Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–16

DOI: 10.1177/17470218251317372
qjep.sagepub.com
Sage

Global measures of syntactic and lexical complexity are not strong predictors of eye-movement patterns in sentence and passage reading

Victor Kuperman[1], Dalmo Buzato[2] and Rui Rothe-Neves[2]

Abstract

The link between the cognitive effort of word processing and the eye-movement patterns elicited by that word is well established in psycholinguistic research using eye-tracking. Yet less evidence or consensus exists regarding whether the same link exists between linguistic complexity measures of a sentence or passage and eye movements registered at the sentence or passage level. This article focuses on "global" measures of syntactic and lexical complexity, i.e., the measures that characterise the structure of the sentence or passage rather than aggregate lexical properties of individual words. We selected several commonly used global complexity measures and tested their predictive power against sentence- and passage-level eye movements in samples of text reading from 13 languages represented in the Multilingual Eye Movement Corpus (MECO). While some syntactic or lexical complexity measures elicited statistically significant effects, they were negligibly small and not of practical relevance for predicting the processing effort either in individual languages or across languages. These findings suggest that the "eye-mind" link known to be valid at the word level may not scale up to larger linguistic units.

Keywords

Eye movements; sentence processing; text reading; syntactic parsing

Received: 16 April 2024; revised: 6 August 2024; accepted: 6 September 2024

The body of eye-movement research on reading has produced vast evidence that eye movements reflect visual and linguistic properties of the words under the gaze (see reviews by Clifton et al., 2007; Rayner, 1998). As codified in the "eye-mind" linking hypothesis by Just and Carpenter (1980, p. 330), "the eye remains fixated on a word as long as the word is being processed." While further work questioned the scope of this hypothesis (see discussion in the article by Rayner, 1998), the link between the cognitive effort of word processing and eye-movement patterns elicited by that word is not in doubt. Three benchmark word-level effects—word length, frequency, and predictability—illustrate this link robustly (e.g., Calvo & Meseguer, 2002; Kliegl et al., 2004). In all written languages considered so far, shorter words, more frequently occurring words, and words that are easier to predict in their linguistic context are processed faster than their longer, less-frequent, or less-predictable counterparts (Kuperman et al., 2024; Rayner, 1998). Yet there is less evidence whether the same link between (1) eye movements and (2) linguistic properties holds "globally," i.e., when both (1) and (2) are defined at the sentence or passage level. This article aims to contribute to this body of knowledge.

Researchers agree that, to process a sentence, readers must parse the sentence's syntactic structure and identify each word's role in this structure. Meanings of individual words and phrases must also be integrated into a unified representation of the sentence meaning and linked to the reader's world knowledge and prior context (e.g., Staub, 2015). At the levels beyond a single sentence, the processes of syntactic parsing and semantic integration are

[1]Department of Linguistics and Languages, McMaster University, Hamilton, Ontario, Canada
[2]Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Corresponding author:
Victor Kuperman, Department of Linguistics and Languages, McMaster University, Togo Salmon Hall 513, 1280 Main Street West, Hamilton, Ontario L8S 4M2, Canada.
Email: vickup@mcmaster.ca

juxtaposed with the need to process the logical and semantic structure of the text and link it to the existing schemas rooted in world knowledge (e.g., Kintsch, 2018). The complexity of the syntactic or semantic structure of a sentence or passage arguably determines the cognitive effort of processing these larger linguistic units, over and above the effort of individual word processing. Existing research has proposed a broad variety of measures of sentence complexity as potential predictors of the sentence-level processing effort (e.g., González-Garduño & Søgaard, 2018; Hollenstein et al., 2022; Nahatame, 2023; Wiechmann et al., 2022). Some of these measures are mathematically derived for sentences from lexical properties defined for individual words in those sentences: e.g., from word length, word and n-gramme frequency, word familiarity, age of acquisition, prevalence, surprisal, transitional probability, semantic ambiguity, and others. Most readability indices, commonly used in relevant studies (discussed in the following sections), fall into the same category because they factor in the lengths and frequencies of individual words in a sentence or passage. The effect of such aggregated measures on the sentence- or passage-processing effort is completely expected given the robust word-level effects of the measures that are being aggregated. For instance, since word length affects word reading times, the mean word length per sentence would also affect sentence-reading time, which is the sum of reading times for words in the sentence. Thus, while demonstrably predictive of sentence- and passage-reading times, the effects of aggregated measures do not directly answer the question of whether the structure or semantics of a sentence or passage as a whole influence sentence-/passage-reading times. Instead, they demonstrate that word-level effects propagate to the sentence level thanks to the incremental word-by-word processing of a sentence.

The present article focuses on the "global" measures of complexity, namely the measures that characterise the entire higher-level linguistic unit (sentence or passage) and represent either the syntactic parse of that unit or its lexical composition. An example of a global measure is the maximum length of the syntactic dependencies in a sentence. This metric is not contingent on the properties of any specific word but instead reflects the sentence structure. Prior work considering global measures, often alongside aggregated word properties, reports mixed findings regarding their predictive role. To give a few examples, Kapteijns and Hintz (2021) report that the effect of syntactic complexity measures on self-paced sentence-reading times was negligible when word-level measures were included in the regression models. Furthermore, syntactic and lexical sentence-level complexity measures ranked relatively low in terms of predictive power, compared to word features and readability indices, in analyses by Wiechmann et al. (2022), which used the ensemble of 107 linguistic features to predict eye movements in text reading. Virtually none of the 14 syntactic features examined in Nahatame's (2023) study of the first- and second-language text reading reached significance as predictors of eye movements, while several word-level features did. Thus, the question of whether the "eye-mind" link between complexity and processing effort holds at the level of the sentence or passage, as it does at the local word level, is far from settled. Our first goal is to examine this question by considering the effects of very commonly used syntactic and semantic global complexity measures on the eye-movement reading record at the levels of sentences and passages.

The second goal of this article is to examine whether the effects that global complexity may have on the sentence- and passage-processing effort vary across languages. The rationale for the cross-linguistic approach is that languages of the world show astounding variability in their morphosyntactic properties, which affect both the syntactic structure of a sentence and its lexical complexity. While measures of complexity can be calculated for every language, it stands to reason that individual effects of those measures may have a different relative contribution to explaining variance in specific languages. Prior research expressed interest in cross-linguistic effects of sentence complexity; see, e.g., a comparison of eye-tracking data in four languages in the studies by Hollenstein et al. (2021) and Wiechmann et al. (2022) and two languages in the study by González-Garduño and Søgaard (2018). Furthermore, Liversedge et al. (2016, 2024) compared sentence-reading times across Chinese, English, and Finnish and discussed whether or not these reading times were statistically identical across the languages; see also the article by Schroeder et al. (2022). Our second goal contributes to this body of research.

Our two focal questions—the strength of the "eye-mind" link beyond the word level and its potential variability across languages—are of interest for several empirical, methodological, and computational areas of research in the literature on reading. One such area, fueled by the recent advances in predictive language models, is the burgeoning investigation of text readability. Production of texts that align in readability, complexity, and accessibility with the language and reading proficiency of the intended audience is a long-standing challenge for educational materials, public messaging, marketing, technical and business writing, and multiple other areas (see the article by DuBay, 2004 for a review). Prior studies in natural language processing have proposed a myriad of quantitative linguistic measures of word, sentence, and passage complexity and developed automatic tools assessing readability of texts for first- and second-language speakers of various languages (e.g., Crossley et al., 2016; Rysová et al., 2016; Sato et al., 2008). Cognitive plausibility of the complexity measures has long been at the forefront of this research area (e.g., Crossley et al., 2008; DuBay, 2004). Yet, it is only recently that the area has recognised

the eye-movement record of the cognitive effort during reading as a source of a useful constraint on the psychological validity of the proposed measures (e.g., Green, 2014; Klerke et al., 2015, and citations mentioned earlier). From this perspective, our cross-linguistic study of the extremely commonly used measures of sentence and passage complexity contributes to the existing literature on text readability by testing whether these measures are predictive of the text-reading behaviour.

