because Collins-Thompson investigated studies in computational methods for readability assessment published up to 2014, readability studies that were conducted in 2014 and later have been reviewed briefly subsequently. Previous studies considered the automated assessment of text readability from different dimensions, such as Balyan et al. [1] and Crossley et al. [48].

One text feature that plays a significant role in readability and has been widely used in many studies is text cohesion and coherence [6,49–51]. A number of studies such as Baazeem [6], Mesgar and Strube [7,37,49] and Zhang et al. [52] have shown a strong correlation between text cohesion and manual readability judgments.

Recently, neural networks (NNs) have shown impressive advancements in several domains, such as computer vision (CV) and NLP [46,53]. In NLP, NNs have shown advanced performance in semantic-related tasks when sufficient data are provided [46]. Given the advancement achieved from deep learning in many NLP tasks, it can successfully improve readability assessment tasks. Yet, surprisingly, deep neural models have not been employed significantly for readability evaluation tasks similar to their current use in other NLP tasks, such as question answering (QA) and machine translation.

For English, NNs have been proposed for the purpose of readability measurement since 1994 [53]. The use of NNs in readability research has not evolved significantly since that time; still in use are standard machine learning classifiers and their associated feature engineering [46]. However, in the past few years, some research has modeled text readability in terms of coherence using NNs, such as Li and Hovy [54], Mesgar and Strube [37], Logeswaran et al. [51], Xu et al. [50], Azpiazu and Pera [45], and Martinc et al. [46].

3.2. Arabic Text Readability Assessment Studies

As observed by Marie-Sainte et al. [55], although readability prediction has been studied for a long time for English and certain other foreign languages, Arabic readability has only recently been considered by the Arab NLP community. Cavalli-Sforza et al. [30] reviewed the current trends in Arabic automated text readability assessment. Some Arabic readability assessment studies that successfully transformed many of the used features in other languages to Arabic [43] are Fouad and Atyah [22], Forsyth [43], Shen et al. [56], and Nassiri et al. [57]. Given that Arabic readability research is still in the very early stages, [58] promoted the investigation of this topic by comparing Arabic readability tools from different usability dimensions and proposing some enhancements.

Al Jarrah [18] represents the first attempt to use NNs for Arabic readability assessment by comparing their performances to a number of traditional machine learning approaches. The most recent Arabic readability assessment study is the study of Khallaf and Sharoff [59]. By modeling sentences by using sentence embeddings only, they found that fine-tuned Arabic-bidirectional encoder representations from transformers (BERT) show the best

performance in difficulty prediction tasks compared with the other embedding types. Because of the similarity between all BERT-like models in their architecture, the authors attributed the Arabic-BERT model's better performance to its training corpora.

3.3. Eye-Tracking in Reading Studies

Eye-tracking is capable of discerning problems with processing in the course of naturalistic reading [11]. Cognitive psychologists have long known that the study of eye movement configurations assists in the comprehension of cognitive processes associated with reading, comprehension, and related problems [8]. For this reason, eye-tracking research has been extensively explored within cognitive science in order to expand scholarship pertaining to both reading processes and difficulties [9,60], as per [13,61–64].

The research conducted by Just and Carpenter [13] explored eye movement patterns and cognitive processing loads in students' readings extracted from scientific papers. Just and Carpenter's eye–mind assumption hypothesizes that the eye lingers on each word for the duration of time it takes to process the word. The garden path theory of sentence comprehension was explored in the research of Frazier and Rayner [61], who proposed that recovery from preliminary sentence misinterpretation is typically achieved through a limited number of definable strategies. Moreover, there is a somewhat rigid correlation between the location and length of fixations and the processing operations linked to the specific text. This was verified in the work of Rayner et al. [63], which demonstrated that there is a correlation between text difficulty and fixation count and duration.

The most common reading time metrics comprise first fixation duration, first-pass time, and total reading time. These measurements were all investigated in Liversedge et al. [64]. They recommended that psycholinguistic experiments include two additional time measures: regression path reading times and re-reading times. Using these new measurements, which determine temporally adjacent fixations, helps researchers to precisely understand eye movement patterns linked to the encountering of textual difficulties.

Different eye movement recording systems have been developed over the last twenty years, thereby rendering the collection of quality recording a more viable proposition [65]. Therefore, eye-tracking and its implications have been employed in multiple linguistics and psycholinguistics studies to identify the text spans that attract or deter eye movements [40], such as [14,39,40,65–68].

3.4. Eye-Tracking in Text Readability Assessment

This section sheds light on some of the last five years' work on integrating eye-tracking data in readability prediction models using machine learning methods, benefiting from the relationship between eye movements and human reading.

A "readability" score indicates a document's recommended audience [16]. Using users' eye movements allows more personalization when predicting reading levels; consequently, appropriate texts are recommended to readers by considering different factors [8,15]. A number of studies have targeted this problem, such as Chen et al. [15], Copeland et al. [12], and Vajjala et al. [8]. Because reader effort is a significant factor in reading and comprehension, Mishra and Bhattacharyya [17] developed an approach for quantifying reading effort called "scanpath complexity". Garain et al.'s [16] work aimed to detect reader-specific difficult words in a document.

However, although using eye-tracking measures have been proven effective in examining human–computer interactions [12], gathering enough real-time reading behavior data and building human-annotated training corpora remains time and labor intensive [8]. Consequently, whereas eye-tracking reading times could be used as indicators of increased linguistic complexity [11], researchers have begun to look toward predicting reading times through linguistic and psycholinguistic features instead of taking them from the eye data directly. The authors of [8] noted that "models based on automatically predicted reading times present themselves as an attractive alternative to the current methods" to meet NLP

demands for automatic complexity measures. Examples of studies that use this concept are Singh et al. [11], Gonzalez-Garduno & Søgaard [20,41], and Leal et al. [19].

The reading research literature notes that words that are "predictable" take less time to read and may even be skipped over [69]. The quantification of this word prediction is known as a surprisal measure. Given this benefit, the surprisal psycholinguistics feature has received special attention for modeling human sentence processing. Motivated by the excellent perplexity results accomplished by the use of NNs [70], Goodkind and Bicknell [71] investigated the degree to which surprisal estimated from computational language models (LMs) could be trusted and how LM quality and type variations affect surprisal's predictive power. Another study by Aurnhammer and Frank [72] compared the performance of variants of recurrent neural networks (RNNs) in estimating word surprisal values for words used in texts that were shown in self-paced reading, eye-tracking, and electroencephalography experiments. Merkx and Frank [69] compared RNNs, particularly gated recurrent units (GRUs), as representative of RNNs, and transformer-based LMs in predicting human reading data following the approach of [72]. Complementing and enhancing the previous research [69,71,72], Wilcox et al. [70] performed a much deeper investigation of the models' prediction behaviors by studying how their language models affect the estimated surprisal and subsequently the estimated human gaze using different datasets with different sizes.

