Running head: Predicting Processing Effort

Full Title: Predicting Processing Effort During L1 and L2 Reading: The Relationship Between Text Linguistic Features and Eye Movements

Author: Shingo Nahatame (University of Tsukuba)

Acknowledgements: This research was supported by JSPS KAKENHI Grant Number 20K00827.

Address for correspondence: School of Education, Institute of Human Sciences, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305–8577, Japan. Email: nahatame.shingo.gp@u.tsukuba.ac.jp

Abstract

Researchers have taken great interest in the assessment of text readability. This study expands on this research by developing readability models that predict the processing effort involved during first language (L1) and second language (L2) text reading. Employing natural language processing tools, the study focused on assessing complex linguistic features of texts, and these features were used to explain the variance in processing effort, as evidenced by eye movement data for L1 or L2 readers of English that were extracted from an open eye-tracking corpus. Results indicated that regression models using the indices of complex linguistic features provided better performance in predicting processing effort for both L1 and L2 reading than the models using simple linguistic features (word and sentence length). Furthermore, many of the predictive variables were lexical features for both L1 and L2 reading, emphasizing the importance of decoding for fluent reading regardless of the language used.

Keywords: reading, eye tracking, readability, natural language processing, decoding

Introduction

Matching text difficulty to the abilities of language learners helps them better understand the text and improve their reading skills over time (McNamara, Graesser, McCarthy & Cai, 2014; Mesmer, 2008). Therefore, assessing how easily written texts can be read and understood has been an important issue among researchers, educators, and publishers, leading to the development of a large number of formulas to assess text readability (particularly in the English language). Although most readability formulas were developed for first language (L1) users, some of them have also been used for second language (L2) users (Greenfield, 1999, 2004).

However, researchers have noted that traditional readability formulas only consider surface linguistic features such as word and sentence length, resulting in weak construct and theoretical validity (e.g., Bertram & Newman, 1981; Crossley, Greenfield & McNamara, 2008). In addition, previous readability studies are limited in that most have focused solely on comprehension and have not considered processing, which is an important aspect of readability (Crossley, Skalicky & Dascalu, 2019). Moreover, they have been conducted in either the L1 or L2 reading context, which makes it difficult to deepen our understanding of the nature of bilingual reading.

To address these issues, the current study aims to build readability models that predict how much cognitive effort is required to process text (i.e., processing effort) in L1 and L2 reading, respectively. This study was based on two previous studies: Crossley et al. (2019) and Nahatame (2021). Specifically, it assessed an extensive range of text linguistic features by employing advanced natural language processing (NLP) tools (as in Crossley et al., 2019), which were then used to explain the variance in the processing effort as evidenced by eye movements (as in Nahatame, 2021) during L1 and L2 reading. The findings will not only offer insights into the readability assessment of L1 and L2 texts but also shed light on

the common or different linguistic features that are closely related to the cognitive processes of L1 and L2 reading.

Background

The background of the current study is two-fold: (a) the theoretical account for the cognitive processes of reading in relation to text linguistic features and (b) a review of previous studies on text readability that motivated the current study.

Cognitive processes of reading and text linguistic features

Multiple cognitive processes during reading

Reading comprises multiple cognitive processes, including word recognition, syntactic parsing, meaning encoding in the form of propositions, text-model formation, situation-model building, and inferencing (Grabe, 2009). Lower-level processing, such as word recognition, has the potential to be strongly automatized; therefore, automatizing such processing is crucial for fluent reading. On the other hand, higher-level processing, such as inferencing, is more conscious and is required for building coherent text comprehension because it is often activated beyond the sentence level. Similarly, Graesser and McNamara (2011) identified the six different levels: words, syntax, the explicit textbase, the referential situation model, the discourse genre and rhetorical structure, and the pragmatic communication level (see also McNamara et al., 2014). They suggested the importance of considering these levels when assessing text difficulty, and this model has been the theoretical foundation for some text analysis tools mentioned later (see the Assessing text readability section).

Although the cognitive processes described above are common regardless of the languages used for reading (i.e., L1 or L2), it is well known that L2 reading is more

cognitively effortful than L1 reading (Dirix, Vander, De Bruyne, Brysbaert & Duyck, 2019). One of the theoretical accounts for this is the resource hypothesis, which suggests that L2 processing demands greater cognitive load on working memory (Sandoval, Gollan, Ferreira & Salmon, 2010). In L2 reading, cognitive resources must be heavily allocated for lower-level processing, and therefore, few remain for higher-level processing (Horiba, 1996, 2000; Morishima, 2013). Another instance is the weaker links hypothesis (Gollan, Montoya, Cera & Sandoval, 2008), which proposes that for bilinguals, the frequency of use of lexical items is divided between the two languages, resulting in L2 word representations that are as weak and less detailed as low-frequency L1 words. Eye-tracking studies provided empirical evidence that L2 reading is less efficient than L1 reading, showing longer reading times, shorter saccades, and less frequent word skipping for L2 reading than for L1 (Cop, Drieghe & Duyck, 2015; Dirix et al., 2019; Kuperman et al., 2022; Nisbet, Bertram, Erlinghagen, Pieczykolan & Kuperman, 2021).

However, it is worth noting that studies have also suggested the factors underlying the similarity of L1 and L2 reading behavior. Kuperman et al. (2022) found that the variance in L2 reading fluency, as assessed by eye movement data, is explained by L1 reading fluency to the greatest extent. Nisbet et al. (2021) statistically demonstrated that when L2 English readers had reading component skill levels similar to those of L1 readers (i.e., spelling skills, vocabulary size, and exposure to print) and their L1 was linguistically close to English, they showed similar eye-tracking reading times as the L1 English readers.

Word recognition

Because the current study considered the first four of the six levels proposed by Graesser and McNamara's (2011) model in the text analysis (see the Method section for the rationale for this decision), this and the following sections provide an in-depth review of the cognitive processes activated at these levels (i.e., word, sentence, textbase, and situation

model) in relation to text linguistic features.

The most fundamental process of reading is word recognition (or, more specifically, decoding). Efficient word recognition is essential for fluent and successful reading comprehension (Jeon & Yamashita, 2014; Koda, 2005; Yamashita, 2013). It is well attested that the frequency of the occurrence of a word in a language is a robust predictor of word recognition; for instance, highly frequent words are fixated for a shorter period of time than less frequent words in both L1 and L2 reading (Cop, Keuleers, Drieghe & Duyck, 2015; Whitford & Titone, 2017). This is in line with E-Z reader, an influential model of eye-movement control (Reichle et al., 1999). The model assumes the large effects of linguistic features such as word frequency, length, and predictability on the early stage of lexical processing.

Previous studies have also investigated the effects of word properties beyond frequency, including concreteness, familiarity, age of acquisition (AoA), word neighborhood, word association, and contextual distinctiveness. In general, research has indicated that words are processed more quickly when they are more familiar, more concrete, and acquired earlier (Chaffin et al., 2001; Dirix & Duyck, 2017; Juhasz & Rayner, 2003), when they have more phonological and orthographic neighbors (Yarkoni, Balota & Yap, 2008; Yap & Balota, 2009), and when they are associated with more words (Buchanan, Westbury & Burgess, 2001) and used with a broader range of context words (McDonald & Shillcock, 2001).

Several of these factors have also been found to influence L2 word reading. Kim, Crossley, and Skalicky (2018) showed that L2 words are read faster when they are more frequent, less concrete, and orthographically indistinct. Other studies have indicated that word frequency (Cop, Keuleers et al., 2015) and L1/L2 AoA (Dirix & Duyck, 2017) influence eye movement patterns during L2 word reading. In addition, these features (i.e., orthographic distinctiveness, frequency, and AoA) indicated greater effects for L2 and lower-

proficiency readers than for L1 and higher-proficiency readers.

In addition to single-word processing, the processing of multi-word units has received a fair amount of attention in recent studies. Research has provided converging evidence that frequent sequences of multi words (i.e., formulaic language) have a processing advantage over less frequent sequences of words (i.e., non-formulaic language) for L1 speakers (e.g., Tabossi, Fanari & Wolf, 2009; Siyanova-Chanturia, Conklin & Schmitt, 2011). Evidence also indicates the effects of phrasal frequency on the processing of formulaic sequences (Li, Warrington, Pagán, Paterson & Wang, 2021; Sonbul, 2015). Although the findings are mixed for L2 speakers, some studies suggest that learners with higher L2 proficiency are more likely to enjoy the advantage provided by formulaic language (e.g., Conklin & Schmitt, 2008; Underwood, Schmitt & Galpin, 2004).

Syntactic parsing

Syntactic parsing also plays a role in successful reading comprehension, such that syntactic structures and complexity have an impact on text processing. For instance, object-relative center-embedded sentences are found to induce longer reading times and more regressive eye movements than subject-relative center-embedded sentences (Holmes & O'Regan, 1981).

Syntactic processing patterns can differ between L1 and L2 speakers, which is known as the shallow structure hypothesis (Clahsen & Felser, 2006). The hypothesis posits that L1 speakers make use of both structural processing (computation of syntactic structure) and shallow processing (dependence on pragmatic and lexical information), whereas L2 learners tend to rely more on the latter. As a result, L2 readers can be less sensitive to syntactic structures than L1 readers during text processing (Marinis, Roberts, Felser & Clahsen, 2005).

Discourse processing

While word recognition and syntactic parsing are conducted, information extracted from the words and sentences constructs meaning units called semantic propositions (Grabe, 2009). These units are connected to each other based on their semantic overlap (Kintsch, 1998), which helps readers construct larger patterns of meaning for discourse comprehension (Nahatame, 2018, 2020) such as the explicit textbase and referential situation model (McNamara et al., 2014).