A second area of relevance to our study is computational modelling of oculomotor control during reading. Virtually all such models have word recognition at the core of their architectures and lexical properties (e.g., word length, frequency, and predictability in context) as their key parameters (see the review by Reichle, 2021). Yet perhaps the most influential models do not account or only partly account for higher-level properties of the sentence and passage (Vasishth et al., 2013), even though these properties determine the ease of linking a word "into a syntactic structure, generating a context-appropriate semantic representation, and incorporating its meaning into a discourse model" (Reichle et al., 2009, pp. 5–6). The E-Z Reader model has introduced a module for post-lexical processing of the word in one of its iterations (Reichle et al., 2009) responsible for all the cognitive operations in the quotation provided earlier. These cognitive operations, which include syntactic structuring, semantic representation, and discourse model incorporation, are crucial for understanding the cognitive effort involved in reading. Yet the model has not proposed operationalisation of the relevant parameters. The SWIFT model of oculomotor control (Engbert et al., 2005) only accounts for the effects of linguistic units beyond the word level to the extent that they influence the word's predictability in context. Yet the Sentence-Processing and Eye-Movement Activation-Coupled Model (SEAM, Rabe et al., 2024) has successfully integrated the architecture of SWIFT with the sentence-processing model, which partly defines the cognitive effort of reading as the cost of completing linguistic dependencies in the sentence.

In sum, correctly and fully accounting for linguistic complexity beyond the word level is an ongoing challenge for models of eye-movement control. To meet this challenge, it is important to determine whether the global properties of sentences and passages affect the behavioural costs of the incremental word-by-word recognition and parsing. Our study may facilitate this research area by pointing out which existing ways to operationalise such complexity offer viable candidates for the model parameters. The cross-linguistic nature of our study is relevant to another implicit challenge that current computational models are to face: They are being developed for, and tested against, a single language. There is growing evidence that many facets of the reading behaviour are universal across multiple written languages (Kuperman et al., 2024; Li et al., 2014, 2022 Liversedge et al., 2016, 2024; Siegelman et al., 2022). It is an open question whether the existing model architectures can account for the natural variety of written languages and writing systems and what adjustments to their parameters would such extension require. Cross-linguistic data from the analyses below help answer this question.

Finally, our study aims to make a methodological contribution. The rich literature on sentence processing has examined at least some of the complexity measures we consider here (or their mathematical variations) and reported their effects on behavioural measures, e.g., self-paced or eye-tracking reading times. Many relevant studies make use of sentence reading as a task, while the present study is based on passage reading. The format in which sentences are presented—either in isolation without prior context or in a context-rich running text passage—demonstrably influences how the words in those sentences, and the entire sentences, are read. The pioneering study by Radach et al. (2008) presents same sentences in isolation or in context and reports substantial differences in the spatial and temporal eye-movement patterns as well as in the strength-word frequency effects as a function of the presentation format. It stands to reason that discrepancies may also be observed when the effect of the same measure of complexity is examined in sentence reading vs. passage reading. Specifically, we predict that the presence of the passage-level context into which the sentence meaning must be integrated would diminish the effect of the sentence-level complexity measures because this context might disambiguate some of syntactic and semantic uncertainties that the sentence would present if read in isolation. This prediction is partly driven by the analogy with studies that consider word processing in isolation (e.g., lexical decision task) vs. word processing in sentence context (e.g., eye-tracking). Behavioural responses to the stand-alone unit like a word tend to amplify the effect of the properties related to this word, while the effect of the same properties is attenuated for words appearing in context: See the studies by Coskun et al. (2023) and Kuperman et al. (2013) for examples of the word frequency and length effects and morphological priming. Our consideration of linguistic complexity at the sentence and passage levels will push forward the research agenda formulated by Radach et al. (2008) and test whether results obtained from sentence-reading studies align with those obtained in the task that involves passage reading for comprehension.

The present study uses the Multilingual Eye Movements Corpus (MECO; Siegelman et al., 2022), a database that contains text-reading eye-tracking data from 13 typologically and genetically diverse languages, including alphabetic and abjad writing systems. Specifically, we use sentence- and passage-level eye-tracking data as behavioural outcomes of processing effort. We also process texts

Table 1. Sample sizes for passage- and sentence-level analyses by language, as well as self-rated ability for speaking, listening, and reading comprehension.

Language	Language code	N participants	N passages	N sentences	Age (range)	Years of education (*SD*)	Speaking (*SD*)	Listening (*SD*)	Reading (*SD*)
Dutch	Du	45	363	3,618	22.69 (19–30)	16.12 (2.81)	9.47 (0.69)	9.56 (0.62)	9.6 (0.58)
Estonian	Ee	52	462	3,364	21.04 (18–28)	15.76 (1.7)	10 (0)	10 (0)	10 (0)
English	En	46	482	4,189	22.23 (18–30)	14.51 (2.56)	9.31 (0.90)	9.64 (0.56)	9.46 (0.79)
Finnish	Fi	49	534	3,201	24.29 (19–35)	15.04 (2.71)	9.67 (0.59)	9.84 (0.47)	9.82 (0.44)
German	Ge	45	447	4,304	23.76 (18–39)	15.88 (2.75)	9.5 (0.69)	9.59 (0.63)	9.41 (0.72)
Greek	Gr	45	355	2,696	22.84 (18–30)	17.04 (2.5)	9 (0.88)	9.67 (0.6)	9.73 (0.58)
Hebrew	He	47	406	3,615	24.04 (18–29)	12.82 (1.37)	9.68 (0.56)	9.79 (0.41)	9.6 (0.54)
Italian	It	54	491	3,735	22.83 (19–30)	16.72 (2.15)	9.59 (0.71)	9.76 (0.55)	9.76 (0.51)
Korean	Ko	32	238	1,742	21.97 (19–25)	12.98 (2.13)	8.53 (1.5)	8.78 (1.31)	8.69 (1.09)
Norwegian	No	42	359	3,386	25.69 (19–30)	15.33 (3.27)	9.31 (1.7)	9.33 (1.6)	9.21 (1.7)
Russian	Ru	46	447	3,625	24.26 (18–45)	15.45 (2.06)	9.38 (1.41)	9.69 (1.08)	9.46 (1.47)
Spanish	Sp	48	432	3,028	23.04 (18–30)	19.48 (3.8)	9.73 (0.61)	9.73 (0.64)	9.58 (0.79)
Turkish	Tr	29	224	1,924	23.69 (20–29)	17.34 (2.38)	9.41 (0.73)	9.66 (0.61)	9.34 (1.59)

in all languages to evaluate well-established syntactic and lexical complexity measures at the sentence and passage levels. The critical analyses address the two goals of the study by pitting the complexity measures against behavioural outcomes and estimating both the effects of those complexity measures and their interaction with language over and above control variables.

Methods

Eye-tracking data: participants, procedure, and materials

This article analyses eye-tracking text-reading data from the first release of the MECO database (Siegelman et al., 2022). Specifically, we use first-language reading data (MECO L1) from 582 participants, representing 13 languages: Dutch (labelled "du"), English (en), Estonian (ee), Finnish (fi), German (ge), Greek (gr), Hebrew (he), Italian (it), Korean (ko), Norwegian (no), Russian (ru), Spanish (sp), and Turkish (tr). All participants were university students and native speakers of the official language(s) of the country where the testing took place; see the study by Siegelman et al. (2022) for details on participating lab sites. Due to natural cross-linguistic differences, no single test could be administered that would quantify individual or group differences in any aspect of language proficiency. However, Siegelman et al. (2022) administered an identical instrument in all sites, which provides demographic information and subjective self-ratings of speaking, listening comprehension, and reading comprehension abilities of readers in their first language, i.e., the abridged version of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007). Table 1 reports sample sizes from each participating laboratory and language sample, as well as relevant demographic and language use data.