3.5. Discussion

Based on the reviewed research, eye-tracking measures can assess text readability, and they are worth further investigation [19]. However, the current readability prediction and eye-tracking literature has some limitations. There is still a gap in deciding about the advantages of integrating different types of eye-tracking measures that have different reading indicators. The majority of the studies in the current literature seem to focus on eye-tracking reading measures such as first-pass reading times and total reading times. Ignoring some eye movement measures during human reading and focusing only on a few of them seem not to reflect a reader's behavior or the encountered difficulties at different reading points [17,39,64,73,74]. Additionally, it is unclear what eye-tracking measures are the most predictive of readability. Additionally, there is still a gap in investigating how different machine learning methods, with different learning behaviors, can capture texts' reading difficulties and thus predict the appropriate readability level using these eye-tracking data. Moreover, there is a lack of available real-time human reading corpora, which will affect further advancement in this domain. This shortage is very obvious for Arabic, which suffers from a scarcity of available corpora, compared to languages with richer resources [30].

4. Materials and Methods

The Arabic language has been defined by its own rules and has its own characteristics and challenges. Because Arabic readability prediction is much more recent compared to other languages [30], the use of behavioral signals, specifically eye-tracking data, has not been considered for quantifying Arabic readability measures. There is still a gap in determining how using eye movement data could benefit Arabic readability prediction, which requires further study. Thus, we will limit our scope to investigating the potential integration of Arabic readability assessment and gaze data. Consequently, we target the following research questions:

1. What is the effect of using eye-tracking features in assessing Arabic text readability?
2. Which eye movement features contribute the most to Arabic readability prediction?
3. Which ML model has better performance in modeling human reading difficulties and, thus, better readability prediction using these eye-tracking data?

To investigate the potential of using eye-tracking features in Arabic readability assessment tasks and gain insight into their usefulness as indicators of Arabic text difficulty, we

conducted an exploratory study. We summarize the main steps of the study in Figure 1. The details of the study are as follows:

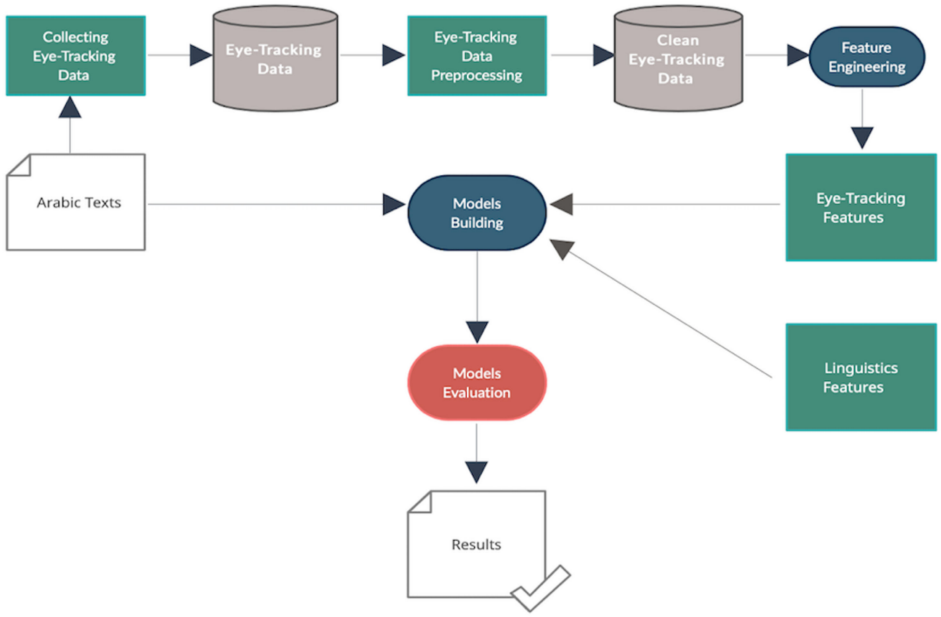

Figure 1. Research methodology summary.

4.1. Materials

We used six classical Arabic poems from [75] as experimental texts. These texts had varying readability levels based on the era in which they were written. Because Arabic has evolved and changed over the centuries, the two used categories of texts differed in terms of their structure and word complexity. Difficult texts were represented by poems written in the pre-Islamic era. There were three difficult-to-read poems, each containing eight sentences and approximately 38 words. Easy-to-read texts were represented by poems written in the Andalusian era. There were three easy-to-read poems, each containing eight sentences and approximately 36 words.

Additionally, to benchmark and quantify the readability level of the used texts, we used OSMAN [29], version 3.0 (released 2020), as a state-of-the-art tool for Arabic readability assessment. This tool was developed at Lancaster University, Lancaster, United Kingdom, as a customized Arabic version of some traditional English readability formulas [29].

Table 2 delineates the main characteristics of the text used in the study using OSMAN, where lower scores imply a lower readability level, whereas higher values imply a higher readability level. As indicated by [29], Faseeh words are words that have six or more characters and end with any of the following letters: ‘ء’, ‘ئ’,‘وء’, ‘ذ’, ‘ظ’, ‘وا’, ‘ون’. We used Arabic text with diacritics before calculating the OSMAN readability scores. Adding diacritics to the Arabic text was based on the OSMAN recommendation and a previous study [76], which used eye movement to examine the effect of three Arabic text forms (total diacritics, partial diacritics, and without diacritics) on text ambiguity and understanding of the text. The results of this study showed that using diacritics throughout the text affected the eye movement of the participants because it improved their comprehension and reduced the lexicon ambiguity of the Arabic text compared to the other two text forms.

Table 2. Summary of the main features of the texts using OSMAN.

File Name	OSMAN Readability Score	Sentences Count	Words Count	Syllables Count	Faseeh Words Count
		Andalusian Era Poems			
Text 1	157.15	8	38	48	0
Text 2	144.58	8	33	54	0
Text 3	163.60	8	39	36	1
		Pre-Islamic Era Poems			
Text 4	101.99	8	37	120	1
Text 5	123.92	8	39	102	2
Text 6	114.10	8	40	113	1

4.2. Participants

We collected data of the eye movements of 41 participants (19 males, 22 females, aged 25–50 years) in February and March 2021. All participants were educated, native Arabic speakers with good Arabic reading fluency; thus, we assumed that the texts should match their reading abilities. They all had nearly identical cultural backgrounds, but they came from different specialties (humanitarian, scientific, and medical). We collected some demographic data and their reading habits, such as Arabic reading frequency. Each participant read and signed a consent form. All participants' data were anonymous, and there was no connection between the personal information provided and the data. To overcome difficulties associated with the eye-tracker limitations illustrated in the background section, we adopted an exclusion criterion of vision problems. To avoid the possibility of tracking loss or error (the calibration problem), we asked the participants not to wear reflecting eyeglasses, contact lenses, or makeup.