Th coherence in discourse comprehension can be facilitated by text cohesion. This indicates the presence of text elements that make the relationship between sentences or ideas explicit. If the text was rewritten to be more cohesive by replacing ambiguous pronouns with nouns, adding connectives, and rearranging the order of text information, it generally improves students' text comprehension (Britton & Gülgöz, 1991; Ozuru, Dempsey & McNamara, 2009). Although discourse processing is often limited in L2 reading (Horiba, 1996, 2000; Morishima, 2013), some studies have indicated that cohesion features such as the use of connectives and sentence relatedness affect eye movements during L2 reading (e.g., Zufferey, Mak, Degand & Sanders, 2015; Nahatame, 2022).

In summary, empirical evidence has indicated that various text linguistic features affect multiple cognitive processing during L1 and L2 reading (as often assessed by eye movements). However, most studies have focused on individual linguistic features (e.g., word frequency); the few that focused on multiple linguistic features only examined features at one level (e.g., lexical features), and not at multiple levels (e.g., lexical, syntactic, and cohesion features). It is possible that linguistic features at a particular level control the effect of other levels and that some levels are more influential than others. Given this, it is worth examining the effects of text linguistic features at multiple levels on bilingual reading processing within a single study.

Assessing text readability

According to Richards and Schmidt (2013), text readability is defined as "how easily written materials can be read and understood" (p. 482). Dale and Chall (1949), a classic readability study, include in their definition "the extent to which readers understand the text, read it at an optimal speed, and find it interesting" (p. 23). Importantly, these definitions encompass a notion of text processing (i.e., how quickly and comfortably a text can be read) in addition to comprehension.

Researchers have proposed numerous formulas to assess English text readability, most of which are traditional formulas based on simple linguistic features (e.g., Dale & Chall, 1949; Flesch, 1948; Kincaid, Fishburne, Rogers & Chissom, 1975). Others are newer formulas or models that employ advanced computational techniques to assess more complex linguistic features (e.g., Crossley et al., 2008; Crossley et al., 2019; De Clercq & Hoste, 2016; Feng, Jansche, Huenerfauth & Elhadad, 2010; Pitler & Nenkova, 2008).

Traditional methods for assessing readability formulas

Among the traditional formulas, the most widely adopted measures are the Flesch formulas: Flesch Reading Ease (Flesch, 1948) and Flesch-Kincaid Grade Level (Kincaid et al., 1975). These formulas rely on two simple linguistic features: word length (the number of syllables) and sentence length (the number of words). Research has reported high correlations between text difficulty as assessed by these formulas and comprehension as assessed by reading tests (typically cloze tests) for both L1 (Fry, 1989; Hamsik, 1984) and L2 reading (Greenfield, 1999; cf. Brown, 1999).[1]

For simple algorithms, these traditional formulas have been commonly used for reading materials and standardized reading tests. However, they have also received criticism for their weak construct and theoretical validity (e.g., Bertram & Newman, 1981; Crossley et al., 2008). The traditional formulas rely on only two levels of linguistic features (i.e., word length and sentence length), and these features are only weak proxies of word recognition and

syntactic parsing. In addition, traditional formulas do not consider text cohesion, which plays a role in discourse processing. To address these issues, recent studies have attempted to develop new formulas that are more theoretically valid and perform better than traditional ones by employing more sophisticated and innovative linguistic techniques.

Using complex linguistic features for readability assessment

Studies in computer science have aimed to improve readability prediction by assessing more complex, fine-grained linguistic features using NLP techniques. These studies showed that more complex features, such as word probability, syntactic structures based on parse trees, and discourse elements based on semantic overlap, are helpful for the prediction of readability (e.g., De Clercq & Hoste, 2016; Feng et al., 2010; Piler & Nenkova, 2008; see also Collins-Thompson's [2014] review). Nevertheless, in many studies, the most reliable features remain lexical in nature (Pitler & Nenkova, 2008; Kate, Luo, Patwardhan, Franz, Florian, Mooney, Roukos & Welty, 2010).

In the field of cognitive science, a series of studies conducted by Crossley and his colleagues developed readability formulas that build on the theoretical account of the reading process and include indices of complex linguistic features obtained from several NLP tools. Crossley et al. (2008) proposed a readability formula for L2 reading, termed the *Coh-Metrix L2 Reading Index* (CML2RI), that relies on three indices obtained from the web-based NLP tool Coh-Metrix (McNamara et al., 2014). These indices are word frequency, syntactic similarity of the sentences, and word overlap between adjacent sentences. Several studies have found that CML2RI performs better than traditional formulas for predicting the cloze test scores of L2 readers (Crossley et al., 2008) and simplification levels of L2 reading texts (Crossley, Allen & McNamara, 2011). Other studies also found that (more) simplified L2 texts were characterized by more frequent and familiar words, less complex syntactic structures, and higher levels of lexical and semantic overlap between sentences (e.g.,

Crossley, Greenfield & McNamara, 2012).

In more recent studies, Crossley and his colleagues adopted a crowdsourcing approach to develop readability models using a large dataset (Crossley, Heintz, Choi, Batchelor, Karimi, & Malatinszky, 2022; Crossley, Skalicky, Dascalu, McNamara & Kyle, 2017; Crossley et al., 2019). They recruited online participants (L1 English users) through a website and asked them to judge which of the two texts was easier to understand (comprehension difficulty) and could be read more quickly (reading speed). In Crossley et al.'s (2019) study, the linguistic features of the texts were obtained by employing five advanced freely available NLP tools: the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015), the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle & Crossley, 2018), the Tool for the Automatic Analysis of Text Cohesion (TAACO; Crossley, Kyle & McNamara, 2016), the Sentiment Analysis and Social Cognition Engine (SEANCE; Crossley, Kyle & McNamara, 2017), and ReaderBench (Dascalu, Dessus, Bianco, Trausan-Matu & Nardy, 2014). They were then used to develop regression models for predicting judgments of comprehension and reading speed. The comprehension model, the Crowdsourced Algorithm of Reading Comprehension (CAREC), contained 13 indices, including measures of lexical sophistication (e.g., age of acquisition, frequency, and imageability), n-gram features[2] (e.g., range and frequency of two-word sequences [*bigrams*] or three-word sequences [*trigrams*] in a corpus), cohesion (e.g., lexical overlap at the paragraph and sentence levels and lemma type count for content words), and sentiment (positive adjectives). On the other hand, the reading speed model, the Crowdsourced Algorithm of Reading Speed (CARES), contained nine indices, including measures of text structure (e.g., number of content lemmas and function words), lexical sophistication (e.g., word naming response times and word concreteness), and syntactic complexity (e.g., complex nominals per T-unit). They found that both

comprehension and reading speed models accounted for a significantly larger amount of variance in judgments than did traditional formulas. In addition, other studies (Crossley et al., 2017, 2022) reported that models of comprehension and processing were strongly informed by indices related to decoding (e.g., AoA, range, concreteness for words; Crossley et al., 2017, 2022).

Text readability and processing effort

As noted in the introduction, most readability studies adopted objective measures of comprehension as assessed by reading tests despite the fact that the definitions of readability include the notion of text processing (Dale & Chall, 1949; Richards & Schmidt, 2013). Empirical evidence suggests that processing does not necessarily lead to comprehension: the eye-tracking reading times of particular sentences did not correlate with the accuracy of comprehension questions on these sentences (Dirix et al., 2019; Yeari et al., 2015). Similarly, Kuperman et al. (2022) reported weak correlations at best for eye movement measures and comprehension accuracy in L2 reading. Given these findings, text readability deserves to be further investigated in terms of how easily text can be processed.

Although not so many, some studies have examined the relationship between text difficulty and processing. For instance, Rayner et al. (2006) demonstrated that readers' subjective ratings of overall text difficulty correlated with their eye movement measures, indicating more and longer fixations for more difficult text. Rets and Rogaten (2021) recorded the eye movements of adult L2 English users while reading authentic and simplified texts. Their analyses showed that the simplified texts induced longer initial processing (as indicated by first-pass fixation duration) and shorter later processing (as indicated by second-pass fixation duration) of the sentences than did the authentic texts, suggesting that text simplification induced more effective initial processing and, therefore, less necessity for rereading.

Although these studies provide insight into the relationship between processing effort and text difficulty, only a limited number of studies have examined processing in relation to text difficulty as assessed by readability formulas. One such study is that of Ardoin et al. (2005), which indicated a moderate relationship between readability as evaluated by traditional formulas and reading fluency as assessed by the number of words students read correctly. Another example is Crossley et al. (2019), as described above, which developed a readability model specifically designed for text processing and demonstrated higher correlations with processing than traditional readability formulas. However, the model was derived from pairwise judgments of reading speed. Thus, as noted in Crossley et al.'s (2019) study itself, "it is an open question as to how accurately these reflect effortful processing on the part of readers" and "it was impossible to assess how long readers spent on each text because the texts needed to be displayed next to each other for the comparisons" (p. 556).

To address this issue and explore more objective measures of text processing, Nahatame (2021) examined processing effort as reflected by eye movements during L2 reading. This study examined the performance of several readability formulas, including both traditional and new (e.g., the Flesch formulas, CML2RI, CAREC, and CARES), to predict global reading measures such as average fixation duration and saccade length. The results indicated that the new formulas performed better than the traditional ones in predicting some eye-movement measures. However, all of the formulas failed to significantly predict all eye-movement measures and to show consistent results in a pair of experiments, suggesting "the difficulty of providing strong and consistent performances for predicting processing effort using existing readability formulas" (p. 33). In addition, the use of a holistic scale of readability in this study did not allow for examination of what specific linguistic features included in the formulas (e.g., lexical, syntactic, and cohesion features) contributed to the prediction of processing effort.