All participants silently read 12 expository texts in their L1 for comprehension and were instructed that they will answer comprehension questions after each text. Five of the texts were translated into each L1 from English, while the remaining seven were originally written in the L1s on the set of topics that were kept consistent across languages (see the article by Siegelman et al., 2022, for comparability of text complexity and translation quality). Each text was followed by four yes/no comprehension questions. Comprehension was generally very high (80% or above, on average). The reading-for-comprehension task that the MECO project (Siegelman et al., 2022) administered is only one of several possible tasks that involve reading. The goal of reading (e.g., comprehension, information verification, skimming, proof-reading, or editing) influences the eye-movement behaviour during reading (e.g., Hyönä & Kaakinen, 2019; Laycock, 1955; Radach et al., 2008). Since alternatives to MECO that would implement tasks other than reading for comprehension are not currently available, the findings and conclusions of this article are constrained to reading for comprehension, which is arguably the most commonly used task in psycholinguistic research of reading. We return to this point in the General Discussion section.

All participating laboratories followed the same experimental procedure. Eye movements were recorded with EyeLink Portable Duo, 1,000 or 1,000+ eye-tracking devices (SR Research, Kanata, Ontario, Canada) with a sampling rate of 1,000 Hz. Chin rests were used to minimise head movements. Each text appeared on a separate screen in a 20- or 22-point mono-spaced font (the visual angle subtended by each character varied by the testing site) with 1.5 spacing (for further details, see the study by Siegelman et al., 2022).

Text processing

All texts used as stimuli in the MECO text-reading task in 13 languages were processed using the UDPipe Natural Language Processing toolkit version 2.0 implemented in the R package udpipe, version 0.8.5 (Straka & Straková, 2017). The syntactic models follow the Universal Dependency schema (UD; Nivre, 2015), version 2.5. This version of the UD schema is available in 90 languages, including all languages in MECO. The package performs tokenisation, parts of speech (POS) tagging, lemmatization, and dependency parsing and associates each word in the input texts with the respective token, lemma, POS, and syntactic annotation, i.e., the variables that enable calculation of syntactic and lexical complexity metrics, defined in the following sections. UDPipe is a probabilistic model based on neural networks. While its accuracy is very high, it is an automatic annotation model and shares the advantages and disadvantages of similar computational tools (https://ufal.mff.cuni.cz/udpipe/2/models).

Dependent variables

At the sentence level, we considered first-pass sentence-reading time (the summed duration of all fixations landing in the sentence region of interest before the gaze left the sentence for the first time, labelled *firstpass*); total sentence-reading time (the summed duration of all fixations landing in the sentence region of interest, *total*); reading rate (words per minute, *rate*); and regression index (the binary index of whether there were regressions into the sentence, *reg*). At the passage level, we considered total passage-reading time (*total*), reading rate (*rate*), skip rate (proportion of skipped words out of total, *skip*), and regression rate (proportion of regressive saccades, *reg*).

Independent variables

The existing literature has introduced hundreds of metrics for sentence complexity (see examples in the Introduction). We only consider nine measures commonly found in computational-linguistic and psycholinguistic analyses of texts (e.g., Graesser et al., 2011; Jagaiah et al., 2020; Kyle, 2019). The description of those measures in the following section closely follows the Supplementary materials S1 in the study by Kyröläinen et al. (2023), providing a more detailed motivation and additional references.

One set of variables related to lexical and structural richness of the text relies on type-token ratio (TTR). The original metric, proposed by Johnson (1944), is calculated as the number of word types (unique words) in a text divided by the number of word tokens (all words, unique or repeated) in a text (types word/tokens word). We used this metric based on words (labelled *ttr-w*) and added two more such metrics: one based on POS associated with a given word (types POS/tokens POS, labelled *ttr-p*), and finally, one based on the dependency relation associated with a particular word (types dep/tokens dep, *ttr-d*). Higher values on any of TTR metrics correspond to richer, more diverse use of words, POS, or dependency types. TTR measures were computed for each sentence and, separately, for each passage. We further included the noun-verb ratio: token_nouns/(token_nouns + token_verbs, *nv-ratio*). A higher ratio is found in more elaborate, lexically richer texts. This metric was calculated for each sentence and each passage.

Another set of variables described syntactic complexity. The first variable, *d-ratio*, captures how elaborate the syntactic constructions were in a given sentence, i.e., it calculates the ratio of heads to nodes in a sentence (e.g., Gibson, 1998, 2000). D-ratio thus reflects the presence of adjunct structures and higher valenced verbs. A higher d-ratio value indicates that a sentence contains more content at a given level in the structure, making the sentence syntactically more elaborate. D-ratio was computed for each sentence, and each passage was associated with the mean of d-ratio values of sentences in that passage. The second variable in this set operationalised syntactic complexity in a given sentence by calculating the longest path in a dependency tree for the sentence: longer dependency relations should be more effortful to process (Gibson, 1998, 2000). Each dependency tree was treated as a directed acyclic graph (Oya, 2011; Yadav et al., 2019), and the longest path was calculated using the diameter function in the R package igraph, version 1.2.6, i.e., the length of the longest path (maxu,vd[u, v]) between any two nodes (u, v), where d(u, v) is a distance (Csardi & Nepusz, 2006) and labelled *depth*. Each sentence was associated with the maximum dependency length, and each passage with the mean dependency length across all sentences in the passage. The third and final variable in this set considered the notion of syntactic complexity at the level of the narrative by focusing on the use of complex clauses—such as those involving greater embedding, e.g., via coordination and subordination—relative to syntactically simplex ones (e.g., Beaman, 1984; Givón, 1991). The presence of one of the following nine dependency relations (defined using UD) was used to mark a sentence as complex, otherwise a sentence was considered as simplex: parataxis, xcomp, ccomp, advcl, acl:relcl, acl, conj, cc, mark. This measure is not defined at the sentence level and is only used in the passage-level analyses. For each passage, the number of simplex and complex sentences was calculated, and the sums were divided to compute a ratio (1 − n_simplex/n_complex) where higher values indicated that a given story had relatively more complex sentences. We refer to this variable as embeddedness.

Three more simple metrics of sentence and passage complexity were considered: the number of tokens (in a sentence or passage, *n_tokens*), sentence number in the passage (defined for sentences only, *sentnum*), and the mean length of utterance (Brown, 1973, *mlu*) defined for passages only as the average sentence length in word tokens. The final independent variable was language ID, a categorical variable with 13 levels representing languages included in MECO.

Statistical considerations

We used random forest modelling to identify the relative importance of predictors. As described in the article by Matsuki et al. (2016), this method is a generalisation of the decision tree method, in which the multivariate data undergo recursive binary split according to the value of one of the predictor variables, such that the observations within a partition reach higher homogeneity. Random forests build multiple decision trees using a random sample of observations for each tree and (at each split point) random samples of predictors. The resulting forest of those trees provides fitted values, which are more accurate than those of any single tree (Breiman, 2001). Variable importance is assessed by randomly permuting the values of one predictor across all trees and estimating the loss in prediction accuracy of the forest: Little loss implies low importance. The variable importance metric obtained from the random forest model is contingent on the scale of the dependent variables and other parameters. Thus, estimates of variable importance can only be interpreted as comparative rather than absolute values (Strobl et al., 2009). We used random forests and library *party* (Hothorn & Zeileis, 2015) version 1.3-11 to determine the relative importance of predictors. Following the guidelines of Matsuki et al. (2016) for random forest modelling, parameter *mtry*, which defines the number of randomly sampled predictor variables that are used to select each split point in a tree, can occupy a range of values from square root of the number of predictors to one-third of the number of predictors. In the present analyses, the range of *mtry* is represented by a single value of 3. Another parameter *ntree* defines the number of trees to be built: it was set to 500.

