Cognition 242 (2024) 105636

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Universality in eye movements and reading: A replication with increased power

Simon P. Liversedge [a,*], Henri Olkoniemi [b,c], Chuanli Zang [a,d], Xin Li [d], Guoli Yan [d], Xuejun Bai [d], Jukka Hyönä [c]

[a] University of Central Lancashire, UK
[b] University of Oulu, Finland
[c] University of Turku, Finland
[d] Tianjin Normal University, PR China

ARTICLE INFO

Keywords:
Eye movements
Reading
Universality

ABSTRACT

Liversedge, Drieghe, Li, Yan, Bai and Hyönä (2016) reported an eye movement study that investigated reading in Chinese, Finnish and English (languages with markedly different orthographic characteristics). Analyses of the eye movement records showed robust differences in fine grained characteristics of eye movements between languages, however, overall sentence reading times did not differ. Liversedge et al. interpreted the entire set of results across languages as reflecting universal aspects of processing in reading. However, the study has been criticized as being statistically underpowered (Brysbaert, 2019) given that only 19–21 subjects were tested in each language. Also, given current best practice, the original statistical analyses can be considered to be somewhat weak (e.g., no inclusion of random slopes and no formal comparison of performance between the three languages). Finally, the original study did not include any formal statistical model to assess effects across all three languages simultaneously. To address these (and some other) concerns, we tested at least 80 new subjects in each language and conducted formal statistical modeling of our data across all three languages. To do this, we included an index that captured variability in visual complexity in each language. Unlike the original findings, the new analyses showed shorter total sentence reading times for Chinese relative to Finnish and English readers. The other main findings reported in the original study were consistent. We suggest that the faster reading times for Chinese subjects occurred due to cultural changes that have taken place in the decade or so that lapsed between when the original and current subjects were tested. We maintain our view that the results can be taken to reflect universality in aspects of reading and we evaluate the claims regarding a lack of statistical power that were levelled against the original article.

1. Introduction

Liversedge et al. (2016) reported a study that at the time of publication represented a serious effort to experimentally identify factors that might account for common variance in eye movement behavior during reading in three languages (Chinese, Finnish, and English), all of which have markedly different orthographies. At the core of the investigation was the idea that any such variables might reflect aspects of representation and process in reading that are universal across languages. An important aspect of the study by Liversedge et al. was the repeated translation and backtranslation of the written stimuli across the three languages under investigation. By undertaking this process carefully, it was possible to develop expository texts to be used as experimental stimuli that were maximally comparable in terms of their content and correspondence. This aspect of the study ensured that any differences in eye movements that might be observed across languages could be attributed to differences in the nature of processing rather than being caused by content differences. A primary prediction in this study was that whilst the specific characteristics of eye movements might change for orthographies that represent linguistic information in quite different visual forms, the extraction of the basic meaning for comparable portions of text should be similar regardless of its written form. That is to say, overall reading times for sentences that convey maximally comparable semantic information should be similar across languages (cf.,

* Corresponding author at: School of Psychology and Humanities, University of Central Lancashire, Preston, PR1 2HE, UK.
E-mail addresses: SPLiversedge@uclan.ac.uk (S.P. Liversedge), henri.olkoniemi@oulu.fi (H. Olkoniemi), CZang@uclan.ac.uk (C. Zang), hyona@utu.fi (J. Hyönä).

https://doi.org/10.1016/j.cognition.2023.105636
Received 7 October 2022; Received in revised form 10 September 2023; Accepted 5 October 2023
Available online 17 October 2023

Frost, 2012). Beyond this central claim, Liversedge et al. also used their study to provide eye movement descriptives of reading performance across each of the three orthographies that were their focus. This exploratory aspect of the study aimed to provide a characterization and assess comparability in relation to basic patterns of eye movements that readers make when written language with different orthographic forms is processed.

To make their comparisons, eye movement data were obtained from three groups of native readers who were undergraduate university students: 21 Chinese subjects, 19 Finnish subjects and 19 English subjects. The fact that subjects were selected from comparable populations (university students of a similar age) ensured that the possibility that subject group differences might contribute to any cross-language effects was minimised. Additionally, given that each sentence of each text provided a very comparable unit of linguistic information (in terms of its basic conceptual meaning), analyses were conducted across languages for eye movement measures computed for each sentence. In our analyses, we assessed the extent to which word length, word frequency and the number of words in a sentence captured variance in eye movement measures.