The current study

The current study builds on the two previous studies described above, Crossley et al. (2019) and Nahatame (2021), and aims to develop models that predict the processing effort involved during L1 and L2 text reading. This study analyzed eye movement measures as an indication of processing effort during reading (Nahatame, 2021) and used complex linguistic features as assessed by the NLP tools (Crossley et al., 2019) to explain the variance in the processing effort. The study developed the models for L1 and L2 reading, respectively, and then compared their performance with the models including only simple linguistic features, as used in traditional readability formulas (i.e., word and sentence length). The following two research questions (RQs), each of which consists of two sub-questions, were addressed:

RQ1

(a) Do models with complex linguistic features perform better than models based on simple linguistic features in predicting processing effort during reading?

(b) Does the performance differ between L1 and L2 reading?

RQ2

(a) Which of the lexical, syntactic, and cohesion features are predictive of processing effort during reading?

(b) Do these features differ between L1 and L2 reading?

In regard to RQ1, it was hypothesized that using complex linguistic features would allow for better prediction of processing effort because they are more closely associated with theories of reading than are simple linguistic features (Crossley et al., 2008, 2011, 2019). However, it is difficult to articulate predictions for the difference between L1 and L2 reading due to very limited research on this topic. Nevertheless, given the larger effects of some

complex lexical features for L2 or lower-proficiency readers' text processing (Cop et al., 2015; Dirix & Duyck, 2017; Kim et al., 2018), complex features may be more useful for predicting processing effort in L2 reading than in L1.

In regard to RQ2, given that eye movements are strongly influenced by word length and frequency (Cop et al., 2015; Reichle et al., 1999) and lexical features are useful for estimating readability (Crossley et al., 2017; Crossley et al., 2022; Pitler & Nenkova, 2008), it is reasonable to assume that lexical features are more predictive than other features. As for the difference between L1 and L2, L2 reading models may depend on more lexical indices in comparison to L1 models because L2 reading requires more cognitive resources for lower-level processing than L1 reading (Horiba, 1996, 2000; Morishima, 2013).

In contrast to Nahatame (2021), who included the existing readability indices as a predictor in the model, the current study is more exploratory in that it aimed to develop the readability models by selecting the predictors from a range of complex linguistic features. This approach allows us to better understand the links between individual text linguistic features and their roles in predicting processing effort.

In addition, the current study used eye movement measures as a dependent variable because eye tracking "provides a 'direct' measure of processing effort" (Conklin et al., 2018, p. 6) rather than subjective ratings of reading speed as used by Crossley et al. (2019). Eye tracking also allows assessment of processing effort in more natural reading situations (see the review by Godfroid, 2019; Conklin et al., 2018; Rets, 2021 for more details). Furthermore, using eye tracking advances existing readability research by providing "a promising new source of cues about text difficulty that could be integrated as features in prediction settings, especially in real time" and helping with "estimating individual cognitive difficulty or ease" (Collins-Thompson, 2014, p. 127).

Method

Eye-tracking corpus and eye movement measures

The current study analyzed data from Cop et al.'s (2015) open eye-tracking corpus (available from https://figshare.com/articles/dataset/new_fileset/1482031). The corpus contained eye movement data collected from 14 English monolinguals and 19 unbalanced Dutch (L1)–English (L2) bilinguals reading an Agatha Christie novel. Bilingual participants read half of the novel in their L1 and the rest in their L2 (the order of the languages was counterbalanced). During each trial, the participants read the text that appeared on the computer screen, with a maximum of 145 words presented at a time, and their eye movements were recorded with a tower-mounted EyeLink 1000 (SR Research, Canada). Six eye-movement measures for each sentence were included in the corpus: total reading time, number of fixations, average fixation duration, average saccade length, skipping rate, and regression rate. After eliminating data for overly long sentences, data for 4,649 sentences per participant remained on average.

The current study analyzed the entire dataset for both L1 and L2 reading (cf. Nahatame, 2021). Because the current study explored reading of English texts, it analyzed the data of monolingual (L1 English) reading and bilingual L2 (L2 English) reading. Importantly, the proficiency test in Cop et al.'s (2015) study indicated that English proficiency was clearly lower for bilingual participants than monolinguals (see Cop et al. [2015] for more details).

Among several eye-movement measures available, the current study adopted the total reading time and number of fixations as an indication of processing effort. Nahatame (2021) adopted other eye-movement measures that are less likely to correlate with text length (e.g., average fixation duration, skipping rates, and regression rates), given that some of the readability measures are more sensitive to text length than others. The current study, however, aimed to develop models that directly predict how much time and effort readers spend on the processing of a particular text. Thus, eye-movement measures were selected that

directly reflect reading speed and processing time as dependent variables while including text length (i.e., word count) as a control variable in the models. Total reading time and the number of fixations will increase as text length increases; thus, the focus of the current study is whether and which linguistic features explain additional variance of these eye-movement measures beyond the effects of text length.

Although the original data consisted of sentence-level measures, in this study, the data needed to be reconstructed for larger text units to assess text readability. Thus, they were reaggregated for each trial to match the participants' experience of reading, resulting in the dataset consisting of 588 trials (i.e., texts) per participant for L1 reading and about half as many trials per participant for L2 reading. As shown in Table 1, L2 reading showed longer reading times and more fixations per trial than L1 reading (Cop et al., 2015).

<Insert Table 1 about here>

Total reading time and the number of fixations were highly correlated (r = .93 for L1 reading; r = .95 for L2 reading). To consolidate these measures, a principal component analysis was performed (Godfroid & Hui, 2020), indicating a high proportion of variance for the first components (.97 for L1 reading and .98 for L2 reading). Thus, the current study used the principal component scores for the first components without considering further components as dependent variables in the following analysis. For the statistical model development, these component scores were averaged over L1 or L2 participants for each trial. Each trial included 14 monolinguals (L1 readers) and nine or 10 bilinguals (L2 readers).

Text analysis

Texts used for the 588 trials were extracted from Cop et al.'s (2015) study, each of

which consists of 91.54 words on average (*SD* = 18.26). The texts were analyzed using the following three advanced NLP tools: TAALES 2.2 (Kyle, Crossley & Berger, 2018), TAASSC 1.3.8 (Kyle & Crossley, 2018), and TAACO 2.0.4 (Crossley, Kyle, & Dascalu, 2019). These tools were employed to assess lexical, syntactic, and cohesion features of texts, respectively. These linguistic features correspond to the first four of the six levels of Graesser and McNamara's (2011) model (i.e., word, syntax, textbase, and situation model). The fifth level, discourse genre and rhetorical structure, was not considered because the target texts in this study were extracted from a single novel, making the understanding of genres less important to complete the task. The sixth and highest level, the pragmatic communication level, is often beyond the scope of the computational tools used in cognitive science (McNamara et al., 2014).

Additionally, the Simple Natural Language Processing Tool (SiNLP; Crossley, Allen, Kyle & McNamara, 2014) was employed to compute simple linguistic features (i.e., word count, average word length, and average sentence length) for these texts. All of the tools used in this study are freely available from NLP for the Social Sciences at https://www.linguisticanalysistools.org/. Appendix S1 in the Supplementary Materials provides an overview of the tools used in this study along with the linguistic features, measures, and indices assessed by the tools.

Lexical features

The TAALES computes many indices of lexical sophistication of the text (Kyle et al., 2018). This study adopted indices for word frequency, range, L1 age of acquisition, psycholinguistic word information (e.g., familiarity and concreteness), word response norms (i.e., L1 English users' reaction times and accuracies of the lexical decision and naming tasks), word neighbor information (i.e., orthographic and phonological neighbors), word association (i.e., the number of other words associated with a word), contextual

distinctiveness (i.e., the diversity of contexts in which a word is encountered based on statistical regularities observed in corpora), semantic lexical relations (i.e., polysemy and hypernymy), and n-gram features (i.e., the frequency, range, and association strength of the bigrams and trigrams). Although the TAALES computes some indices for academic language use, they were not considered in this study, given that the current target text was a novel.

Frequency and range indices are calculated from available corpora such as the British National Corpus (BNC Consortium, 2007), SUBTLEXus (Brysbaert & New, 2009), and the Corpus of Contemporary American English (COCA; Davies, 2009). Indices from both the written and spoken sub-corpora of the BNC were used, whereas only those from the spoken and fiction sub-corpora of COCA were used, given the genre of the current target text.

If any lexical features provided the indices of all words, content words, and function words separately, only indices for content and function words were used (i.e., all-word indices were not used) to reduce redundancy (Eguchi & Kyle, 2020) and increase the interpretability of the results. More details on the lexical indices described above can be found in Appendix S1 in the Supplementary Materials along with additional references (see also Kyle & Crossley, 2015; Kyle et al., 2018).

Syntactic features

The TAASSC was employed to assess syntactic sophistication and complexity within the text (Kyle & Crossley, 2018). It computes traditional indices of syntactic complexity as well as fine-grained indices of clausal and phrasal complexity. The traditional indices contained 14 indices measured by Lu's (2010) L2 syntactic complexity analyzer (SCA), including the mean length of clauses or T-units, the number of dependent clauses per clause or per T-unit, and the number of complex nominals per clause or per T-unit.