4.3. Setup

We recorded the actual measurements of eye movement tracking using SR Eye-link 1000 plus eye tracker [77] located in the Applied Linguistics Research Lab, Prince Sultan University. We used the Desktop Mount mode, where the eye tracker is placed just below the display monitor that the participant is looking at in order to capture his/her eye movements, with a head support and a viewing distance of 72 cm. The display monitor was a DELL E1917HV with a refresh rate of 60 Hz, a screen resolution of 1366 $\times$ 768 pixels, and a landscape orientation. We designed the actual experiment using SR Experiment Builder software. The used textual contents shown on the screen have the following independent variables: black color on a white background and a font size of 14 points. To make reading the poems easier, we aligned the texts to match the common poem format in Arabic. Because doing this was impossible when entering text, we entered the text as images instead of text when designing the experiment. The text was clear with a large font and enough spacing to be readable from the computer screen.

4.4. Experimental Procedure

All the experiments were conducted under experimental lab conditions. Because we conducted this research to understand the relationship between eye movements and text readability, the procedure involved collecting participants' eye movements. At the same time, they read six Arabic texts, one by one (so each participant was involved in six trials). We asked participants to read aloud to help us guarantee that they actually read the text and to enable us to hear how fast and how well they were reading. Each participant read all of the easy- and difficult-to-read texts. In order to minimize any possible influence of text order on reading [8,67], we randomized the order of the displayed texts in each session.

At the beginning of the experiment procedure, we provided a brief description of the texts to the participants. The session started by getting some information about each participant and his/her Arabic reading frequency, type, and amount per month. After that, the participant sat in front of the monitor and rested his/her head on the head support.

To ensure that we obtained accurate measurements, we conducted a calibration; crosses appeared on the screen and a participant's eye had to follow them.

After successfully passing the nine-point calibration and validation that aimed to capture the actual gaze position on the display screen (a good calibration status for starting the experiment is to obtain a symmetric three-by-three grid and a good validation has an average error less than 0.5 and maximum error < 1), the instructions screen appeared. Subsequently, the participants read six texts in the following sequence: drift correction, reading 1st text (Trial 1), drift correction, reading 2nd text (Trial 2), and so on, until drift correction, reading 6th text (Trial 6). After finishing the reading on a screen, they pressed any keyboard button to move on. Since the reading time for each text was open and not restricted to a specific time, a participant had to go to the next page immediately after finishing reading and before giving his/her feedback so that the device did not collect irrelevant data. Performing drift correction after each reading was important to avoid drift in the gaze position [78]. Thus, any problem in the gaze position would be detected.

In addition to the previous reading task, we requested participants to verbally give their subjective readability about each text reading on a scale of four readability levels: Easy (E), Medium near to Easy (M/E), Medium near to Difficult (M/D), and Difficult (D). Each session lasted for approximately one hour, including calibration and the subsequent subjective readability questions.

4.5. Pre-Processing of the Eye-Tracking Data

We used SR Data Viewer software to display, manipulate, and analyze the data collected by the eye tracker during the experiments. Although the majority of reviewed studies have focused on assessing readability at the document and sentence level, we narrowed our attention to measure readability on the word level. The reason is that based on the observed behaviors of the participants during reading and based on the feedback or readability judgment from the participants, difficulty in reading was sourced primarily to the difficulty of the words. The participants did not struggle with the texts' structure, even though the participants may have been unfamiliar with the structure. Thus, for each text, we tokenized all texts word by word by defining each word to be an interest area (IA), which is the area under analysis in the screen [10]. There were 221 tokens in total and a vocabulary size of 186 words.

Additionally, based on the observed behavior of some participants who scanned the screen randomly and looked at white space at the beginning of the reading, we removed these fixations as they were not part of any interest areas. Additionally, we adjusted drifted fixation events manually to match the interest areas. Based on previous research, "if a portion of text causes the reader difficulty, then this difficulty can spill over and affect processing of subsequent text" [64]. Thus, we did not remove stop words in order to capture the effect of reading easy and difficult words. Additionally, although there have been some suggestions of merging fixations of less than a minimum threshold, we limited ourselves to the manual correction of drifted fixation events to avoid any loss of significant data.

Figure 2 shows examples of the output of the eye tracker during a participant's readings of two texts with different readability levels. The text on the right has a higher readability score compared with the text on the left. Each orange box represents an interest area—a word, in our case. The red circles represent the fixations, where the size of the circle correlates with the time spent in this fixation (fixation duration), and the yellow arrows represent the saccades. Note that the Arabic language is read from right-to-left.

Figure 2. Examples of the eye-tracker output from two texts that show fixations and saccades of one of the participants when reading the two texts. The text on the right is a poem written in the Andalusian era with a higher readability score than the text on the left, which is a poem written in the pre-Islamic era. Orange boxes represent interest areas, the red circles represent the fixations, the numbers above these fixations represent their durations, and the yellow arrows represent the saccades.

As shown in Figure 2, a less readable text has shorter saccades, more fixations, and more regression, which may imply reading difficulty. More specifically, they demonstrate that it is possible to establish the frequency, duration, and exact times of word fixations during reading, along with data related to revisited text when challenges arise [12].

Additionally, eye fixations could be seen in fixation maps or heatmaps (Figure 3) that aggregate the fixations of all the participants when reading texts with low and high readability scores and show the words that are fixated upon and the duration of these fixations.

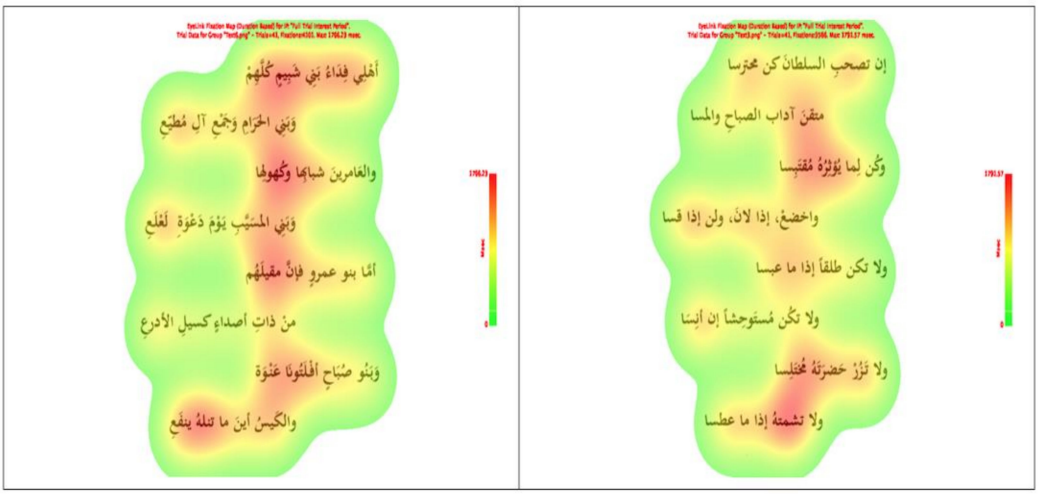

Figure 3. Examples of heatmaps built from two texts by averaging the fixations of all participants when looking at the two texts. The text on the right is a text, which is a poem written in the Andalusian era, with a higher readability score than the text on the left, which is a poem written in the pre-Islamic era.