In line with Liversedge et al.'s predictions, analyses provided no evidence for differences in total sentence reading times across languages. The absence of an effect for total time was particularly striking given that there were robust differences between languages in more fine-grained aspects of eye movement behavior (e.g., mean fixation duration, number of fixations, forward saccade extent, as well as skipping, refixation and regression rates). Additionally, word length, word frequency and the number of words in a sentence accounted for common variance across the languages. Liversedge et al. took the findings as evidence to suggest that whilst unique characteristics of the orthography produce differences in moment-to-moment aspects of eye movement control in reading, it was also the case that basic lexical properties of words were key indices of linguistic processes across languages. These word properties may be candidates for universals with respect to the nature of representation and process. Finally, the results were also taken to reflect how the linguistic density of an orthography modulates the balance in information exchange between the visual encoding and linguistic processing systems. For orthographies that are relatively linguistically dense (e.g., Chinese), delivery of information from the visual encoding system to the linguistic processing system can occur rapidly whilst later processing within the linguistic system proceeds more slowly, acting as something of a bottle neck. In contrast, for orthographies that are relatively linguistically sparse (e.g., Finnish), the counterpart situation exists. Here, visual encoding has to occur across multiple successive fixations through text that is substantially horizontally extended, with the consequence that the rate of visual delivery acts as the bottleneck constraining the rate at which subsequent linguistic processing can occur.

1.1. Weaknesses and criticisms of the original study

At the time of publication, the study by Liversedge et al. was novel in its approach in a number of respects. The adoption of stimuli that had been successively translated and then back-translated to investigate cross-linguistic differences in reading was a unique characteristic in relation to existing cross-linguistic eye movement studies. Also, the use of Linear Mixed-Effect Models (LMM) to quantify effects across languages was considered forward looking. And an eye movement investigation of three languages with quite different orthographic characteristics had not been conducted before.

Nonetheless, it was also the case that there were empirical weaknesses to the study. For example, prior to testing their subjects, Liversedge et al. did not undertake any formal power computations to establish the appropriate number of subjects and the number of stimuli that would be necessary to provide a robust test of their hypotheses. Also, the primary test for effects of language was conducted using ANOVA, not LMM. Additionally, the LMMs that were adopted in the analyses whilst relatively advanced at the time of testing, in current context, the field has advanced and new practises in model building have emerged (e.g., Meteyard & Davies, 2020). Finally, the (arguably) central theoretical claim in the study rested on a null effect, that is, a lack of difference in total reading times across languages. The absence of such an effect was taken to reflect comparability in the overall time taken for the extraction of sentential meaning, and to some extent, evidence for universality in process. However, Liversedge et al. failed to undertake Bayesian (or other) analyses to evaluate the evidence in favor of the null and given that the number of subjects that was tested in each language was relatively low, the null effect might plausibly have occurred due to a lack of statistical power (Brysbaert, 2019). These are clearly significant and important empirical weaknesses.

1.2. Objectives of the present study

The present study represents an effort to further explore the issues that Liversedge et al. investigated whilst also addressing earlier weaknesses and adopting a more thorough and sophisticated approach to our work. Thus, we had a number of objectives in undertaking this work. First, given that the study by Liversedge et al. has been criticized for being “severely underpowered” (Brysbaert, 2019, p. 58), we wished to test a larger number of subjects in each of the languages than we originally tested in order to ensure that we had greater, and sufficient, power in our analyses to justify the theoretical claims we wish to make. Brysbaert pointed out that studies with the number of subjects that Liversedge et al. tested (between 19 and 21 Ss in each language group) had insufficient power to detect effect sizes of $d = 0.4$, an effect magnitude that has been suggested to be the mean effect size in psychology (Stanley, Carter, & Doucouliagos, 2018). Brysbaert further argued that for such an effect size, it would be necessary to test “at least 100 individuals per language” (Brysbaert, 2019, p. 58; see also Brysbaert & Stevens, 2018). Thus, our first objective in the current study was to assess this suggestion and to ensure that we tested an adequate number of subjects such that we would meet this stipulation in respect of power. For this reason, at the outset, one of our objectives in the present study was to test 100 subjects in each language, that is, 80 subjects beyond those tested in the original experiment. Given that our experimental methods and stimuli remained identical to the original, it would then be possible to combine the new and original data sets to deliver 100 subjects per language.