The fine-grained indices of clausal complexity included the average number of specific structures (e.g., clausal subject and direct object) per clause within the text. There are

also two general indices of clausal complexity: the mean and standard deviation of the number of dependents per clause. Standard deviation measures indicate the variety of syntax within a text.

Fine-grained indices of phrasal complexity are categorized into three types (Kyle & Crossley, 2018). The first indicates the average number of dependents per specific phrase type (e.g., direct objects) and the average number of dependents for all phrase types as well as their standard deviations. The second represents the incidence of specific dependent types (e.g., adjective modifiers), regardless of the noun phrase type in which they occur. The third is the average occurrence of specific dependent types in specific types of noun phrases (e.g., adjective modifiers occurring in direct object phrases). This study used phrasal complexity indices that do not count pronoun noun phrases, given that the use of pronouns as phrases may bias the counts of dependents (Kyle & Crossley, 2018).

Cohesion features

The TAACO reports on several types of cohesion indices (Crossley & Kyle, 2016; Crossley et al., 2019). The current study focused on the following four categories: lexical overlap, semantic overlap, connectives, and type-token ratio (TTR). Lexical overlap computes a variety of types of lemma overlap (e.g., all lemma overlap and content word lemma overlap) between sentences or paragraphs. Semantic overlap measures the semantic similarity between sentences or paragraphs by employing NLP techniques such as latent semantic analysis (LSA; Landauer, Foltz & Laham, 1998), latent Dirichlet allocation (LDA; Blei, Ng & Jordan, 2003) and Word2Vec (Mikolov, Chen, Corrado & Dean, 2013). Connectives indicate the occurrence of several classes of connectives in the text (e.g., positive, negative, causal, additive, and temporal). TTR measures the diversity of words in the text by dividing the number of types by the number of tokens. Several different TTR indices are reported by the TAACO, including simple, content word, and lemma TTRs.

Connective measures assess cohesion at the local level (between adjacent sentences), whereas lexical and semantic overlap measures assess cohesion at both the local and global levels (between adjacent sentences and between adjacent paragraphs). However, because the current target texts are relatively short and do not necessarily consist of paragraphs, this study adopted only the overlap measures for local cohesion. The TTR indices assess overall text cohesion.

Simple linguistic features

The SiNLP (Crossley et al., 2014) was employed to compute indices of the following simple linguistic features: word count (i.e., the total number of words in the text), average word length (i.e., the average number of letters per word),[3] and average sentence length (i.e., the average number of words per sentence). These indices were used to develop the models for comparison to the models that include complex linguistic features, although word count was also used as a control variable in the complex models.

Statistical analysis

The statistical analysis was primarily conducted with R 4.0.5 (R Core Team, 2021). Multiple regression models were constructed using indices of complex linguistic features of target texts to explain the variance in eye movement measures for L1 and L2 reading. This study used multiple regression models so that the results could be compared with those of Crossley et al.'s (2019) study, which also developed multiple regression readability models, and considering their simplicity, interpretability, and greater theoretical commitments.

The models were developed following the procedures employed by Crossley et al. (2019). Prior to the model construction, the dataset was divided into training and test sets using a 67/33 split, resulting in 388 texts in the training set and 200 texts in the test set, to allow for cross-validation of the models. The model was first derived from the training set

and then applied to the test set to assess the generalizability.

Indices of complex linguistic features were then selected for developing the models. First, all indices were checked for normality, and those that exceeded 2.0 kurtosis or skewness values was removed (George & Mallery, 2010). Then, any index that did not indicate a meaningful ($r > .10$) and significant relationship ($p < .05$) with the component scores of eye movement measures was excluded. Finally, if any multiple indices were highly collinear ($r > .90$), only the index with the strongest relationship to eye movement measures was retained. The remaining indices were then entered into a stepwise multiple regression based on the Akaike information criterion (AIC). The models also included word count as a control variable given that the adopted eye movement measures (i.e., total reading time and the number of fixations) are closely related to text length. If any of the indices in the model demonstrated suppression (i.e., the bivariate correlation and the beta weight had opposite signs; Tabachnick & Fidell, 2013), the model was run again without the indices. This process was repeated until there were no suppressed indices in the models.

After obtaining the final model, a posteriori approach informed by model criticism was conducted (Baayen & Milin, 2010) to alleviate the effects of outliers. The residuals were inspected for the model, and the observations with large residuals (in excess of 3.0 *SD*s in this case), which were poorly predicated by the analysis, were removed (less than 2% of the data). This approach is advantageous in that it potentially leaves a larger portion of the dataset intact while still improving the model fit (Godfroid, 2019). Relative weight analysis was then performed for the final model to partition explained variance among the predictors and assess their relative importance (Tonidandel & LeBreton, 2011).[5] The analysis was conducted by employing Mizumoto's (2022) R-based web application from the langtest.jp, available at http://langtest.jp/shiny/relimp/.

The models constructed using complex linguistic features were then compared with

the models including simple linguistic features. The first simple model included only the word count index, and the second model added the indices of word and sentence length to word count. The index of sentence length in the training set was log-transformed because of the violation of normal distribution. Note that the models with complex linguistic features are likely to include more parameters (i.e., linguistic indices) than the simple linguistic-feature models. In general, the model better fits the observed data if it has a larger number of parameters. However, if the model is too complex, it will not be effective in predicting future phenomena. Thus, this study compared the complex and simple linguistic feature models with the AIC, which is an evaluation criterion for deciding the better model in terms of the prediction. It defines the goodness of fit of the model to the data as the maximum log likelihood and incorporates the number of free parameters of the model as a penalty for the complexity of the model. The model that exhibits the smaller AIC value is considered better. The Fisher *r*-to-*z* transformation was also conducted using Weiss's (2011) calculator to examine the differences in the correlations with eye movement measures reported by the models with complex and simple linguistic features.

Results

L1 reading

The model with complex linguistic features

After controlling for normality, correlations, and multicollinearity, 44 linguistic indices remained for L1 reading. These indices were entered into a stepwise multiple regression, along with the control variable of word count. The initial model included 16 indices. After controlling the suppression effects, 10 indices remained in the model, including the indices of word count, familiarity of content words, frequency of function words, bigram

and trigram frequency, association for function words (types and tokes), phonological neighborhood of content words, standard deviations for the number of dependents per direct object, and average number of prepositions per clause. After conducting the model criticism, the final model was significant, F (10, 371) = 176.20, $p < .001$, $r = .91$, $R^2 = .83$, AIC = 345.60, explaining 83% of the variance in eye movement measures, and four indices in the model were significant predictors[4]: word count, bigram frequency, trigram frequency, and the average number of prepositions per clause (see Table 2). The effects of these indices indicated that processing effort increased when the text included more words, less frequently used bigrams and trigrams, and more prepositions per clause.

<Insert Table 2 about here>

Relative wight analysis indicated that of the variance explained by the model (i.e., 83% of the overall variance), word count accounted for 72%, indicating its major role in this model. Nevertheless, the combination of the other indices explained the remaining 11% of the variance. Of these indices, bigram and trigram frequency contributed the most to the explanation, accounting for 3.9% of the variance in combination (see Appendix S3-2 in the Supplementary Materials for detailed results).[6]

When the model was applied to the test dataset, it reported $r = .92$ and $R^2 = .84$. This demonstrated that the combination of 10 indices accounted for 84% of the variance in the eye movement measures found in the test set and did not support the overfit of the constructed model.

Comparison with simple linguistic feature models

The first simple model that included only word count accounted for 71% of the variance of eye-movement measures in L1 reading, F (1, 382) = 942.10, $p < .001$, $r = .84$, $R^2 = .71$, AIC = 521.74. The second simple model that included word and sentence length in addition to word count accounted for 79% of the variance, F (3, 379) = 479.40, $p < .001$, r

= .89 R^2 = .79, AIC = 401.12, showing the significant effect of both word length and sentence length (see Table 3).

<Insert Table 3 about here>

Nevertheless, the model with complex linguistic features indicated a much smaller AIC score than did the simple models, providing support for the better prediction ability of the complex linguistic feature model. In addition, Fisher *r*-to-*z* transformation indicated that the complex model explained a significantly greater amount of variance than the first simple model ($z = 3.97$, $p < .001$). However, the difference failed to approach significance for the second simple model ($z = 1.39$, $p = .164$).

L2 reading

The model with complex linguistic features

After controlling for normality, correlations, and multicollinearity, 43 linguistic indices remained for L2 reading. Similar to the L1 reading model, they were entered into a stepwise multiple regression along with the control variable of word count. The initial model included 17 indices. After controlling the suppression effects, 12 indices remained in the model, including the indices of word count; familiarity, concreteness, and frequency of content words; frequency, association (types and tokes), and contextual distinctiveness of function words; bigram and trigram frequency; bigram association strength; and the proportion of subordinate clauses. After conducting the model criticism, the final model was significant, F (12, 370) = 108.30, $p < .001$, $r = .88$, $R^2 = .78$, AIC = 457.41, explaining 78% of the variance in eye movement measures, and five indices in the model were significant: word count, frequency of function words, bigram association strength, bigram frequency, and trigram frequency (see Table 4). The model indicated that processing effort increased when

the text included more words, less frequent function words, more strongly associated bigrams, and less frequent bigrams and trigrams.