Statistical software environment R version 4.2.2 was used for data processing and analyses. Library lmerTest 3.1-3 (Kuznetsova et al., 2017) was used to fit (generalised) generalised linear mixed-effects regression models with either the Gaussian or binomial distribution family, contingent on the type of dependent variable. All predictors were normalised, such that their effect sizes (ES) were given on the same scale, per one unit of the predictor's standard deviation. Passage ID, language, and participant ID were included as random intercepts in every model. A further inclusion of random slopes led to convergence issues, and the slopes were removed from the models. After fitting each model with the Gaussian family, we identified outliers as datapoints with absolute scaled residuals over 2.5 standard deviations. These outliers were removed, and the model was refitted (Baayen & Milin, 2010).

Availability

Data and code for language models are found at: https://github.com/dalmobuzato/meco-complexity-ud

Results and discussion

This section reports findings derived from the passage-level analyses of eye-tracking data, followed by the analyses of sentence-reading data. Data-trimming procedures are described in the study by Siegelman et al. (2022), and the original sentence- and passage-level eye-tracking data reanalyzed are available at the MECO repository https://osf.io/3527a/. In addition, we removed data points from the sentence dataset that represented unrealistically fast and slow reading rates (the top 1% and the bottom 1% of the reading rate distribution). The following punctuation marks were markers of sentence boundaries in the eye-tracking data and during text processing: Full stop, semicolon, exclamation, and question mark. Twelve out of 156 unique sentences (7.7%) across all languages were misaligned and removed from further consideration. Table 1 reports the resulting sample sizes for all languages considered in the article. Descriptive statistics for all dependent and independent variables (mean, *SD*, min and max values), represented by language, are reported in the online Supplementary materials at https://github.com/dalmobuzato/meco-complexity-ud.

Sentence-level analyses

Table 2 presents the correlation matrix for sentence-level data. This matrix indicates strong correlations between several metrics of syntactic and lexical complexity ($|r| > 0.7$), especially correlations between the sentence length in word tokens and TTR values in both sentence and passage data. Including many correlated predictors into regression models may lead to collinearity and loss of precision in parameter estimation (see the study by Dormann et al., 2013 for a review). For this reason, before fitting regression models, we took an additional step to reduce the dimensionality of the data and the number of predictors. The correlation matrix for passage-level data is provided in the online Supplemental Materials at https://github.com/dalmobuzato/meco-complexity-ud.

In this step, all relevant independent variables were entered into the random forest model as predictors of each of the sentence-reading eye movements, see the Methods section for a brief description of the method and parameter estimation. The output of each model is the estimated

Table 2. Correlation matrix for sentence-level data.

	firstpass	Total	rate	reg	sentnum	d_ratio	depth	n_tokens	nv_ratio	ttr_w	ttr_d	ttr_p
firstpass	NA	0.87	−0.26	0.04	−0.06	0.22	0.43	0.66	0.09	−0.32	−0.56	−0.51
total	0	NA	−0.27	0.22	−0.13	0.25	0.52	0.8	0.07	−0.45	−0.67	−0.6
rate	0	0	NA	−0.25	0.26	0.09	0.15	0.18	−0.14	−0.15	−0.16	−0.17
reg	0.162	0	0	NA	−0.6	0.02	0.08	0.07	0.02	−0.20	−0.10	−0.10
sentnum	0.034	0	0	0	NA	0.03	−0.01	0.04	0	0.03	0.01	0.02
d_ratio	0	0	0.001	0.432	0.338	NA	−0.23	0.37	0.14	−0.34	−0.25	0.00
depth	0	0	0	0.003	0.742	0	NA	0.59	−0.12	−0.27	−0.47	−0.59
n_tokens	0	0	0	0.009	0.112	0	0	NA	0.07	−0.60	−0.74	−0.68
nv_ratio	0.001	0.011	0	0.467	0.941	0	0	0.011	NA	−0.13	−0.16	−0.11
ttr_w	0	0	0	0	0.244	0	0	0	0	NA	0.53	0.41
ttr_d	0	0	0	0	0.839	0	0	0	0	0	NA	0.8
ttr_p	0	0	0	0	0.521	0.918	0	0	0	0	0	NA

Pearson's correlation coefficients are reported above the diagonal, and *p*-values below. The *p*-value $< .001$ is reported as "0."

relative importance of predictors of the given dependent variable. Figure 1 visualises results of random forest analyses for sentence data.

Figure 1 illustrates the dominance of simple control variables across all eye-movement measures: the most important predictor for reading times was sentence length (n_tokens), and that for reading rate and regression likelihood was sentence position in the passage (sentnum). These findings are intuitive, as further evidenced by the regression models reported in the following sections. Reading times for longer sentences are longer. In addition, readers speed up when they move further into a longer text (e.g., Kuperman et al., 2018), which explains the relative importance of sentence position in the passage. Finally, it stands to reason that sentences at the beginning of the passage are more frequently targeted by regressions than those towards the end of the passage, hence the effect of sentence position on regression likelihood. All other predictors, tapping into facets of syntactic and lexical complexity, were estimated as much less important. For comparability, we extract three most important predictors from each random forest models. These predictors were introduced into the regression model fitted to respective eye-movement measures.

Table 3 reports the outputs of generalised mixed-effects models for first-pass sentence-reading time, total sentence-reading time, reading rate, and regression likelihood. Since the analysis of sentence-reading data is overpowered (with $N > 40{,}000$), many of the selected predictors reached the nominal 5% threshold of statistical significance. Yet our main interest lies in the practical relevance of those predictors rather than their statistical significance (for elaboration of this distinction, see, e.g., the study by Gelman & Stern, 2006). To evaluate the practical relevance of each effect, we used the estimated beta coefficients of respective models; see reports of ES in Table 4. Since all predictors were standardised (z-transformed), the standardised beta coefficients reflect the amount of change in the continuous dependent variable as a function of a difference in one unit of standard deviation in the predictor. For instance, a difference of 1 *SD* of the number of tokens comes with an estimated increase in 1,105 ms in first-pass sentence-reading time, see the first column of Table 3. Since the standard deviation of first-pass sentence-reading time is 2,781 ms, the difference of 1 *SD* in the number of tokens corresponds to a change in (1,105/2,781 =) 0.40 units of *SD* of first-pass reading time.

To further illustrate the meaning of this ES, we make an informal analogy of the regression coefficients to another statistical test for which conventional thresholds exist that define a small, medium, or large effect. If two groups of stimuli were selected that differed from each other in 1 *SD* of the predictor variable (e.g., shorter and longer sentences) and the difference in reading times between those two groups were 0.4 *SD*, as is our finding above, this ES would correspond to Cohen's *d* of 0.4. This is the ES that is both most commonly found in psychological research (Camerer et al., 2018; Open Science Collaboration, 2015) and the one that signals a minimum practical relevance (Brysbaert, 2019). ES estimated for the two remaining predictors of first-pass reading time—TTR based on syntactic dependencies and POS (ttr-d and ttr-p)—are negligibly small, accounting for a change in 0.08 *SD* and 0.01 *SD* in the dependent variable, respectively. We make use of this informal analogy of the standardised regression coefficients and Cohen's *d* throughout the article for illustrative purposes only.