4.6. Feature Engineering

In order to construct our training data, we extracted the aggregated values of many eye measures across all participants for each word in a given text using Data Viewer. These features offered a quantitative measure of assessment for each word in the vocabulary. Table 3 shows the selected eye-tracking features and their descriptions as per [78]. We selected these features based on their indications in reading, as mentioned in the background section and based on several previous studies [8,12,14,40,63,68].

Table 3. Used eye-tracking features for each interest area (IA).

Eye-Tracking Feature	Description
IA_DWELL_TIME	The summation of all fixations' durations that occurred on a specific IA in a trial
IA_DWELL_TIME_%	The percentage of the total trial time that a participant devoted to fixating on a specific IA
IA_FIRST_FIXATION_DURATION	The duration of a participant's fixation on a specific interest area for the first time in a trial
IA_FIXATION_COUNT	The total number of fixations for a participant on a specific IA in a trial
IA_FIX_COUNT_%	The percentage of the total number of fixations in a trial that a participant devoted to a specific IA
IA_RUN_COUNT	The count of how many times a participant entered a specific interest area left (runs) in a trial
IA_VISITED_TRIAL_%	The percentage of trials that have at least one fixation on a specific IA
IA_REVISIT_TRIAL_%	The percentage of trials that have one or more run of fixations on a specific IA

Note that if a token appeared multiple times in the same text or appeared in many texts, we averaged the values of all its occurrences to obtain a single feature vector for this token. We also used OSMAN to obtain the readability prediction of each word in the vocabulary (OSMAN Score) and to extract two linguistic features: the number of syllables (No_of_Syllables), and whether the word is a Faseeh word or not (Is_Faseeh). We entered each word with its diacritics, exactly as shown previously during the experiments.

4.7. Modeling

For model construction, we used the Waikato Environment for Knowledge Analysis (WEKA) [79], an open-source ML tool developed at the University of Waikato, Hamilton, New Zealand. After normalizing and standardizing all the linguistics and the selected eye-tracking features, we built models using different ML regression algorithms: linear regression (LR), implementation of the support vector machine (SVM) for regression (SMOreg), multi-layer perceptron (MLP), and two types of decision trees: M5P tree and reduced error pruning tree (REPTree). We used each algorithm to build two models in two rounds. We used all the 186-word types in our vocabulary in both rounds but with different features in the prediction analysis. In the first round, we used the features extracted from OSMAN by having the linguistic features No_of_Syllables and Is_Faseeh as predictor variables and the OSMAN score as the predicted value. In the second round, for each word, we used the same features as we did in the first round but combined them with the extracted eye movement measures. Similarly, the OSMAN score was the predicted value.

5. Results and Discussion

This section discusses the results of using eye-tracking measures to predict readability at the word and text levels.

5.1. Readability Prediction at the Word Level

To answer our research questions, in both rounds, we compared the performance of the regression models with different performance measures: correlation coefficient (CC) between the predicted and the actual readability score. The actual readability score is the score obtained from OSMAN, and thus it is based on the linguistic features used by OSMAN only without considering the eye-tracking data. The predicted score is the score obtained from the regression model. Other performance measures are mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and root relative squared error (RRSE). The results are shown in Table 4.

Table 4. The results of using different regression techniques with and without eye-tracking data.

Reg. Model	Using Gaze Data	CC	MAE	RMSE	RAE (%)	RRSE (%)
LR	x	0.96	11.98	13.83	28.86	26.56
	✓	0.98	9.14	11.50	21.99	22.05
SMO-reg	x	0.96	11.47	14.50	27.62	27.86
	✓	0.98	8.88	11.67	21.36	22.38
MLP	x	0.93	15.09	19.48	36.33	37.43
	✓	0.95	12.43	15.87	29.90	30.42
M5P Tree	x	0.96	12.44	14.70	29.97	28.23
	✓	0.97	9.78	12.73	23.54	24.40
REP Tree	x	0.95	12.19	16.76	29.31	32.13
	✓	0.91	16.51	21.69	39.83	41.68

As shown in Table 4, using eye-tracking data in the prediction task improved the correlation coefficient between the predicted and the actual readability scores and decreased all error types. This observation is evident for non-tree-based models, LR, SMO-reg, and MLP, compared to the M5P Tree model. The improvement in the correlation coefficient and the decrease of all error types in the readability prediction task implied that there was a relationship between using eye movement measures and the ability of the used algorithms to model words' difficulty level more accurately. However, in the case of the REPTree, we can observe that it did not perform similarly to the other models, as its performance decreased when using gaze data.

5.1.1. Features' Contribution

Regarding our second question about the features that contribute the most to Arabic readability prediction, we used the Correlation Ranking Filter in WEKA. We ranked the features in the first round based on their correlation with the predicted value (OSMAN score) as shown in Figure 4. (Is_Faseeh) was shown to have more correlation with the predicted OSMAN score than the number of syllables (No_of_Syllables). In the second round, it was found that some eye features, which are IA_REVISIT_TRIAL_%, IA_FIRST_FIXATION_DURATION, and IA_RUN_COUNT were ranked as the three top correlated attributes, followed by Is_Faseeh as a linguistic feature.

Based on this, we can assume that participants tend to fixate on a word and revisit previous parts of the text when facing difficult words. The rest of the eye measures (IA_DWELL_TIME, IA_DWELL_TIME_%, IA_FIXATION_COUNT, IA_FIX_COUNT_%, and IA_VISITED_TRIAL_%) reflect the late processing stage of readers, which justify their ranking below (IA_REVISIT_TRIAL_%, IA_FIRST_FIXATION_DURATION, and IA_RUN_COUNT). We also noticed that although used vocabularies affect readability significantly [80], No_of_Syllables was always the lowest-ranked attribute in both rounds, indicating that it does not affect the predicted score. This might be because, as mentioned in [81], "an Arabic word could contain more than three syllables and still be considered a non-complex word" [29].