<Insert Table 4 about here>

Relative weight analysis found that of the variance explained by the model (i.e., 78% of the overall variance), word count accounted for approximately 67%, indicating its major role in this model. Nevertheless, the combination of the other indices explained the remaining 11% of the variance. Of these indices, the bigram and trigram frequency contributed the most to the explanation, accounting for 3.7% of the variance in combination (see Appendix S3-3 in the Supplementary Materials for detailed results).

The model was then applied to the test dataset and reported $r = .88$, $R^2 = .78$. This demonstrated that the combination of 12 indices accounted for 78% of the variance in the eye movement measures found in the test set and did not support the overfit of the constructed model.

Comparison with simple linguistic feature models

The first simple model (i.e., word count model) accounted for 67% of the variance in eye movement measures for L2 reading, F (1, 381) = 763.00, $p < .001$, $r = .82$, $R^2 = .67$, AIC = 587.06. The second simple model (i.e., word and sentence length model) accounted for 71% of the variance, F (3, 380) = 305.50, $p < .001$, $r = .84$, $R^2 = .71$, AIC = 546.72, indicating the significant effect of only word length (see Table 5).

<Insert Table 5 about here>

Similar to L1 reading, the complex model indicated a much smaller AIC score than those of the simple models, providing support for better prediction of the complex model. Moreover, the comparison using Fisher *r*-to-*z* transformation indicated that the model with

complex linguistic features explained a significantly greater amount of variance than both the first simple model ($z = 3.30$ $p < .001$) and the second simple model ($z = 2.23$, $p = .026$).

Discussion

Predicting reading processing effort using text linguistic features (RQ1)

(a) The use of complex linguistic features

RQ1 compared the performance of the models that used complex linguistic features with that of the models that used simple linguistic features. The analysis yielded significant models including several complex linguistic features in addition to word count, which explained 83% and 78% of the variance in eye movement measures (i.e., the combination of total reading times and the number of fixations) for L1 and L2 reading, respectively (see Appendix S4-1 in the Supplementary Materials for an explanation of this large variance accounted for by the models). These complex feature models provided better performance in predicting processing effort for both L1 and L2 reading than either the word count models or the models including word and sentence length in addition to word count. In addition, the complex models explained approximately 12% and 11% more of the variance in eye movement measures than did the word count models and 4% and 7% more than did the word and sentence length models for L1 and L2 reading, respectively. Although the difference between the models was less than expected, it is reasonable given that word length is a determining factor for eye movements during reading (Cop et al., 2015; Reichle et al., 1999).

There is no doubt that readers need more time and effort to process text as it increases in length. Therefore, it is not surprising that the word count accounted for the large variance in the processing effort calculated from the total reading times and the number of fixations. However, the inclusion of indices of complex linguistic features in addition to word count significantly improved the performance of the models. Moreover, this improvement

was greater than when adding the indices of word and sentence length to the word count model. Thus, these results support the notion that more complex linguistic features, rather than simple linguistic features, are more useful for the accurate estimation of readability (e.g., Crossley et al., 2011; Crossley et al., 2019; Piler & Nenkova, 2008). Importantly, the current study extended this view from readability for comprehension to readability for processing. This is noteworthy given that comprehension does not always correlate with processing (Dirix et al., 2019; Kuperman et al., 2022).

(b) The comparison of L1 and L2 reading models

As described above, the complex linguistic features were useful for the prediction of processing effort during both L1 and L2 reading. However, the difference in the variance explained by the complex and simple feature (word and sentence length) models failed to reach significance for L1 reading, whereas it was significant for L2 reading. This suggests that the use of complex linguistic features is more important for estimating processing effort involved during L2 reading than L1, which is partly in line with the previous finding that the effect of some complex linguistic features (e.g., word frequency) can be larger for L2 or lower-proficiency readers (Cop et al., 2015; Dirix & Duyck, 2017; Kim et al., 2018).

In addition, the variance explained by the L2 reading models was always lower by 4% to 8% compared to the L1 reading models. The remaining variance for L2 reading might be explained by individual differences such as L2 proficiency, given the large differences in L2 proficiency in this dataset compared to L1 proficiency (see Cop et al., 2015). This suggests the benefits of including the variable of proficiency or the interaction of proficiency and linguistic features for L2 readability models (see also the Limitations and future directions).

Linguistic features that are predictive of reading processing effort (RQ2)

(a) The importance of lexical features

RQ2 concerned which linguistic features are predictive of reading processing effort. Although this study assessed lexical, syntactic, and cohesion features, many of the indices included in the final models were lexical in nature (e.g., word frequency, familiarity, concreteness, association, and n-gram features) for both L1 and L2 reading. These results suggest that after controlling for the effects of text length, processing effort during text reading is more likely to be explained by lexical features than by syntactic and cohesion features. This supports the view that lexical indices are the most predictive variables in readability models (e.g., Crossley et al., 2022; Crossley, Skalicky et al., 2017; Pitler & Nenkova, 2008). Again, it is worth mentioning that the current study extended this view from readability for comprehension to readability for processing. In addition, the importance of lexical features in modeling eye movements during reading partly concurs with E-Z reader (Reichle et al., 1999), which assumes a large effect of word length and frequency on eye movements (see Appendix S4-2 in the Supplementary Materials for further discussion compared with E-Z reader as well as Graesser and McNamara's [2011] model).

When comparing the current models with Crossley et al.'s (2019) model (based on the human judgments of reading speed), they are similar in that both contained several lexical indices but no cohesion indices. However, Crossley et al.'s model included the index of naming response times, which can be a direct measure of ease of lexical processing similar to lexical decision times, whereas no such indices were included in the current eye-movement-based models. Nevertheless, this seems reasonable given the small variance shared between lexical decision times and some eye movement measures (Dirix et al., 2019; Kuperman et al., 2013).

(b) The comparison of L1 and L2 reading models

L2 reading showed much more and longer fixations than L1 reading (see Table 1), supporting the long-standing notion that L2 reading requires more cognitive effort than L1 reading (Cop et al., 2015; Kuperman et al., 2022; Nisbet et al., 2021). However, despite such a prominent difference, lexical features played a dominant role for estimating both L1 and L2 reading effort after controlling the text length effect. Although this is not congruent with the expectation that L2 reading models depend on lexical indices to a great extent than L1 reading models, the finding emphasizes the importance of decoding for fluent reading regardless of the language used (Koda, 2005; Yamashita, 2013).

Recent studies on bilingual reading have suggested that L1 and L2 reading fluency, as assessed by eye movement measures, are explained by readers' cognitive speed of information processing and reading component skill (Kuperman et al., 2022; Nisbet et al., 2021). The current study is similar to these studies in that it suggests the cognitive factor that explains L1 and L2 reading fluency. This study found that both L1 and L2 reading fluency largely depend on the decoding difficulty of the text as assessed by lexical sophistication, which suggests the cognitive process underlying the similarity of L1 and L2 reading.

Focusing on individual linguistic features, indices of bigram and trigram frequency were included in both L1 and L2 reading models, and they were significant predictors. In addition, except for the word count, these indices contributed the most to both models, as shown by the relative weight analysis, although care should be taken in interpreting this result given the sampling error in rank order of weights and overlap of confidence intervals between the predictors (Mizumoto, 2022; Tonidandel & LeBreton, 2011). Taken together, these results suggest that the frequency of multiword sequences, rather than the frequency of individual words, plays a role in explaining the processing effort for both L1 and L2 reading. Processing effort decreased when the text included highly frequent bigrams and trigrams (see Appendix S5 in the Supplementary Materials for example texts with higher and lower scores for bigram

and trigram frequency), which is in accordance with the view of previous studies on formulaic language (e.g., Li et al., 2021; Tabossi et al., 2009; Siyanova-Chanturia et al., 2011; Sonbul, 2015). Additionally, it also supports the view that formulaic language is processed more quickly even by L2 speakers, particularly those with higher L2 proficiency (e.g., Conklin & Schmitt, 2008; Underwood et al., 2004). Given that 17 out of 19 bilinguals in the current dataset were classified as upper-intermediate to advanced L2 users (see Cop et al., 2015 for more details), it is reasonable to assume that most of the bilinguals in the current dataset were proficient enough in their L2 to enjoy the advantage of formulaic language.

On the other hand, the index of bigram association strength (based on the fiction sub-corpus) was only included and significant in the L2 reading model. Given that bigram frequency was also a significant predictor in the L2 model, this result supports the view that "n-gram frequency and strength-of-association indices may capture related but different aspects of collocational knowledge" (Kyle et al., 2018, p. 1041). Processing effort increased for the text that contained more strongly associated bigrams (the bivariate correlation was also positive). Although there are possible explanations for this finding (see Appendix S4-3 in the Supplementary Materials), further investigation is required to determine whether and why strongly associated bigrams require more time and effort to process in L2 reading.

Furthermore, the L1 model included the two indices of syntactic features, the standard deviations for the number of dependents per direct object and the average number of prepositions per clause, which explained 2.7% of the variance in combination. On the other hand, the L2 model included an index of syntactic features, the proportion of subordinate clauses in the text, which explained only 1% of the variance. These differences may indicate that syntactic features are related to L1 text processing to a greater extent than L2 text processing, as suggested by the shallow structure hypothesis (Clahsen & Felser, 2006). However, given that the variance explained by these syntactic features was small compared to that explained by other linguistic features (i.e., lexical features) in both language models, it is

plausible to interpret this result as suggesting that syntactic features are less related to processing effort than lexical features in both L1 and L2 reading.