The pattern was similar for total sentence-reading time. While all three predictors were estimated as statistically significant (at the 5% threshold), only sentence length in word tokens elicited a substantially medium-size effect on this eye-movement measure (an increase in 0.48 *SD* per 1 *SD* difference in the predictor). Decreases in total sentence-reading time associated with an increase in ttr-p and

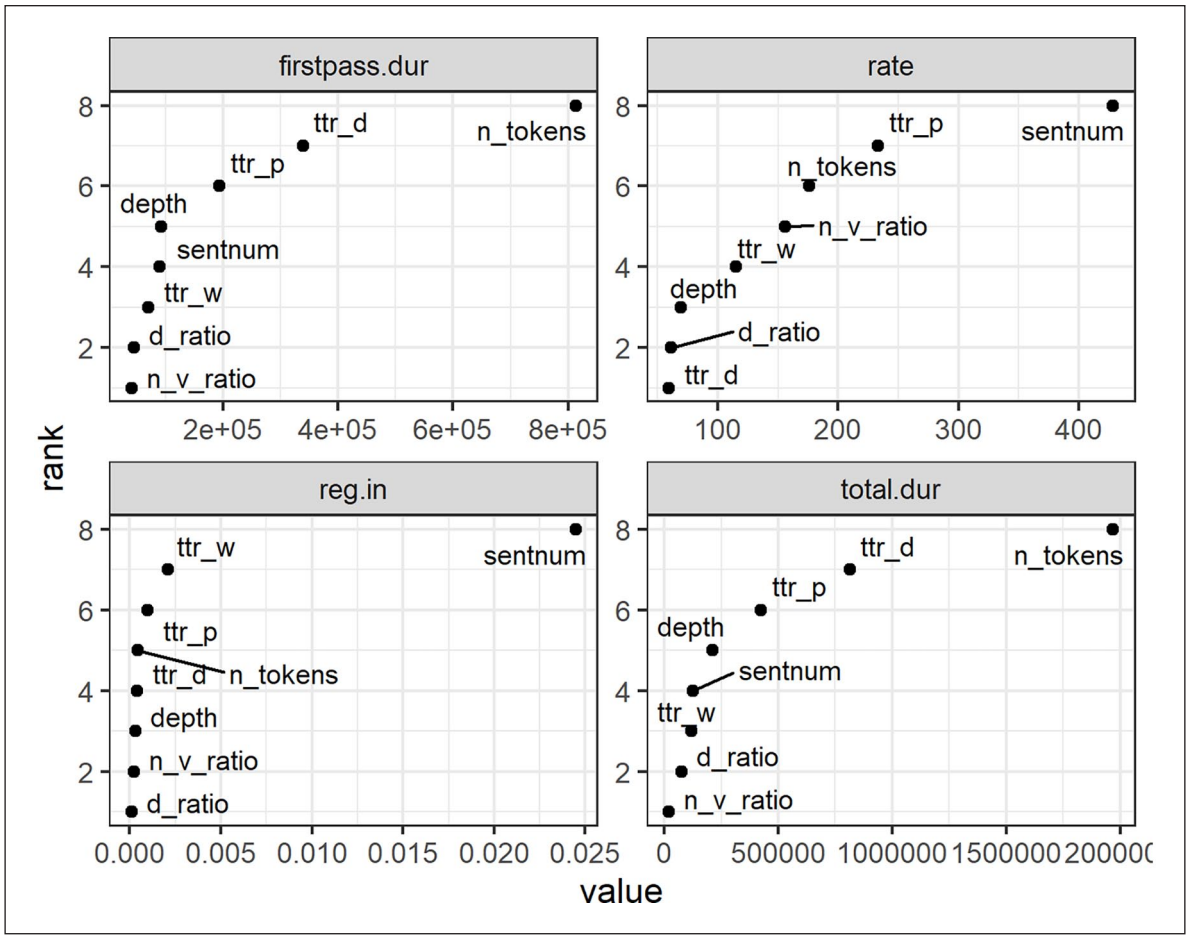

Figure 1. Relative importance (on the x-axis) of predictors for first-pass sentence-reading time, reading rate, regression likelihood, and total sentence-reading time.

ttr-d were extremely small in size (0.09 *SD* and 0.01 *SD*, respectively), below Cohen's *d* of 0.20, conventionally associated with a small-size effect (Cohen, 1992).

Among predictors of reading rate, sentence position in passage (sentnum) elicited the strongest effect out of all selected predictors: Sentences further into the passage were read at a higher rate (Kuperman et al., 2018). Yet even this effect (0.13 *SD*) was weaker than the conventional threshold for small-size effects ($d = 0.2$). While the regression model indicated that reading rate is significantly higher in longer sentences and in sentences with a higher type-token POS ratio, these effects were negligibly small.

Since standard deviation is not meaningful for binary variable like regression likelihood, we could not estimate the ES of its predictors. Still, since all predictors are standardised and thus can be interpreted on the same scale, we note that the negative effect of sentence position in passage on regression likelihood is (−0.689/−0.054=) 12.7 times stronger than the second strongest effect of word-based TTR, see the last column of Table 3.

In sum, analyses of sentence data indicated the dominance of simple control variables: sentence length and sentence position in the passage. No metric of syntactic complexity fell into the top 3 most-important predictors indicated by the random forest models. While measures of lexical complexity such as TTR did, their effects on eye movements defined at the sentence level were minimal. This pattern repeated itself in the analyses of sentence-reading data conducted in specific languages (not shown). In each language sample, control variables produced the most substantial effects in the small-to-medium range, while other predictors showed negligibly minor effects.

Passage-level data

Table 4 summarises the correlation matrix for dependent and independent variables in the passage-level data. As with the sentence analysis mentioned earlier, some of the predictors show strong correlations, indicating potentially dangerous collinearity.

Table 3. Outputs of generalised mixed-effects models for first-pass sentence-reading time, total sentence-reading time, reading rate, and regression likelihood (binomial).

	firstpass	Total	rate	reg
Intercept	3755.651* [3319.310; 4191.991]	5247.318* [4841.076; 5653.561]	236.889* [218.548; 255.231]	−1.001* [−1.240; −0.762]
ttr_p	−214.106* [−252.836; −175.376]; ES = 0.08	−281.317* [−316.994; −245.640]; ES = 0.09	2.602* [1.363; 3.840]; ES = 0.02	−0.044* [−0.073; −0.014]
ttr_d	24.436 [−14.792; 63.663]; ES = 0.01	−40.480* [−76.675; −4.284]; ES = 0.01		
n_tokens	1104.730* [1069.353; 1140.107]; ES = 0.40	1554.544* [1522.436; 1586.653]; ES = 0.48	9.876* [8.597; 11.156]; ES = 0.09	
sentnum			15.041* [14.267; 15.815]; ES = 0.13	−0.689* [−0.714; −0.664]
ttr_w				−0.054* [−0.083; −0.024]
AIC	746204.707	739954.440	478958.204	45767.022
BIC	746273.748	740023.482	479027.237	45827.611
Log Likelihood	−373094.354	−369969.220	−239471.102	−22876.511
Num. obs.	41363	41367	41320	42427
Num. groups: subject	582	582	582	582
Num. groups: language	13	13	13	13
Num. groups: passage	12	12	12	12
Var: subject (Intercept)	732320.155	1471991.275	2810.123	0.434
Var: language (Intercept)	602408.448	469955.206	883.503	0.154
Var: passage (Intercept)	21771.554	49352.692	173.656	0.025
Var: Residual	3852050.221	3268276.716	6016.829	

The 95% confidence intervals are reported in square brackets: Asterisks indicate that the null hypothesis value is outside the confidence interval (p-value $< .05$). ES = effect size.