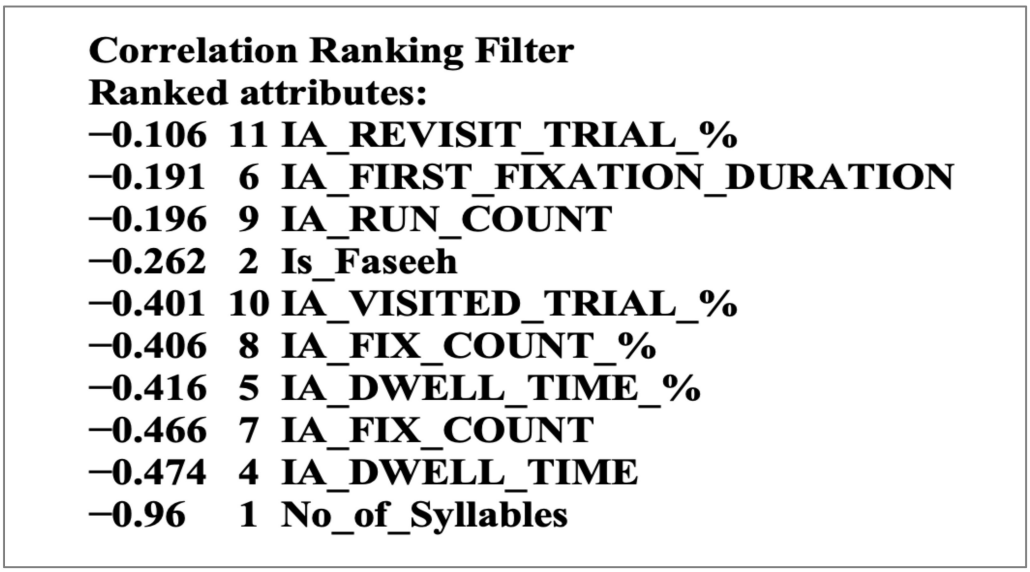

Figure 4. Correlation ranking attribute evaluation (with eye data).

These results indicate that eye-tracking could be a valuable technique for monitoring the cognitive activities associated with reading and explaining the correlation between cognitive processes and eye movement.

5.1.2. Model Comparison

Finally, to answer our last question, we compared model performance using WEKA by setting linear regression as our base model. The results are illustrated in Tables 5 and 6, with and without using eye-tracking data, respectively.

Table 5. Model comparison without using eye-tracking data. The asterisk (*) indicates statistically significantly worse and (v) indicates statistically significantly better.

Performance Measure	LR	MLP	SMOreg	M5P	REPTree
CC	0.20	0.94 v	0.96 v	0.85 v	0.00 *
MAE	40.69	17.69 *	14.29 *	38.45 *	41.55
RMSE	50.15	21.87 *	15.75 *	50.39	51.37
RAE	97.94	43.37 *	34.97 *	92.20 *	100.00
RRSE	97.62	43.18 *	30.96 *	97.82	100.00

As can be seen from Table 5, when we used only linguistic features, MLP, SMOreg, and M5P showed statistically significant improvement over linear regression in their correlation coefficients between the predicted values and the actual values. In contrast, REPTree had a significantly worse correlation coefficient than linear regression. MLP and SMOreg showed a significant decrease for different types of error, and SMOreg had better performance (lower errors) than did MLP.

Table 6. Model comparison when using eye-tracking data. The asterisk (*) indicates statistically significantly worse and (v) indicates statistically significantly better.

Performance Measure	LR	MLP	SMOreg	M5P	REPTree
CC	0.14	0.76 v	0.93 v	−0.02	0.00
MAE	12.43	8.00 *	4.90 *	12.68 v	12.58
RMSE	15.43	9.94 *	6.25 *	15.60	15.63
RAE	98.76	64.20 *	39.28 *	100.92 v	100.00
RRSE	98.76	63.86 *	40.51 *	100.05	100.00

For decision trees, M5P showed significant improvement in the mean absolute error and the relative absolute error only, and REPTree did not show any significant difference compared to linear regression. Thus, the best-performing model using linguistic features only was SMOreg, followed by MLP.

We can see from Table 6 that when we used both linguistics and eye-tracking features, MLP and SMOreg showed a statistically significant improvement over linear regression in their correlation coefficients. In contrast, neither decision tree showed a significant difference compared to linear regression. For different types of error, similar to Table 5, MLP and SMOreg showed a significant decrease, and SMOreg had better performance (lower errors) than did MLP. For decision trees, M5P showed a significant increase in the mean absolute error and the relative absolute error, and REPTree did not show any significant difference compared to linear regression. Thus, similar to the previous results, the best-performing model in our dataset and the features used was SMOreg, followed by MLP.

5.2. Readability Prediction at the Text Level

In addition to extracted eye-tracking measures at the word level, we extracted several eye-tracking measures aggregated at the text level, as shown in Table 7.

Table 7. Used eye-tracking features for each text.

Eye-Tracking Feature	Description
FIXATION_COUNT (FC)	The average number of fixations in a text.
FIXATION_DURATION_MEAN (FDM)	The average duration of fixations (in milliseconds) a participant makes in a text.
SACCADE_AMPLITUDE_MEAN (SAM)	Average size (in degrees of visual angle) of all saccades in the trial group. "This is calculated by summing up the amplitude of all saccades in the trial group divided by the total number of saccades in the group" [78].
SACCADE_COUNT (SC)	The average number of saccades in a text.
RUN_COUNT (RC)	The average count of how many times a participant makes runs of fixations in a text.

In addition, we took the common subjective readabilities from all the participants for all texts. Table 8 shows all extracted eye-tracking measures and common labels assigned by participants.

Table 8. Texts' aggregated eye-tracking measures and common subjective readabilities assigned by participants.

Texts	FC	FDM	SAM	SC	RC	Participants' Common Subjective Readability
Text 1	80.9	276.01	1.79	81.15	41.88	E
Text 2	89.39	293.92	1.65	89.54	49.59	M/E
Text 3	86.98	280.19	1.75	87.22	49.68	E
Text 4	102.63	290.63	1.53	102.88	53.56	M/D
Text 5	108.15	282.38	1.74	108.37	54.1	M/D
Text 6	100.02	287.06	1.66	100.24	53.12	M/D

Using the eye-tracking measures in Table 8 and taking their average as shown in Table 9, we tried to observe how collected readers' behaviors using their eye movements were correlated with their subjective readabilities and with the OSMAN scores assigned in Table 1.

Table 9. Average text features.

Texts	FC	FDM	SAM	SC	RC	Participants' Common Subjective Readability
Texts 1–3	85.76	283.37	1.73	85.97	47.05	E
Texts 4–6	103.6	286.69	1.64	103.83	53.59	M/D

As we can see from Table 9, the results validate our initial assumption about the readability level of the labeled texts based on the era and assigned OSMAN scores in Table 1 (thus, poems written in the Andalusian era are easier to read than poems written in the pre-Islamic era). There is a correlation between these scores, the texts' aggregated eye features, and the participants' feedback about these texts.