Conclusion and implications

The findings of the current study are summarized as three points. First, the use of several complex linguistic features in addition to word count in the readability models led to a better prediction of the processing effort during both L1 and L2 reading as evidenced by eye movements. These features were more useful for prediction than were simple linguistic features, such as word and sentence length, particularly for L2 reading. Second, most of the linguistic indices that were useful for prediction were lexical features for both L1 and L2 reading. Third, except for the word count, the frequency of multiword sequences (bigrams and trigrams) is likely to play a key role in predicting both L1 and L2 reading processing effort.

Although the results showed some differences between L1 and L2 reading (e.g., L2 reading was much more effortful than L1, and a few different indices explained the L1 and L2 reading behavior), this study concludes the similar qualitative relationship between processing effort and text linguistic features for L1 and L2 reading. Importantly, this similarity raises the possibility of building a single readability model that estimates the processing effort for both L1 and L2 reading (i.e., readability can be scaled similarly regardless of the L1 or L2), which will be of practical use.

The findings of this study suggest that future investigation of readability for processing should consider the effects of more varied and complex linguistic features and that the use of lexical sophistication indices, in particular, will be helpful in providing a more accurate readability assessment for both L1 and L2 texts. Although a similar idea has been already proposed by several studies (e.g., Crossley et al., 2011, 2022), most are related to readability for comprehension and not to processing itself. The current findings deepen our

understanding of text readability in terms of the cognitive difficulty that readers experience in real time (Collins-Thompson, 2014).

From a theoretical perspective, this study emphasizes the importance of decoding for both L1 and L2 fluent reading (Koda, 2005; Yamashita, 2013). It also sheds light on the processing of multiword units. Although extant studies have already provided evidence for the processing advantage of formulaic language in both L1 and L2 reading (e.g., Conklin & Schmitt, 2008; Siyanova et al., 2011; Sonbul, 2015; Tabossi et al., 2009; Underwood et al., 2004), the current findings are distinct from these investigations in that the effects of phrasal frequency remained significant for both L1 and L2 models even after controlling for the effects of several other linguistic features. Thus, continuing to explore the processing of multiword units may be key to understanding how fluent bilingual reading can be achieved.

Limitations and future directions

This study has several limitations, offering promising directions for future studies. First, it targeted the processing of a single novel, which limits the generalizability of the current findings. A novel text typically includes many conversations, which do not necessarily illustrate syntactic complexity and higher levels of cohesion. This might explain why lexical features, rather than syntactic and cohesion features, played an important role in the current readability models.

In relation, attention should be paid to the language registers reflected by the corpus used for readability assessment. Although this study examined the indices of n-gram features based on the spoken and fiction sub-corpora of COCA, the indices selected for the final models were those from the fiction sub-corpus. Given the genre of the current target text (i.e., a novel), this suggests the importance of selecting appropriate language registers for readability assessment. Therefore, it may be reasonable to develop readability models using

specific language registers according to the text genre being assessed or to incorporate genre-related features in the readability models (Collins-Thompson, 2014). Similarly, it would be beneficial to use a corpus that more accurately approximates the exposure experienced by target readers. For example, given that the monolingual readers in the current dataset are British and that the L2 readers are taught British English as well, the use of SUBTLEX-UK rather than SUBTLEX-US may contribute to explaining additional variance in processing effort (van Heuven et al., 2014).

Second, a cross-validation of the current models should be conducted with either other eye-tracking corpus data or a small set of new data. Although the current study demonstrated that the models based on the training set predicted the variance in the test set well, the eye-tracking reading times are quite consistent on the individual level in the current dataset (see Dirix et al., 2019). For this purpose, the Multilingual Eye-Movements Corpus (MECO) for L1 (Siegelman et al., 2022) and L2 reading (Kuperman et al., 2022), the latest publicly available eye movement data for bilingual reading, might be a possible data source.

Third, this study did not consider reader variables, such as proficiency in the target language, reading component skills, L1 background, motivation, and strategy use. Although few readability models do not include these variables because readability is usually computed without assuming the specific audience, it is beneficial to discuss and examine the role of these variables in text readability assessment (Collins-Thompson, 2014; Rets & Rogaten, 2021). Additionality, it may be worth considering the interaction effect of reader and text variables (Cop et al., 2015; Dirix & Duyck, 2017; Kim et al., 2018).

Fourth, this study used global reading measures, specifically, total time and fixation count, as dependent variables. Thus, the current models did not discriminate between early (or lower-level) and late (or higher-level) processing and should be interpreted as estimating cognitive effort at both lower- and higher-levels of processing. Given this limitation, the use

of early and late processing measures, separately, will allow us to obtain more nuanced insight into the processing effort and text linguistic features (see Appendix S4-4 in the Supporting Materials for more details).

Finally, although this study adopted generalized linear models (multiple regression) as a modeling approach, several alternatives exist (e.g., generalized linear mixed-effects model and tree-based model). In particular, recent linguists find tree-based approaches (particularly random forests) appealing as an alternative to regression approaches, although each approach has its own advantages and disadvantages (see Gries [2021] for more detailed discussion). In addition, there are several other text features that can be obtained via more advanced linguistic techniques and used for the readability model (e.g., De Clercq & Hoste, 2016). According to Collins-Thompson (2014), the choice of features can be more important than the choice of modeling approaches in readability assessment. Further discussion and investigations are needed to determine more appropriate and effective methods for developing readability models.

Competing interests

The author declares none.

Data Availably Statement

The materials and data that support the findings of this study are openly available in OSF at https://osf.io/uzhxr/?view_only=481a251a8fb0442d9b66f8e037d9c6ab.

Supplementary Materials

For supplementary materials accompanying this paper, visit xxxxxxxxxxxxxxxxxxxxxx.

References

Ardoin, SP, Suldo, SM, Witt, J, Aldrich, S, & McDonald, E (2005) Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly*, 20, 1–22.

https://doi.org/10.1521/scpq.20.1.1.64193

Baayen, RH and Milin, P (2010) Analyzing reaction times. *International Journal of Psychological Research* 3, 12–28. https://doi.org/10.21500/20112084.807

Bertram, B and Newman, S (1981) Why readability formulas fail (Report No. 28). Illinois University, Urbana: Center for the Study of Reading (Eric Document Service No. ED205915).

Blei, DM, Ng, AY and Jordan, MI (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

BNC Consortium, (2007) The British National Corpus, version 3. BNC Consortium. Retrieved from www.natcorp.ox.ac.uk

Britton, BK and Gülgöz, S (1991) Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology* 83(3), 329–345. https://doi.org/10.1037/0022-0663.83.3.329

Brown, JD, (1998) An EFL readability index. *JALT Journal* 20, 7–36.

Brysbaert, M and New, B (2009) Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41, 977–990. https://doi.org/10.3758/BRM.41.4.977

Buchanan, L, Westbury, C and Burgess, C (2001) Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review* 8(3), 531–544. https://doi.org/10.3758/BF03196189

Chaffin, R, Morris, RK and Seely, RE (2001) Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(1), 225–235. https://doi.org/10.1037/0278-7393.27.1.225

Clahsen, H and Felser, C (2006) How native-like is non-native language processing? *Trends in Cognitive Sciences* 10(12), 564–570. https://doi.org/10.1016/j.tics.2006.10.002

Collins-Thompson, K (2014) Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165(2), 97–135. https://doi.org/10.1075/itl.165.2.01col

Conklin, K and Schmitt, N (2008) Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29(1), 7289. https://doi.org/10.1093/applin/amm022

Conklin, K, Pellicer-Sánchez, A and Carrol, G (2018) *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.

Cop, U, Drieghe, D and Duyck, W (2015) Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PLOS One* 10, e0134008. https://doi.org/10.1371/journal.pone.0134008

Cop, U, Dirix, N, Drieghe, D and Duyck, W (2017) Presenting GECO: An eye tracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods* 49(2), 602–615. https://doi.org/10.3758/s13428-016-0734-0

Cop, U, Keuleers, E, Drieghe, D and Duyck, W (2015) Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review* 22(5), 1216–1234. https://doi.org/10.3758/s13423-015-0819-2

Crossley, SA, Allen, LK, Kyle, K and McNamara, DS (2014) Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes* 51(5-6), 511–534. https://doi.org/10.1080/0163853X.2014.910723

Crossley, SA, Allen, DB and McNamara, DS (2011) Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language* 23, 84–101.

Crossley, SA, Allen, D and McNamara, DS (2012) Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research* 16, 89–108. https://doi.org/10.1177/1362168811423456

Crossley, SA, Greenfield, J and McNamara, DS (2008) Assessing text readability using cognitively based indices. *TESOL Quarterly* 42, 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, SA, Heintz, A, Choi, J, Batchelor, J, Karimi, M and Malatinszky, A (2022) A large-scaled corpus for assessing text readability. *Behavioral Research Methods*. https://doi.org/10.3758/s13428-022-01802-x

Crossley, SA., Kyle, K and Dascalu, M (2019) The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavioral Research Methods* 50, 1030–1046. https://doi.org/10.3758/s13428-018-1142-4

Crossley, SA, Kyle, K and McNamara, DS (2016) The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48, 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, SA, Kyle, K and McNamara, DS (2017) Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods* 49, 803–821. https://doi.org/10.3758/s13428-016-0743-z

Crossley, SA, Skalicky, S and Dascalu, M (2019) Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading* 42, 541–561. https://doi.org/10.1111/1467-9817.12283

Crossley, SA, Skalicky, S, Dascalu, M, McNamara, DS and Kyle, K (2017) Predicting text comprehension, processing, and familiarity in adult readers: New approaches to

readability formulas. *Discourse Processes* 54(5-6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Dale, E and Chall, JS (1949) The concept of readability. *Elementary English* 26, 19–26.