Table 4. Correlation matrix for passage-level data.

	total	reg	rate	skip	d_ratio	depth	embed	mlu	n_tokens	ttr_d	ttr_p	ttr_w	nv_ratio
total	NA	0.39	−0.56	0.01	0.31	−0.02	0	0.22	0.48	−0.39	−0.29	−0.38	0.17
reg	0	NA	−0.18	0.49	0.04	−0.08	0.01	0.12	0.13	−0.22	−0.26	−0.2	−0.14
rate	0	0.022	NA	0.58	−0.1	0.27	−0.16	0.24	0.19	−0.29	−0.38	−0.2	−0.3
skip	0.939	0	0	NA	0.04	0.22	−0.14	0.37	0.42	−0.51	−0.5	−0.51	−0.21
d_ratio	0	0.616	0.227	0.612	NA	−0.4	0.13	−0.43	0.36	−0.21	−0.15	−0.29	0.25
depth	0.816	0.339	0.001	0.005	0	NA	−0.26	0.44	0.28	−0.47	−0.5	−0.09	−0.32
embed	0.974	0.945	0.051	0.077	0.108	0.001	NA	−0.21	−0.08	0.11	0.2	0.08	0.01
mlu	0.005	0.126	0.003	0	0	0	0.007	NA	0.54	−0.4	−0.46	−0.56	−0.06
n_tokens	0	0.106	0.017	0	0	0	0.346	0	NA	−0.73	−0.76	−0.83	0.12
ttr_d	0	0.006	0	0	0.008	0	0.181	0	0	NA	0.77	0.61	−0.03
ttr_p	0	0.001	0	0	0.068	0	0.012	0	0	0	NA	0.54	0.18
ttr_w	0	0.012	0.014	0	0	0.255	0.332	0	0	0	0	NA	−0.23
nv_ratio	0.031	0.081	0	0.007	0.002	0	0.883	0.458	0.135	0.708	0.027	0.003	NA

Pearson's correlation coefficients are reported above the diagonal, and *p*-values below. The *p*-value $<.001$ is reported as "0."

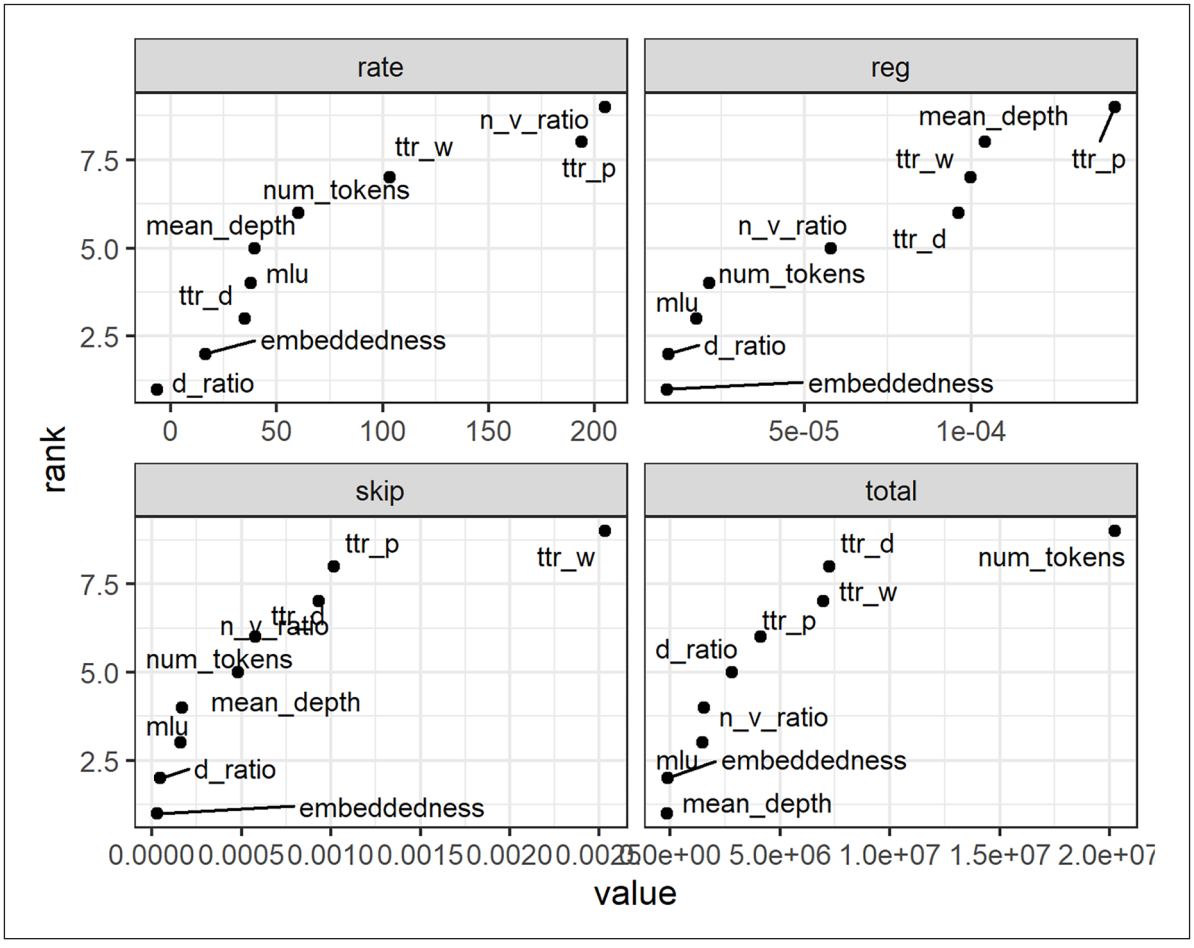

Figure 2. Relative importance (on the x-axis) of predictors for total passage-reading time, reading rate, regression rate, and skipping rate.

The random forest analyses aimed to reduce the dimensionality of the data and identify the most important predictors of each of the eye-movement measures. Figure 2 visualises the results.

As with prior analyses, we selected three predictors with the highest relative importance from each random forest analysis to be included in respective regression models. Table 5 presents outputs of those regression

Table 5. Outputs of generalised mixed-effects models for first-pass sentence-reading time, total sentence-reading time, reading rate, and regression likelihood (binomial).

	total	rate	reg	skip
Intercept	47933.510* [44023.129; 51843.892]	222.454* [204.227; 240.681]	0.245* [0.223; 0.266]	0.440* [0.396; 0.485]
ttr_w	3551.256* [2787.834; 4314.679]; ES = 0.20	−12.663* [−15.913; −9.413]; ES = 0.17	−0.001 [−0.006; 0.003]; ES = 0.01	−0.014* [−0.023; −0.004]; ES = 0.08
ttr_d	−62.915 [−688.636; 562.806]; ES = 0.00			0.005 [−0.003; 0.013]; ES = 0.03
num_tokens	7129.540* [6477.091; 7781.988]; ES = 0.41			
ttr_p		−3.142* [−5.094; −1.191]; ES = 0.04	0.004* [0.001; 0.007]; ES = 0.04	−0.002 [−0.008; 0.005]; ES = 0.01
n_v_ratio		0.918 [−0.479; 2.316]; ES = 0.01		
mean_depth			−0.002 [−0.005; 0.000]; ES = 0.02	
AIC	106512.928	51091.240	−15766.810	−9233.383
BIC	106565.237	51143.555	−15714.433	−9181.100
Log Likelihood	−53248.464	−25537.620	7891.405	4624.692
Num. obs.	5,108	5,112	5,152	5,092
Num. groups: uniform_id	580	580	580	580
Num. groups: lang	13	13	13	13
Num. groups: trialid	12	12	12	12
Var: uniform_id (Intercept)	175269597.571	2966.952	0.006	0.009
Var: lang (Intercept)	37824997.719	875.419	0.001	0.006
Var: trialid (Intercept)	8998179.951	164.302	0.000	0.001
Var: Residual	44250786.571	856.760	0.002	0.007

The 95% confidence intervals are reported in square brackets: Asterisks indicate that the null hypothesis value is outside the confidence interval. ES = effect size.

models, including estimates of the ES implemented as described earlier.