It is clear that when comparing the two text categories, texts 1 to 3 had a lower fixation count, shorter fixation duration, larger amplitude of saccades (caused by the simplification of the texts [40]), lower saccade count, and lower run count, which indicated fewer regressions compared to texts 4 to 6. All these findings are well-matched with previous findings by Rayner [74], reported by [8], which conclude that readers' comprehension difficulties could be reflected by longer fixation duration, shorter saccades, and more revisits of parts in the text (regressions).

Although in readability assessment research, text readability is done by having annotated data and then training an ML model to predict the readability of the document or sentence, in our approach, we did the opposite by starting with unlabeled texts, assessing the readability at the word level by collecting eye movement data and then labeling texts. Thus, we can say that texts 1 to 3 are easy to read, and texts 4 to 6 are medium near to difficult, or more generally more difficult than other texts.

Because these texts are basically composed of the words that have been used at the word-level readability prediction and there was a correlation between the readability level of these texts (as measured by the OSMAN tool) and the eye-tracking measures obtained for those texts, we can infer that the generated readability scores at the word level are acceptable.

6. Conclusions and Future Work

This paper demonstrates an exploratory experiment in which we collected human reading data during the reading of Arabic texts, used eye-tracking features from the reading literature, and analyzed the results. The results show an improvement in the readability prediction task. The degree of this improvement varied based on the type of algorithm used, which indicates that a relationship exists between eye-tracking features and readability prediction. Further investigation of this observation is needed from different dimensions, such as the nature of the prediction algorithm and its associated parameters. Additionally, eye-movement measures were found to be superior to linguistic features when combined, which may indicate the ability of these features to reflect the readability level of the text in a more natural and precise way.

Our results do have some limitations, such as the insufficiency of data and the number of features used and the ML models used not being state-of-the-art. No generalization could be obtained about this relationship and its strength. However, based on the behaviors of different regression models and the error decrease, we infer a relationship. This study motivates us to explore this subarea of readability assessment more deeply and examine the significance of using gaze data in Arabic readability assessment.

For future work, because a very important success factor for readability studies is the availability of sufficiently large and graded gold-standard datasets, which is not the case for Arabic, we plan to build an improved Arabic human reading corpus similar to the widely used Dundee corpus [82]. Additionally, we plan to feature engineer eye-tracking data to understand their effects on readability assessment, particularly for Arabic. Offering an open-source Arabic human reading corpus will be an excellent contribution for Arabic

that will open the door for broader utilization of eye-tracking for readability assessment of many NLP tasks, such as text simplification, machine translation, text summarization, etc.

Additionally, regarding the ML model, NNs have shown significant advancement in NLP, where word embedding programs, such as word2vec and GloVe, have enabled the accurate modeling of the semantic relationships between words. Using these relationships in readability assessments has shown encouraging results [37,49]. Additionally, RNNs have been used widely in psycholinguistics and NLP research, particularly reading research. It will be worth using different RNN variations (simple recurrent networks (SRNs), long short-term memory networks (LSTMs), and GRUs) and observe how they could model human cognitive processing more accurately over traditional ML models such as SVM, naïve Bayes, and decision trees. Thus, it is worth investigating these neural methods for Arabic readability prediction, particularly with eye-tracking features. Another alternative is transformers that could be fine-tuned to target the readability prediction task.

Lastly, we aim to use different texts, such as news articles or textbooks, to observe how modeling human cognitive processing differs depending on the text used.

Author Contributions: Conceptualization, I.B., H.A.-K. and A.A.-S.; methodology, I.B., H.A.-K. and A.A.-S.; software, I.B.; validation, I.B. and H.A.-K.; formal analysis, I.B. and H.A.-K.; investigation, I.B., H.A.-K. and A.A.-S.; resources, I.B. and H.A.-K.; data curation, I.B.; writing—original draft preparation, I.B.; writing—review and editing, H.A.-K. and A.A.-S.; visualization, I.B.; supervision, H.A.-K. and A.A.-S.; project administration, I.B., H.A.-K. and A.A.-S.; funding acquisition, H.A.-K. and A.A.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deanship of Scientific Research in King Saud University through the initiative of DSR Graduate Students Research Support (GSR).

Institutional Review Board Statement: The study was conducted according to the Prince Sultan University's Institutional Review Board (IRB) rules and regulations and was approved by the Ethics Committee of Prince Sultan University (PSU IRB-2021-01-0067, 20 February 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study, including the experimental texts, participants' details, and collected eye-movement data for all tokens, are available on request from the corresponding author. The data are not publicly available due to ethical and privacy issues.

Acknowledgments: The authors would like to thank the Deanship of Scientific Research in King Saud University for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR). Additionally, the authors would like to thank the Applied Linguistics Research Lab at Prince Sultan University, Riyadh, Saudi Arabia, for facilitating eye-tracking data collection.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Balyan, R.; McCarthy, K.S.; McNamara, D.S. Comparing Machine Learning Classification Approaches for Predicting Expository Text Difficulty. In Proceedings of the Thirty-First International Flairs Conference, Melbourne, FL, USA, 21–23 May 2018; pp. 421–426.
2. Collins-Thompson, K. Computational assessment of text readability: A survey of current and future research. *ITL-Int. J. Appl. Linguist.* **2014**, *165*, 97–135. [CrossRef]
3. Dale, E.; Chall, J.S. The Concept of Readability. *Elem. Engl.* **1949**, *26*, 19–26.
4. Alotaibi, S.; Alyahya, M.; Al-Khalifa, H.; Alageel, S.; Abanmy, N. Readability of Arabic Medicine Information Leaflets: A Machine Learning Approach. *Procedia Comput. Sci.* **2016**, *82*, 122–126. [CrossRef]
5. Feng, L.; Elhadad, N.; Huenerfauth, M. Cognitively motivated features for readability assessment. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009; pp. 229–237.
6. Baazeem, I. Analysing the Effects of Latent Semantic Analysis Parameters on Plain Language Visualisation. Master's Thesis, Queensland University, Brisbane, QLD, Australia, 2015.
7. Mesgar, M.; Strube, M. Graph-based coherence modeling for assessing readability. In Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, Denver, CO, USA, 4–5 June 2015; pp. 309–318.