Davies, M (2009) The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14, 159–190. https://doi.org/10.1075/ijcl.14.2.02da

Dascalu, M, Dessus, P, Bianco, M, Trausan-Matu, S and Nardy, A (2014) Mining texts, learner productions and strategies with ReaderBench. In Peña-Ayala, A (ed.), *Educational data mining: Applications and trends*. Cham, Switzerland: Springer, pp. 345–377.

De Clercq, O and Hoste, V (2016) All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics* 42(3), 457–490. https://doi.org/10.1162/COLI_a_00255

Dirix, N and Duyck W (2017) The first-and second-language age of acquisition effect in first-and second-language book reading. *Journal of Memory and Language* 97, 103–120. https://doi.org/10.1016/j.jml.2017.07.012

Dirix, N, Vander Beken, H, De Bruyne, E, Brysbaert, M and Duyck, W (2019) Reading text when studying in a second language: An eye-tracking study. *Reading Research Quarterly* 55(3), 371–397. https://doi.org/10.1002/rrq.277

Eguchi, M and Kyle, K (2020) Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal* 104(2), 381–400. https://doi.org/10.1111/modl.12637

Feng, L, Jansche, M, Huenerfauth, M and Elhadad, N (2010) A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 276–284). USA: Association for

Computational Linguistics.

Flesch, R (1948) A new readability yardstick. *Journal of Applied Psychology* 32, 221–233. https://doi.org/10.1037/h0057532

Fry, EB (1989) Reading formulas – maligned but valid. *Journal of Reading* 32, 292–297.

Garner, J, Crossley, S and Kyle, K (2019) N-gram measures and L2 writing proficiency. *System* 80, 176–187. https://doi.org/10.1016/j.system.2018.12.001

George, D and Mallery, P (2010) *SPSS for Windows step by step: A simple guide and reference 17.0 Update* (10th ed.). Boston, MA: Pearson.

Godfroid, A (2019) *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. New York, NY: Routledge.

Godfroid, A and Hui, B (2020) Five common pitfalls in eye-tracking research. *Second Language Research* 36(3), 277–305. https://doi.org/10.1177/0267658320921218

Gollan, TH, Montoya, RI, Cera, C and Sandoval, TC (2008) More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language* 58(3), 787–814. https ://doi.org/10.1016/j.jml.2007.07.001

Grabe, W (2009) *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.

Graesser, AC and McNamara, DS (2011) Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science* 3(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Greenfield, G (1999) Classic readability formulas in an EFL context: Are they valid for Japanese speakers? Unpublished doctoral dissertation, Temple University, Philadelphia, PA, United States (University Microfilms No. 99–38670).

Greenfield, J (2004) Readability formulas for EFL. *JALT Journal* 26, 5–24.

Gries, ST (2021) (Generalized linear) Mixed-effects modeling: A learner corpus example.

Language Learning. https://doi.org/10.1111/lang.12448

Hamsik, MJ (1984) Reading, readability, and the ESL reader. Unpublished doctoral dissertation, The Florida State University, U.S. Hoffman, P, Lambon Ralph, MA, & Rogers, TT (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45, 718–730. https://doi.org/doi:10.3758/s13428-012-0278-x

Holmes, VM and O'Regan, JK (1981) Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior* 20, 417–430. https://doi.org/10.1016/S0022-5371(81)90533-8

Horiba, Y (1996) Comprehension processes in L2 reading: Language competence, textual coherence, and inferences. *Studies in Second Language Acquisition* 18, 433–473. https://doi.org/10.1017/S0272263100015370

Horiba, Y (2000) Reader control in reading: Effects of language competence, text type, and task. *Discourse Processes* 29(3), 223–267. https://doi.org/10.1207/S15326950dp2903_3

Jeon, EH and Yamashita, J (2014) L2 reading comprehension and its correlates: A meta-analysis. *Language Learning* 64, 160–212. https://doi.org/10.1111/lang.12034

Juhasz, BJ and Rayner, K (2003) Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 29(6), 1312–1318. https://doi.org/10.1037/0278-7393.29.6.1312

Just, MA and Carpenter, PA (1980) A theory of reading: From eye fixations to comprehension. *Psychological Review* 87, 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kate, RJ, Luo, X, Patwardhan, S, Franz, M, Florian, R, Mooney, RJ, Roukos, R, and Welty,

C (2010) Learning to predict readability using diverse linguistic features. In *In Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 546–554). USA: Association for Computational Linguistics.

Kim, M, Crossley, SA and Skalicky, S (2018) Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing* 31, 1155–1180. https://doi.org/10.1007/s11145-018-9833-x

Kincaid, JP, Fishburne, RP, Rogers, RL and Chissom, BS (1975) *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Research Branch Report 8–75. Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Kintsch, W (1998) *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Koda, K (2005) *Insights into second language reading: A cross-linguistic approach.* New York, NY: Cambridge University Press.

Kuperman, V, Siegelman, N, Schroeder, S, Alexeeva, A, Acartürk, C, Amenta, S, ... Usual, KA (2022) Text reading in English as a second language: Evidence from the multilingual eye-movements corpus (MECO). *Studies in Second Language Acquisition*. https://doi.org/ 10.1017/S0272263121000954

Kyle, K and Crossley, SA (2015) Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49, 757–786. https://doi.org/10.1002/tesq.194

Kyle, K and Crossley, SA (2018) Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal* 102, 333–349. https://doi.org/10.1111/modl.12468

Kyle, K, Crossley, S and Berger, C (2018) The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Landauer, TK, Foltz, PW and Laham, D (1998) An introduction to latent semantic analysis. *Discourse Processes* 25(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lu, X (2010) Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Li, H, Warrington, KL, Pagán, A, Paterson, KB and Wang, X (2021) Independent effects of collocation strength and contextual predictability on eye movements in reading. *Language, Cognition and Neuroscience,* 1–9. https://doi.org/10.1080/23273798.2021.1922726

Marinis, T, Roberts, L, Felser, C and Clahsen, H (2005) Gaps in second language sentence processing. *Studies in Second Language Acquisition,* 27(1), 53–78. https://doi.org/10.1017/S0272263105050035

McDonald, SA and Shillcock, RC (2001) Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech* 44(3), 295–322. https://doi.org/10.1177/00238309010440030101

McNamara, DS, Graesser, AC, McCarthy, PM and Cai, Z (2014) *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

Mesmer, HAE (2008) *Tools for matching readers to texts: Research-based practices*. New York, NY: Guilford Press.

Mikolov, T, Chen, K, Corrado, G and Dean, J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv,* 1301–3781.

Mizumoto, A (2022). Calculating the relative importance of multiple regression predictor

variables using dominance analysis and random forests. *Language Learning* (Early View). https://doi.org/10.1111/lang.12518

Morishima, Y (2013) Allocation of limited cognitive resources during text comprehension in a second language. *Discourse Processes* 50, 577–597. https://doi.org/10.1080/0163853X.2013.846964

Nahatame, S (2018) Comprehension and processing of paired sentences in second language reading: A comparison of causal and semantic relatedness. *Modern Language Journal* 102, 392–415. https://doi.org/10.1111/modl.12466

Nahatame, S (2020) Revisiting second language readers' memory for narrative texts: The role of causal and semantic text relations. *Reading Psychology* 41(8)*,* 753–777. https://doi.org/10.1080/02702711.2020.1768986

Nahatame, S (2021) Text readability and processing effort in second language reading: A computational and eye-tracking investigation. *Language Learning*, 71(4), 1004–1043. doi.org/10.1111/lang.12455.

Nahatame, S (2022) Causal and semantic relations in second language text processing: An eye-tracking study. *Reading in a Foreign Language*, 34(1), 91–115. https://nflrc.hawaii.edu/rfl/item/546

Nisbet, K, Bertram, R, Erlinghagen, C, Pieczykolan, A and Kuperman, V (2021) Quantifying the difference in reading fluency between L1 and L2 readers of English. *Studies in Second Language Acquisition,* 1–28. https://doi:10.1017/S0272263121000279

Ozuru, Y, Dempsey, K and McNamara, DS (2009) Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction* 19(3), 228–242. https://doi.org/10.1016/j.learninstruc.2008.04.003

Pitler, E and Nenkova, A (2008) Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural*

Language Processing (pp. 186–195). USA: Association for Computational Linguistics.

Rayner, K (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, 372–422. https://doi.org/:10.1037/0033-2909.124.3.372

Rayner, K, Chace, K, Slattery, TJ and Ashby, J (2006) Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading* 10, 241–255. https://doi.org/ 10.1207/s1532799xssr1003_3

R Core Team (2021) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/

Reichle, ED, Rayner, K and Pollatsek, A (1999) Eye movement control in reading: Accounting for initial fixation locations and refixations within the EZ Reader model. *Vision Research* 39(26), 4403–4411. https://doi.org/10.1016/S0042-6989(99)00152-2

Rets, I (2021) Linguistic accessibility of Open Educational Resources: Text Simplification as an aid to non-native readers of English. Doctoral dissertation, the Open University. http://oro.open.ac.uk/75140/1/Rets_thesis_ORO.pdf

Rets, I and Rogaten, J (2021) To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. *Journal of Computer Assisted Learning* 37(3), 705–717. https://doi.org/10.1111/jcal.12517

Richards, JC and Schmidt, RW (2013) *Longman dictionary of language teaching and applied linguistics*. New York, NY: Routledge.