If the results and conclusions reported by Liversedge et al. are reliable, then at the very least, basic pattern of effects in respect of the primary theoretical claim (i.e., the null effects for total sentence reading time) should replicate. This represents a quite straightforward and strong prediction. However, as we have indicated, we acknowledge the power concerns associated with the original study, and therefore, it is quite possible that the effects reported by Liversedge et al. may not be reliable, and of course, if this is the case, then an alternative pattern of effects will occur. It is extremely difficult to predict the particular form of any alternative pattern of effects given that there are no other studies to date that have reported comparative total sentence reading time data for text reading in Chinese, Finnish, and English readers. We note that Siegelman, Schroeder, Acartürk, et al. (2022) reported total reading times for individual words (though data for sentences and passages for texts with, and without, comparable semantic content are available in an OSF repository), and that Brysbaert (2019) reported silent reading rates (based on an estimated words per minute metric) across these languages. However, at best, these measures only approximate total sentence reading time and differences in word lengths across languages in their studies complicate matters further with respect to direct comparisons. Given this, if the pattern of effects reported by Liversedge et al. does not occur, then we remain agnostic with respect to our predictions of the particular pattern of alternative effects.

A second important objective of the current study to which we have

already alluded, was to establish whether we could replicate our findings. The work that was originally reported by Liversedge et al. was formative and exploratory. We, therefore, felt a scientific responsibility to ensure that our results were trustworthy.

Our third objective was to adopt a more sophisticated and refined approach in our analyses of the data in the experiment. Whilst Liversedge et al. adopted a LMM approach to the analyses of their data, an analytical method that was considered relatively advanced at the time, the linear mixed models that were computed included solely random intercepts. Current best practice with respect to linear mixed models requires that random slopes are also included in the models (e.g., Barr, Levy, Scheepers, & Tily, 2013; Meteyard & Davies, 2020). For this reason, we adopted this approach in the current investigation. More importantly, when Liversedge et al. undertook their analyses, they did not compute a statistical model that included all three of the languages. Liversedge et al. argued that since Finnish and English are alphabetic languages whereas Chinese is a character-based orthography, it was very difficult to undertake direct comparisons between the former and the latter due to differences in their nature and visual complexity (Liversedge et al., referred to this as the "apples and pears" problem). To be clear, the metric that might allow direct comparison between Finnish and English (e.g., the number of letters in a word) could not be computed for Chinese. On this basis, the Chinese data were never considered alongside the Finnish and English data in the same statistical model, meaning that there was no simultaneous formal statistical comparison of performance across all three languages.

In the present investigation, we sought to return to this issue and explore whether it might be possible to move beyond this position by computing a comparability metric of visual complexity for Chinese, Finnish, and English that might allow for direct three-language comparisons. Of course, traditionally, in alphabetic languages, the visual complexity of a word is indexed by its number of constituent letters (i.e., word length). Indeed, word length and visual complexity are perfectly confounded in most alphabetic languages, that is, as word length increases, so the visual complexity of a word increases (see Fu, Liversedge, Bai, Moosa, & Zang, 2023). In contrast, in Chinese, the visual complexity of a word is usually indexed by the number of strokes that comprise its constituent characters. Words can be formed from one or more characters, and in turn, characters are formed from strokes. Characters may be comprised of a very small number of strokes (one or two), through to a large number of strokes (thirty or even more). It is important to note, though, that alphabetic letters are also comprised of strokes. For example, the letter "l" is formed from a single stroke, the letter "t" is formed from two strokes and the letter "k" is formed from three strokes. To this extent, one shared characteristic of words across languages is that they are formed from strokes and the visual complexity of a word in each language might therefore be computed on the basis of the number of its constituent strokes (note that, where applicable, diacritical marks, e.g., umlauts like Ä, may be counted as 2 strokes). For this reason, we computed the stroke complexity of each word in our texts for each language and then took the average word complexity for each sentence and included this metric in our analyses. Accordingly, note that we did not include the word length metric reported by Liversedge et al. that was originally incorporated into the linear mixed model for Finnish and English. Thus, through the inclusion of visual complexity, alongside the lexical frequency metric that was also employed by Liversedge et al., we were able to build linear mixed models that incorporated all three languages. Assuming the stroke complexity metric to adequately capture the visual density of words across languages, then this approach represents a significant advance on the approach that Liversedge et al. adopted.