8. Vajjala, S.; Meurers, D.; Eitel, A.; Scheiter, K. 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), Osaka, Japan, 11 December 2016; pp. 38–48.
9. Vajjala, S.; Lucic, I. On understanding the relation between expert annotations of text readability and target reader comprehension. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics. Florence, Italy, 2 August 2019; pp. 349–359.
10. Mathias, S.; Kanojia, D.; Mishra, A.; Bhattacharya, P. A Survey on Using Gaze Behaviour for Natural Language Processing. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Survey Track, Yokohama, Japan, 11–17 July 2020; pp. 4907–4913.
11. Singh, A.D.; Mehta, P.; Husain, S.; Rajkumar, R. Quantifying sentence complexity based on eye-tracking measures. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), Osaka, Japan, 11 December 2016; pp. 202–212.
12. Copeland, L.; Gedeon, T.; Caldwell, S. Effects of text difficulty and readers on predicting reading comprehension from eye movements. In Proceedings of the 2015 6th IEEE International Conference on Cognitive Info communications (Cog. Info. Com.), Gyor, Hungary, 19–21 October 2015; pp. 407–412.
13. Just, M.A.; Carpenter, P.A. A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* **1980**, *87*, 329–354. [CrossRef]
14. Atvars, A. Eye movement analyses for obtaining Readability Formula for Latvian texts for primary school. *Procedia Comput. Sci.* **2017**, *104*, 477–484. [CrossRef]
15. Chen, Y.; Zhang, W.; Song, D.; Zhang, P.; Ren, Q.; Hou, Y. Inferring Document Readability by Integrating Text and Eye Movement Features. In Proceedings of the SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research, Santiago, Chile, 2 December 2015.
16. Garain, U.; Pandit, O.; Augereau, O.; Okoso, A.; Kise, K. Identification of reader specific difficult words by analyzing eye gaze and document content. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 1346–1351.
17. Mishra, A.; Bhattacharyya, P. Scanpath Complexity: Modeling Reading/Annotation Effort Using Gaze Information. In *Cognitively Inspired Natural Language Processing. Cognitive Intelligence and Robotics*; Robotics: Singapore, 2018; pp. 77–98.
18. Al Jarrah, E.Q.A. Using Language Features to Enhance Measuring the Readability of Arabic Text. Master's Thesis, Yarmouk University, Irbid, Jordan, 2017.
19. Leal, S.E.; Vieira, J.M.M.; Rodrigues, E.D.S.; Teixeira, E.N.; Aluísio, S. Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches. In Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics. Barcelona, Spain, 8–13 December 2020; pp. 5821–5831.
20. Gonzalez-Garduño, A.V.; Søgaard, A. Learning to predict readability using eye-movement data from natives and learners. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5118–5124.
21. Litsas, C.; Mastropavlou, M.; Symvonis, A. Text classification for children with dyslexia employing user modelling techniques. In Proceedings of the IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications, Chania, Greece, 7–9 July 2014; pp. 1–6.
22. Fouad, M.M.; Atyah, M.A. MLAR: Machine Learning based System for Measuring the Readability of Online Arabic News. *Int. J. Comput. Appl.* **2016**, *154*, 29–33. [CrossRef]
23. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221–233. [CrossRef]
24. Gunning, R. *The Technique of Clear Writing*; McGraw-Hill Book Company: New York, NY, USA, 1968.
25. Chall, J.S.; Dale, E. *Readability Revisited: The New Dale-Chall Readability Formula*; Brookline Books: Cambridge, MA, USA, 1995.
26. Laughlin, G.H.M. SMOG grading-a new readability formula. *J. Read.* **1969**, *12*, 639–646.
27. Coleman, M.; Liau, T.L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **1975**, *60*, 283–284. [CrossRef]
28. Kincaid, J.P.; Fishburne, R.P.; Rogers, R.L.; Chissom, B.S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*; University of Central Florida: Orlando, FL, USA, 1975.
29. El-Haj, M.; Rayson, P. OSMAN—A Novel Arabic Readability Metric. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 8–13 December 2020; pp. 250–255.
30. Cavalli-Sforza, V.; Saddiki, H.; Nassiri, N. Arabic Readability Research: Current State and Future Directions. *Procedia Comput. Sci.* **2018**, *142*, 38–49. [CrossRef]
31. Dawood, B. The Relationship between Readability and Selected Language Variables. Ph.D. Thesis, Baghdad University, Iraq, Baghdad, 1977.
32. Al-Heeti, K.N. Judgment analysis technique applied to readability prediction of Arabic reading material. Ph.D. Thesis, University of Northern Colorado, Greeley, CO, USA, 1985.
33. Daud, N.M.; Hassan, H.; Aziz, N.A. A corpus-based readability formula for estimate of Arabic texts reading difficulty. *World Appl. Sci. J.* **2013**, *21*, 168–173.
34. Al Tamimi, A.K.A.; Jaradat, M.; Al-Jarrah, N.; Ghanem, S. AARI: Automatic Arabic readability index. *Int. Arab J. Inf. Technol.* **2014**, *11*, 370–378.

35. Ghani, K.A.; Noh, A.S.; Yusoff, N.M.R.N.; Hussein, N.H. Developing Readability Computational Formula for Arabic Reading Materials Among Non-native Students in Malaysia. *Importance New Technol.* **2021**, *194*, 2041–2057. [CrossRef]
36. Chen, X.; Meurers, D. Characterizing text difficulty with word frequencies. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, San Diego, CA, USA, 16 June 2016; pp. 84–94.
37. Mesgar, M.; Strube, M. A neural local coherence model for text quality assessment. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4328–4339.
38. Al-Edaily, A.; Al-Wabil, A.; Al-Ohali, Y. Interactive Screening for Learning Difficulties: Analyzing Visual Patterns of Reading Arabic Scripts with Eye Tracking. In *HCI 2013: HCI International 2013—Posters' Extended Abstracts*; Stephanidis, C., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 3–7.
39. Conklin, K.; Pellicer-Sanchez, A. Using eye-tracking in applied linguistics and second language research. *Second. Lang. Res.* **2016**, *32*, 453–467. [CrossRef]
40. Grabar, N.; Farce, E.; Sparrow, L. Study of readability of health documents with eye-tracking approaches. In Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA), Tilburg, The Netherlands, 8 November 2018; pp. 10–20.
41. Gonzalez-Garduno, A.V.; Søgaard, A. Using gaze to predict text readability. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Copenhagen, Denmark, 8 September 2017; pp. 438–443.
42. Al-Ajlan, A.A.; Al-Khalifa, H.S.; Al-Salman, A.S. Towards the development of an automatic readability measurements for Arabic language. In Proceedings of the 2008 Third International Conference on Digital Information Management, London, UK, 13–16 November 2008; pp. 506–511.
43. Forsyth, J.N. Automatic Readability Prediction for Modern Standard Arabic. Ph.D. Thesis, Brigham Young University, Provo, UT, USA, 2014.
44. Rello, L. DysWebxia: A Text Accessibility Model for People with Dyslexia. Ph.D. Thesis, Pompeu Fabra University, Barcelona, Spain, 2014.
45. Azpiazu, I.M.; Pera, M.S. Multiattentive Recurrent Neural Network Architecture for Multilingual Readability Assessment. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 421–436. [CrossRef]
46. Martinc, M.; Pollak, S.; Robnik-Šikonja, M. Supervised and unsupervised neural approaches to text readability. *Comput. Linguist.* **2021**, *47*, 141–179. [CrossRef]
47. De Oliveira, A.M.; Germano, G.; Capellini, S.A. Comparison of Reading Performance in Students with Developmental Dyslexia by Sex. *Paidéia* **2017**, *27*, 306–313. [CrossRef]
48. Crossley, S.A.; Allen, L.K.; Snow, E.L.; McNamara, D.S. Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *J. Educ. Data Min.* **2016**, *8*, 1–19.
49. Mesgar, M.; Strube, M. Lexical coherence graph modeling using word embeddings. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1414–1423.
50. Xu, P.; Saghir, H.; Kang, J.S.; Long, T.; Bose, A.J.; Cao, Y.; Cheung, J.C.K. A cross-domain transferable neural coherence model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July 2019; pp. 678–687.
51. Logeswaran, L.; Lee, H.; Radev, D. Sentence Ordering and Coherence Modeling using Recurrent Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5285–5292.
52. Zhang, M.; Feng, V.W.; Qin, B.; Hirst, G.; Liu, T.; Huang, J. Encoding world knowledge in the evaluation of local coherence. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 1087–1096.
53. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
54. Li, J.; Hovy, E. A model of coherence based on distributed sentence representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 2039–2048.
55. Marie-Sainte, S.L.; Alalyani, N.; Alotaibi, S.; Ghouzali, S.; Abunadi, I. Arabic Natural Language Processing and Machine Learning-Based Systems. *IEEE Access* **2018**, *7*, 7011–7020. [CrossRef]
56. Shen, W.; Williams, J.; Marius, T.; Salesky, E. A language-independent approach to automatic text difficulty assessment for second-language learners. In Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, Sofia, Bulgaria, 8 August 2013; pp. 30–38.
57. Nassiri, N.; Lakhouaja, A.; Cavalli-Sforza, V. Arabic L2 readability assessment: Dimensionality reduction study. *J. King Saud. Univ. Comput. Inf. Sci.* **2021**. [CrossRef]
58. Saddiki, H.; Cavalli-Sforza, V.; Bouzoubaa, K. Enhancing Visualization in Readability Reports for Arabic Texts. *Procedia Comput. Sci.* **2017**, *117*, 241–247. [CrossRef]
59. Khallaf, N.; Sharoff, S. Automatic difficulty classification of Arabic sentences. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP), Virtual. Kyiv, Ukraine, 19 April 2021; pp. 105–114.
60. Sood, E.; Tannert, S.; Frassinelli, D.; Bulling, A.; Vu, N.T. Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension. In Proceedings of the 24th Conference on Computational Natural Language Learning, Virtual. 19–20 November 2020; pp. 12–25.