Passage length in word tokens showed a medium-size effect on total passage-reading time (0.41 *SD* of change per 1 *SD* of predictor variable): Longer passages take more time to read. The word-based estimate of TTR (ttr-w) also has an effect that exceeds the small-size threshold (0.2 *SD*). However, a further inspection reveals that the sign of the estimated beta coefficient for ttr-w (3551.256) has the opposite sign relative to the sign of the zero-order Pearson correlation between ttr-w and total passage-reading time ($r = -0.20$). Such reversal of the sign of the effect is sometimes observed in multiple regression models with highly collinear predictors and results from statistical suppression (Friedman & Wall, 2005; Nathans et al., 2012). Statistical suppression is a phenomenon observed in models that contain a strong predictor of the dependent variable and another predictor—which is strongly correlated with the first predictor and weakly with the dependent variable: The weaker predictor may flip the sign, and its estimate is unreliable. The proposed interpretation for suppression is that the effect of a weaker predictor does not hold once the stronger predictor is accounted for (Tabachnick et al., 2013; Wurm & Fisicaro, 2014). In the present data, both conditions for suppression hold, i.e., passage length in tokens and word-based TTR are very strongly negatively correlated ($r = -0.83$), and passage length is strongly positively correlated with total passage-reading time ($r = 0.48$). Even though the effect of the TTR appears to be relatively strong and significant at the nominal level, this effect is an artefact of suppression by passage length and is unreliable.

In all remaining analyses—with reading rate, regression rate, and skipping rate as dependent variables—all selected predictors elicited negligibly small effects. We conclude that passage-level data are less sensitive to both control variables and syntactic and lexical complexity metrics than sentence-level data. As with sentence-based analyses mentioned earlier, only the simple control predictor of passage length in word tokens was observed in all language-specific samples (s observed in all language-specific samples not shown).

General discussion

The literature examining eye movements as predictors of sentence and text processing is vast (see, among many others, the studies by Clifton et al., 2007; Mézière et al., 2023; Rayner, 1998; Staub & Rayner, 2007; Vasishth et al., 2013). While the link between linguistic complexity and processing effort (gauged via eye movements) is firmly established at the word level, the literature shows mixed results in the studies of larger linguistic units, see examples in the Introduction. One goal of this article was to examine a selection of nine commonly used metrics of syntactic and lexical complexity and establish whether eye movements recorded at the sentence and passage levels reflect this complexity. Our focus was on those metrics that are defined over the entire sentence/passage (e.g., the longest path in the syntactic dependency tree) rather than aggregate characteristics of individual words that constitute that sentence/passage (e.g., mean word length). The predictive power of the latter variables is not surprising as it is a direct consequence of word-level effects and incrementality in the processing of large linguistic units (e.g., Levy, 2008). To compute complexity metrics of the former kind, we produced morphosyntactic parses of text data in 13 languages represented in the MECO database of eye-tracking data. Those metrics were pitted against a variety of eye-movement measures defined at the sentence or passage level.

Analyses at the sentence level demonstrated that simple control variables—sentence length in tokens and sentence position in the passage—are by far the most important predictors of relevant eye movements. Regression models further revealed that only these measures elicited small- or medium-size effects on eye movements. These effects are theoretically trivial and not directly reflective of the sentence structure. Furthermore, while occasionally reaching statistical significance in the overpowered dataset, all specialised estimates of syntactic and lexical complexity produced negligibly small effects of no practical relevance.

The outcomes of analyses at the passage level were similar. Only one of the eye-movement measures (total passage-reading time) revealed sensitivity to any of the predictors considered here: Passage length in word tokens elicited a medium-sized positive effect. No other predictor elicited an effect that would qualify as small on any other eye-movement measure.

Our second goal was to determine whether cross-linguistic variability exists in the relative contributions of complexity metrics on eye-movement measures. Yet analyses in all languages replicated the global patterns reported in the previous two paragraphs: Only sentence/passage length and sentence position in the passage produced tangible effects (in the small-to-medium range) on the eye movements.

In sum, our null findings are consistent with the view that sentence or passage structure or lexical composition does not affect eye-movement patterns over and above (aggregated) properties of the individual words that make up those sentences or passages. The observed null effects are relevant for several strands of reading research. First, they are compatible with the view that the eye-mind link hypothesis does not hold beyond the word level. While words presenting with greater "local" linguistic complexity take more effort to process, sentences and passages presenting with greater "global" complexity do not. That is, the processing effort of larger linguistic units can be fully ascribed to the aggregated complexity of the words that make up those units, but not to the syntactic or semantic

complexity of those units per se. For instance, sentence-reading times are affected by the number of words in that sentence but not by the depth of the syntactic tree representing the sentence. This finding, confirmed across 13 languages, supports the strictly incremental accounts of sentence processing (e.g., Levy, 2008).

Another implication of our results is in the constraint that it can posit on existing models and practices. For instance, readability research has proposed a large variety of global computational measures of text complexity, along with the measures that aggregate local word complexity. Our study does not find evidence for psychological validity of those global measures in a set of 13 languages, as none of them affected the sentence- or passage-processing effort to any noticeable degree. Since the eye-movement record is a behavioural reflection of text complexity, these findings suggest that some of the most commonly used complexity measures are suboptimal for characterising the cognitive effort of text reading for comprehension. Similarly, global measures of syntactic or semantic complexity can be ruled out as viable candidates for the computational models that seek to incorporate higher-level cognitive processing (driven by syntactic or discourse properties), in any of the languages considered in this study.

A final, methodological implication is the apparent discrepancy between the effects that selected global measures of complexity produce in passage reading vs. sentence reading. Unlike the study by Radach et al. (2008), this study did not directly compare same sentences presented in isolation vs. the passage-long context. Given prior reports on significant effects that some of the syntactic variables have on eye movements in sentence reading, we speculate that the null effects uncovered in this study are partly due to the difference in presentation format. Just like how words embedded in sentences elicit weaker effects of lexical properties compared to the same words presented in isolation (see references cited earlier), sentences embedded in text passages also attenuate the effects of syntactic and semantic global properties. We believe the cause of attenuation to the same. The available context reduces uncertainty about the sentence being read, informs expectations about upcoming words and structures, and presents an additional task of integrating the sentence meaning into the large logical and semantic discourse structure (Kintsch, 2018; Levy, 2008).

Limitations and future directions

Naturally, the present findings only relate to the commonly used but small selection of syntactic and lexical complexity metrics considered in this article. Many analyses use ensemble approaches that combine hundreds of metrics (e.g., Wiechmann et al., 2022), and new metrics are being regularly proposed (e.g., Sun & Wang, 2024). In addition, texts in the MECO database do not include a broad variety of the relatively rare syntactic constructions that become a frequent subject of the sentence-processing research, e.g., subject- vs. object-relative clauses, cleft sentences, or sentences with lexical or syntactic ambiguity (Demberg & Keller, 2008).

Similarly, the present analyses only represent one task—reading for comprehension—and thus the observed null effects do not necessarily hold for other tasks involving reading. For instance, if text editing for clarity or cohesion rather than comprehension is the reader's goal, it is possible that lexical and syntactic complexity of sentences and passages would predict the reader's behaviour more strongly. Expansion of the present study onto tasks other than reading for comprehension is clearly desirable but is contingent on the future availability of multilingual databases of eye-movement behaviour recorded when these tasks are performed.

This study does not account for individual differences among readers, even though it is a known source of variance (Kuperman et al., 2018). Due to cross-linguistic differences and differences in resource availability, coming up with a single instrument applicable to all 13 languages of MECO did not prove possible in the study by Siegelman et al. (2022). Instead, we used subjective ratings of speaking, listening, and reading proficiency across sites. Consideration of objectively measured component skills of reading may influence the present findings: We relegate this to future research.

Yet we believe that the present results put a noteworthy constraint on the research enterprise that aims to link the effort of processing a sentence or a passage with linguistic properties of those respective units. Across 13 languages, the eye-movement record did not corroborate that link for the sentence- and passage-reading data and complexity metrics in a practically relevant way.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the Social Sciences and Humanities Research Council of Canada Partnered Research Training Grant, 895-2016-1008, (Libben, PI), Insight Grant 435-2021-0657 (Kuperman, PI), the Canada Research Chair (Tier 2; Kuperman, PI), and Grant #316036/2021-8 from the National Council for Scientific and Technological Development—CNPq (Rothe-Neves).