Our fourth objective was to broaden our consideration of different eye movement measures in our analyses. In addition to the four sentence-level reading time measures reported by Liversedge et al. (total fixation time, total number of fixations, average fixation duration, and rightward saccade size), we computed the number of regressions that were made, as well as three additional sentence-level measures of processing time (first-pass forward-fixation time, first-pass rereading time, and look-back time) that allow for more detailed consideration of the types of eye movement behavior that contribute to the overall total sentence reading times (Hyönä, Lorch, & Rinck, 2003). In order to undertake these analyses, it was necessary to re-analyse the original raw eye movement data sets and because we were doing this for the new measures, we also recalculated all the original eye movement measures. Additionally, during these computations, we did not adopt any filtering procedures (i.e., did not filter reading time values >2.5 SD), as recommended when analyzing data with LMMs (Baayen & Milin, 2010). Finally, for each text, we identified word units and on this basis, we computed word skipping and refixation rates.

Our inclusion of these additional eye movement measures allows us to provide a more detailed characterization of any differences in the types of eye movement behavior that readers in each of the languages adopted in order to comprehend the sentences. Specifically, the additional analyses will allow us to determine whether there are differences across the languages in the proportion of total sentence reading time they spend initially processing words, or re-reading text (and re-reading text in different ways). To summarise our predictions, first, we anticipate that the primary finding reported by Liversedge et al. (2016), namely that there were no differences in total sentence reading time between the three languages will be replicated. Additionally, the skipping and refixation measures will show patterns of effects that reflect cross-linguistic differences in the mean length of words. The assumed language differences in skipping and refixation rate are derived from the nature of the languages. In terms of language typology, Chinese and English are analytic languages, whereas Finnish is a synthetic language. The difference lies in how relations between words in sentences are expressed. In analytic languages they are conveyed by function words and word order, whereas in synthetic languages they are expressed by morphological inflections attached to content words. Therefore, words are generally longer in synthetic than analytic languages and are also likely to carry more information. As regards Finnish, up to four different inflections can be attached to nouns. This feature departs significantly from Chinese where nouns have no inflections. English is closer to Chinese in that only the plural inflection may be added to nouns. Similarly, verb morphology is significantly more complex in Finnish than in Chinese or English. Given these differences between languages, we predict that Chinese, which is comprised of quite short words (mainly one or two characters in length) should show increased skipping and reduced refixation rates, whereas Finnish which has significant morphological complexity and is agglutinative, both of which contribute to a relatively long average word length, should produce reduced skipping rates and relatively high refixation rates. We expect that English will lie intermediary to Chinese and Finnish in relation to these two measures. Beyond these predictions, we also anticipated standard effects of word frequency, number of words and effects of (our novel index of) visual complexity. Furthermore, we anticipated a more complex set of interactive effects between these variables that should align with the original findings of Liversedge et al. In sum, it should be clear that the current study and sets of analyses deliver understanding of the nature of eye movement behavior in reading across the languages beyond that gained by Liversedge et al., and therefore, it has the potential to deliver significant additional theoretical insight.

1.3. Power calculations

The stipulation of subject numbers put forward by Brysbaert (2019) was based on a power calculation that assumes independent subject groups and does not capture the number of stimuli (or trials) per condition as a factor contributing to power. In studies investigating eye movements and reading, it is very well accepted that both the number of stimuli, as well as the number of subjects, contribute to the statistical power of an experiment. With this in mind, we sought to consider how

power analyses that capture the contribution of both the number of items as well as the number of subjects in an experiment might differ in their implications for subject numbers relative to those suggested by Brysbaert.