61. Frazier, L.; Rayner, K. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* **1982**, *14*, 178–210. [CrossRef]
62. Clifton, C.; Staub, A.; Rayner, K. Eye movements in reading words and sentences. In *Eye Movements*; Gompel, R.P.G.V., Fischer, M.H., Murray, W.S., Hill, R.L., Eds.; Elsevier: Oxford, UK, 2007; pp. 341–371.
63. Rayner, K.; Chace, K.H.; Slattery, T.; Ashby, J. Eye Movements as Reflections of Comprehension Processes in Reading. *Sci. Stud. Read.* **2006**, *10*, 241–255. [CrossRef]
64. Liversedge, S.P.; Paterson, K.B.; Pickering, M.J. Eye movements and measures of reading time. In *Eye Guidance in Reading and Scene Perception*; Underwood, G., Ed.; Elsevier: Amsterdam, The Netherlands, 1998; pp. 55–75.
65. Schroeder, S.; Hyönä, J.; Liversedge, S.P. Developmental eye-tracking research in reading: Introduction to the special issue. *J. Cogn. Psychol.* **2015**, *27*, 500–510. [CrossRef]
66. Raney, G.E.; Campbell, S.J.; Bovee, J.C. Using Eye Movements to Evaluate the Cognitive Processes Involved in Text Comprehension. *J. Vis. Exp.* **2014**, *83*, e50780. [CrossRef] [PubMed]
67. Sinha, A.; Roy, D.; Chaki, R.; De, B.K.; Saha, S.K. Readability Analysis Based on Cognitive Assessment Using Physiological Sensing. *IEEE Sens. J.* **2019**, *19*, 8127–8135. [CrossRef]
68. Zubov, V.I.; Petrova, T.E. Lexically or grammatically adapted texts: What is easier to process for secondary school children? *Procedia Comput. Sci.* **2020**, *176*, 2117–2124. [CrossRef]
69. Merkx, D.; Frank, S.L. Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv* **2020**, arXiv:2005.09471.
70. Wilcox, E.; Gauthier, J.; Hu, J.; Qian, P.; Levy, R. On the predictive power of neural language models for human real-time comprehension behavior. In Proceedings of the 42nd Annual Meeting of the Cognitive Science Society, Virtual. 29 July–1 August 2020; pp. 1707–1713.
71. Goodkind, A.; Bicknell, K. Predictive power of word surprisal for reading times is a linear function of language model quality. In Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), Salt Lake City, UT, USA, 7 January 2018; pp. 10–18.
72. Aurnhammer, C.; Frank, S.L. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In Proceedings of the 41st Annual Conference of the Cognitive Science Society (CogSci 2019), Montreal, QC, Canada, 24–27 July 2019; pp. 112–118.
73. Clifton, C.; Staub, A. Syntactic influences on eye movements during reading. *Oxf. Handb. Online* **2011**, *3*, 895–910. [CrossRef]
74. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372–422. [CrossRef] [PubMed]
75. Adab: The World Encyclopedia of Arabic Literature. Available online: https://www.adab.com (accessed on 1 February 2021).
76. Bensoltana, D.; Asselah, B. Exploration of Arabic reading, in terms of the vocalization of the text form by registering the eyes movements of pupils. *World J. Neurosci.* **2013**, *3*, 263–268. [CrossRef]
77. S. R. Ltd. SR Research EyeLink. Available online: https://www.sr-research.com (accessed on 20 March 2021).
78. S. R. Ltd. EyeLink Data Viewer User's Manual. Available online: http://sr-research.jp/support/files/dvmanual.pdf (accessed on 20 April 2021).
79. WEKA. The Workbench for Machine Learning. Available online: https://www.cs.waikato.ac.nz/mL/weka/ (accessed on 10 May 2021).
80. Cavalli-Sforza, V.; Mezouar, M.E.; Saddiki, H. Matching an Arabic text to a learners' curriculum. In Proceedings of the 2014 Fifth International Conference on Arabic Language Processing (CITALA 2014), Oujda, Morocco, 26–27 November 2014; pp. 79–88.
81. Al-Khalifa, H.S.; Al-Ajlan, A.A. Automatic readability measurements of the Arabic text: An exploratory study. *Arab. J. Sci. Eng.* **2010**, *35*, 103–124.
82. Barrett, M.; Agic, Ž.; Søgaard, A. The dundee treebank. In Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT14), Warsaw, Poland, 11–12 December 2015; pp. 242–248.