Siyanova-Chanturia, A, Conklin, K and Schmitt, N (2011) Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research* 27(2), 251–272. https://doi.org/10.1177/0267658310382068

Sonbul, S (2015) Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition* 18(3), 419–437. https://doi.org/10.1017/S1366728914000674

Tabachnick, BG and Fidell, LS (2014) *Using multivariate statistics* (6th Ed.). Boston, MA: Pearson Education Limited.

Tabossi, P, Fanari, R and Wolf, K (2009) Why are idioms recognized fast? *Memory & Cognition* 37(4), 529–540. https://doi.org/10.3758/MC.37.4.529

Tonidandel, S and LeBreton, JM (2011) Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology* 26(1), 1–9. https://doi.org/10.1007/s10869-010-9204-3

Underwood, G, Schmitt, N and Galpin, A (2004) The eyes have it: An eye-movement study into the processing of formulaic sequences. In Schmitt, N (ed.), *Formulaic sequences*. Amsterdam, the Netherlands: John Benjamins.

Van Heuven, WJ, Mandera, P, Keuleers, E and 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

Weiss, BA (2011) Fisher's r-to-Z transformation calculator to compare two independent samples [Computer software]. Available from https://blogs.gwu.edu/weissba/teaching/calculators/fishers-z-transformation/

Whitford, V and Titone, D (2017) The effects of word frequency and word predictability during first- and second-language paragraph reading in bilingual older and younger adults. *Psychology and Aging* 32(2), 158–177. https://doi.org/10.1037/pag0000151

Yamashita, J (2013) Word recognition subcomponents and passage level reading in a foreign language. *Reading in a Foreign Language* 25, 52–71.

Yap, MJ and Balota, DA (2009) Visual word recognition of multisyllabic words. *Journal of Memory and Language* 60(4), 502–529. https://doi.org/10.1016/j.jml.2009.02.001

Yarkoni, T, Balota, D and Yap, M (2008) Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review* 15(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Yeari, M, van den Broek, P and Oudega, M (2015) Processing and memory of central versus peripheral information as a function of reading goals: Evidence from eye-movements. *Reading and Writing* 28(8), 1071–1097. https://doi.org/10.1007/s11145-015-9561-4

Zhang, X and Li, W (2021) Effects of n-grams on the rated L2 writing quality of expository essays: A conceptual replication and extension. *System* 97, 102437. https://doi.org/10.1016/j.system.2020.102437

Zufferey, S, Mak, W, Degand, L and Sanders, T (2015) Advanced learners' comprehension of discourse connectives: The role of L1 transfer across on-line and off-line tasks. *Second Language Research* 31, 389–411. https://doi.org/10.1177/0267658315573349

Footnotes

1. Brown (1998) did not demonstrate strong correlations between traditional formulas and L2 English text difficulty. Given this result, Greenfield (2004) suggested that traditional formulas can be more predictive for specific types of L2 text.
2. *N*-grams are contiguous word sequences of *n* number of words. Researchers have most often investigated bigrams (e.g., *very much*) and trigrams (e.g., *a lot of*).
3. Although some traditional readability measures include the number of syllables per word as an index of word length (e.g., Flesch formula), it usually highly correlates with the number of letters per word, which only the SiNLP calculates.
4. The remaining indices were originally significant or approached significance before controlling the suppression effects and conducting model criticism. This is also true for the L2 reading model. Although these processes made some indices insignificant, they were necessary for model improvement and interpretation. The results from the model post-criticism are more reliable and indicate an effect "that is actually supported by the majority of data points" (Baayen & Milin, 2010, p. 26). In addition, the suppressed variables are often removed during the construction of multiple regression models (e.g., Crossley et al., 2022; Kyle et al., 2018). This study found no prominent difference in the variance explained by the current models with and without suppressed variables (83.4% vs. 82.6% for L1; 79.5% vs. 77.8% for L2).
5. The use of the standardized beta coefficients to determine the relative importance of predictors in the regression model has been criticized as the misuse of multiple regression analysis, and some alternative approaches are recommended, including relative weight analysis (see Appendix S3-1 in the Supplementary Materials for more details).
6. Although relative weight analysis and dominance analysis are largely interchangeable with one another, Mizumoto (2022) recommends the use of the latter accompanied by

random forest, a machine learning method. The results of these analyses are also available in Appendices S3-2 and S3-3 in the Supplementary Materials.

Table 1. *Total Reading Time and Number of Fixations (per Trial) for L1 and L2 Reading*

	Total Reading Time (ms)		Number of Fixations	
	M	*SD*	*M*	*SD*
L1	15,039.19	5,313.44	68.33	20.80
L2	18,631.59	6,307.95	81.85	24.89

Note. N = 8,206 for L1 reading; 5,566 for L2 reading.

Table 2. *Multiple Regression Model with Complex Linguistic Features for L1 Reading*

	r	β [95% CI]	*B*	*SE*	*t*	*p*
Word.Count	.85	0.88 [0.82, 0.91]	0.04	0.00	39.15	< .001*
MRC_Familiarity_CW (familiarity of content words)	−.10	−0.01 [−0.05, 0.04]	0.00	0.00	−0.37	.711
SUBTLEXus_Freq_FW (frequency of function words)	−.19	−0.03 [−0.08, 0.02]	0.00	0.00	−1.13	.259
COCA_fiction_bi_prop_70k (bigram frequency)	−.16	−0.18 [−0.25, −0.11]	−2.15	0.40	−5.44	< .001*
COCA_fiction_tri_prop_40k (trigram frequency)	−.15	−0.09 [−0.15, −0.03]	−1.37	0.50	−2.75	.006*
eat_types_FW (association for function words)	.16	0.03 [−0.02, 0.08]	0.01	0.01	1.21	.229
eat_tokens_FW (association for function words)	−.09	−0.03 [−0.07, 0.02]	−0.01	0.00	−1.32	.188
PLD_CW (phonological neighborhood of content words)	.14	0.04 [−0.01, 0.10]	0.21	0.15	1.47	.142
dobj_NN_stdev (*SD* of the number of dependents per direct object)	.16	0.03 [−0.02, 0.07]	0.05	0.05	1.06	.292
prep_per_cl (number of prepositions per clause)	.20	0.05 [0.01, 0.10]	0.30	0.13	2.259	.024*

Note. $N = 382$. $R^2 = .83$. Asterisks indicate that the *p* values are significant at alpha = .05. See Table S2 in the Supplementary Materials for a more detailed description of the indices.

Table 3. *Multiple Regression Model with Simple Linguistic Features for L1 Reading*

	r	β [95% CI]	*B*	*SE*	*t*	*p*
Word.Count	.85	0.90 [0.85, 0.94]	0.05	0.00	37.44	< .001*
word length (number of letters per word)	.13	0.27 [0.22, 0.32]	0.80	0.07	11.12	< .001*
sentence length_Log (number of words per sentence)	.07	−0.06 [−0.11, −0.01]	−0.14	0.05	−2.52	.012*

Note. $N = 383$. $R^2 = .79$. Asterisks indicate that the *p* values are significant at alpha = .05.

Table 4. *Multiple Regression Model with Complex Linguistic Features for L2 Reading*

	r	β	*B*	*SE*	*t*	*p*
Word.Count	.82	0.84 [0.79, 0.89]	0.04	0.00	33.16	< .001*
MRC_Familiarity_CW (familiarity of content words)	−.12	−0.04 [−0.10, 0.02]	0.00	0.00	−1.27	.206
MRC_Concreteness_CW (concreteness of content words)	.12	0.01 [−0.05, 0.08]	0.00	0.00	0.45	.650
SUBTLEXus_Freq_CW (frequency of content words)	−.18	−0.05 [−0.13, 0.02]	0.00	0.00	−1.42	.156
BNC_Spoken_Freq_FW_Log (frequency of function words)	−.11	−0.08 [−0.13, −0.02]	−0.61	0.21	−2.86	.005*
COCA_fiction_bi_DP (bigram association strength)	.15	0.07 [0.01, 0.12]	4.72	2.06	2.29	.022*
COCA_fiction_bi_prop_70k (bigram frequency)	−.15	−0.10 [−0.17, −0.02]	−1.14	0.44	−2.57	.011*
COCA_fiction_tri_prop_40k (trigram frequency)	−.17	−0.15 [−0.22, −0.08]	−2.45	0.60	−4.08	< .001*
eat_types_FW (association of function words)	.19	0.04 [−0.02, 0.10]	0.01	0.01	1.28	.203
eat_tokens_FW (association of function words)	−.10	−0.02 [−0.08, 0.03]	0.00	0.00	−0.86	.392
McD_CD_FW (contextual distinctiveness of function words)	.16	0.04 [−0.01, 0.09]	0.20	0.12	1.67	.096
DC_C (proportion of dependent clauses)	.13	0.04 [−0.01, 0.09]	0.25	0.17	1.46	.144

Note. $N = 383$. $R^2 = .78$. Asterisks indicate that the *p* values are significant at alpha = .05. See Table S3 in the Supplementary Materials for a more detailed description of the indices.

Table 5. *Multiple Regression Model with Simple Linguistic Features for L2 Reading*

	r	β [95% CI]	*B*	*SE*	*t*	*p*
Word.Count	.81	0.84 [0.78, 0.89]	0.04	0.00	29.53	< .001*
word length (number of letters per word)	.09	0.21 [0.16, 0.27]	0.64	0.09	7.43	< .001*
sentence length_Log (number of words per sentence)	.17	0.05 [0.00, 0.11]	0.12	0.07	1.80	.072

Note. $N = 384$. $R^2 = .71$. Asterisks indicate that the *p* values are significant at alpha = .05.