ORCID iD

Victor Kuperman https://orcid.org/0000-0001-8961-3767

References

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 45–80). Ablex.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brown, R. (1973). Development of the first language in the human species. *American Psychologist*, *28*(2), 97–106. https://doi.org/10.1037/h0034209

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16.

Calvo, M. G., & Meseguer, E. (2002). Eye movements and processing stages in reading: Relative contribution of visual, lexical, and contextual factors. *Spanish Journal of Psychology*, *5*, 66–77.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., . . .Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour 2*, 637–644. https://doi.org/10.1038/s41562-018-0399-z

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–372). Elsevier.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Coskun, M., Kuperman, V., & Rueckl, J. (2023). Long-lag repetition priming in natural text reading: No evidence for morphological effects. *The Mental Lexicon*, *18*(1), 1–40.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, *42*(3), 475–493.

Crossley, S. A., Kyle, K., & McNamara, D. S. (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.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5), 1–9.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., . . .Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46.

DuBay, W. (2004). *The principles of readability*. Impact Information.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, *59*(2), 127–136.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126). The MIT Press.

Givón, T. (1991). Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, *15*(2), 335–370.

Gonzalez-Garduno, A., & Søgaard, A. (2018, April). Learning to predict readability using eye-movement data from natives and learners. *Proceedings of the AAAI Conference on Artificial Intelligence 32*(1).

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Green, M. J. (2014). An eye-tracking evaluation of some parser complexity metrics. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)* (p. 3846). Association for Computational Linguistics.

Hollenstein, N., Gonzalez-Dios, I., Beinborn, L., & Jäger, L. (2022, November). Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (pp. 1–15). Association for Computational Linguistics.

Hollenstein, N., Pirovano, F., Zhang, C., Jäger, L., & Beinborn, L. (2021). Multilingual language models predict human reading behavior. *arXiv preprint arXiv:2104.05433*.

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*(1), 3905–3909.

Hyönä, J., & Kaakinen, J. K. (2019). Eye movements during reading. In C. Klein & U. Ettinger (Eds.), *Eye movement research: An introduction to its scientific foundations and applications* (pp. 239–274). Springer.

Jagaiah, T., Olinghouse, N. G., & Kearns, D. M. (2020). Syntactic complexity measures: Variation by genre, grade-level, students' writing abilities, and writing quality. *Reading and Writing*, *33*, 2577–2638. https://doi.org/10.1007/s11145-020-10057-x

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, *56*, 1–15.

Just, M. A., & Carpenter, P. A. (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

Kapteijns, B., & Hintz, F. (2021). Comparing predictors of sentence self-paced reading times: Syntactic complexity versus transitional probability metrics. *PLOS ONE*, *16*(7), Article e0254546. https://doi.org/10.1371/journal.pone.0254546

Kintsch, W. (2018). Revisiting the construction—integration model of text comprehension and its Implications for Instruction. In D. E. Alvermann, N. J. Unrau, M. Sailors &

R. B. Ruddell (Eds.), *Theoretical models and processes of literacy* (pp. 178–203). Routledge.

Klerke, S., Castilho, S., Barrett, M., & Søgaard, A. (2015). Reading metrics for estimating task efficiency with mt output. In *Conference on Empirical Methods in Natural Language Processing* (pp. 6–13). Association for Computational Linguistics.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1–2), 262–284.

Kuperman, V., Drieghe, D., Keuleers, E., & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *Quarterly Journal of Experimental Psychology*, *66*(3), 563–580.

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

Kuperman, V., Schroeder, S., & Gnetov, D. (2024). Word length and frequency effects on text reading are highly similar in 12 alphabetic languages. *Journal of Memory and Language*, *135*, 104497.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454–475). Routledge.

Kyröläinen, A. J., Gillett, J., Karabin, M., Sonnadara, R., & Kuperman, V. (2023). Cognitive and social well-being in older adulthood: The CoSoWELL corpus of written life stories. *Behavior Research Methods*, *55*(6), 2885–2909.

Laycock, F. (1955). Significant characteristics of college students with varying flexibility in reading rate: I. Eye movements in reading prose. *Journal of Experimental Education*, *23*, 311–319.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Li, X., Bicknell, K., Liu, P., Wei, W., & Rayner, K. (2014). Reading is fundamentally similar across disparate writing systems: A systematic characterization of how words and characters influence eye movements in Chinese reading. *Journal of Experimental Psychology: General*, *143*(2), 895.

Li, X., Huang, L., Yao, P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology*, *1*(3), 133–144.

Liversedge, S. P., Drieghe, D., Li, X., Yan, G., Bai, X., & Hyönä, J. (2016). Universality in eye movements and reading: A trilingual investigation. *Cognition*, *147*, 1–20.

Liversedge, S. P., Olkoniemi, H., Zang, C., Li, X., Yan, G., Bai, X., & Hyönä, J. (2024). Universality in eye movements and reading: A replication with increased power. *Cognition*, *242*, 105636.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP–Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*, 940–967.

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, *20*(1), 20–33.

Mézière, D. C., Yu, L., Reichle, E. D., Von Der Malsburg, T., & McArthur, G. (2023). Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, *58*(3), 425–449.

Nahatame, S. (2023). Predicting processing effort during L1 and L2 reading: The relationship between text linguistic features and eye movements. *Bilingualism: Language and Cognition*, *26*(4), 724–737.

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, *17*(9), n9.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 3–16). Springer.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics* (pp. 313–316). Pan-Pacific Association of Applied Linguistics.

Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2024). SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading. *Journal of Memory and Language*, *135*, 104496.

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological Research*, *72*, 675–688.

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

Reichle, E. D. (2021). *Computational models of reading: A handbook*. Oxford University Press.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*, 1–21. https://doi.org/10.3758/PBR.16.1.1

Rysová, K., Rysová, M., & Mírovský, J. (2016, October). Automatic evaluation of surface coherence in L2 texts in Czech. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016)* (pp. 214–228). The Association for Computational Linguistics and Chinese Language Processing

Sato, S., Matsuyoshi, S., & Kondoh, Y. (2008, May). *Automatic assessment of Japanese text readability based on a textbook corpus* [Conference session]. Proceedings of LREC' 08.

Schroeder, S., Häikiö, T., Pagán, A., Dickins, J. H., Hyönä, J., & Liversedge, S. P. (2022). Eye movements of children and adults reading in three different orthographies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(10), 1518.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H. D., Alexeeva, S., Amenta, S., . . .Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, *54*(6), 2843–2863.

Staub, A. (2015). Reading sentences: Syntactic parsing and semantic interpretation. In Pollatsek A., Treiman R. (Eds.), *The Oxford handbook of reading* (pp. 202–216). Oxford University Press.

Staub, A., & Rayner, K. (2007). Eye movements and online comprehension processes. *The Oxford Handbook of Psycholinguistics*, *327*, 342.

Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88–99). Association for Computational Linguistics.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348.

Sun, K., & Wang, R. (2024). Computational sentence-level metrics predicting human sentence comprehension. *arXiv preprint arXiv:2403.15822*.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6, pp. 497–516). Pearson.

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134.

Wiechmann, D., Qiao, Y., Kerz, E., & Mattern, J. (2022). Measuring the impact of (psycho-) linguistic and readability features and their spill over effects on the prediction of eye movement patterns. *arXiv preprint arXiv: 2203.08085*.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in r//egression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48.

Yadav, H., Husain, S., & Futrell, R. (2019). Are formal restrictions on crossing dependencies epiphenominal? In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories* (pp. 2–12). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7802