In our first assessment of power, we undertook simulations based on the original data set reported by Liversedge et al. As noted above, the use of ANOVA rather than LMM to analyse the difference in total sentence reading times between language was considered a shortcoming. Thus, we started by fitting a LMM with Language as fixed effect (see details regarding the model fitting in the Results) to confirm that there were no differences. Indeed, the LMM showed no difference between Chinese and English. However unexpectedly, the model did reveal a difference between Finnish and English, indicating that Finnish was read faster than English, $\beta = -0.18$, 95% CI [−0.35, −0.01], $t = -2.13$. To reiterate, this difference was not obtained in the original study, but we believe that it arose due to our recalculation of the total sentence reading time measures from the original data sets and our decision not to include filtering of the high reading times as was done in the original study (i.e., reading times >2.5 *SD*; see Baayen & Milin, 2010). The model and the reading times can be found in Appendix Table A1. We will return to consider this important effect in more detail in the Discussion.

We used the *simr* package (Green & MacLeod, 2016) in the *R* statistical software (Version 4.1.2; R Core Team, 2020) to estimate the number of subjects required for effective power in our new study (see Brysbaert, 2019). We based our simulation on the simple model using the original data set (see Table A1) and simulated an increased number of subjects to reach $N = 300$, the suggested sufficient number of subjects (Brysbaert, 2019). Then we estimated the number of subjects required to attain 80% power to detect a main effect of Language. The observed differences between the total sentence reading times for both comparisons that we used to estimate the main effect were for English vs. Chinese -371 ms (model estimate = −0.08) and for Finnish vs. English -423 ms (model estimate = −0.18). Power was estimated on the basis of 100 random samples. The analysis demonstrated that using LMM to analyse the data, it would be necessary to test at least a total of 61 subjects (i.e., 21 subjects per language; see Fig. 1). To further investigate the sensitivity of the estimated power, we calculated alternative total sentence reading times for each language in our study. We based these estimates on average sentence lengths for each language (see Table 2) and the words per minute values reported by Brysbaert (2019) for silent reading of language. These were 3399 ms for Chinese, 3683 ms for English (reading time for non-fiction), and 3237 ms for Finnish respectively. Observed differences between the total sentence reading times differed little for the comparison of Finnish vs. English (−446 ms, model estimate = −0.12), and were slightly smaller and in the opposite direction (i.e., Chinese being faster) for the comparison of English vs. Chinese (284 ms, model estimate = 0.09). The fixed effect estimates were fitted separately for each contrast (i.e., English vs Chinese, and Finnish vs. English) and 100 random samples were run to estimate the required number of subjects for both. These analyses showed that a total of 111–150 subjects (i.e., 37–50 subjects per language) would be required to attain 80% power (see Fig. 1).

In should be noted, that in the original study, more complex relationships between language, the number of words in a sentence and average lexical frequency were explored. Of particular interest in respect of these considerations was how such effects in Chinese (an unspaced character-based language) might contrast with effects in Finnish and English (spaced alphabetic languages). Of course, such relationships involved the assessment of interactive influences within LMMs and for such interactive effects an increased level of power beyond that required to test basic effects of language would be necessary. For this reason, we also fitted all the possible three-way interactions that include effect of language (see Results for details related to model fitting), using the original data, and on this basis, we estimated the number of subjects required to detect the observed three-way interactions between English and Chinese (Language × Average Frequency × Number of Words, $\beta = -0.07$) with adequate statistical power (> 80%; this model is reported in Table A2). Based on 100 simulations, we established that a minimum of 39 subjects (i.e., 13 subjects per language) would be necessary to reliably detect this effect, which did not increase the previous estimates.

Overall, the number of subjects was beyond that tested by Liversedge et al., and to reiterate, we acknowledge that the Liversedge et al. study was, therefore, underpowered. However, it is striking that the number of subjects required for adequate power is not substantially greater than the number originally tested (and certainly <100 per language). Thus, our power analyses, simulations and considerations in respect of critique

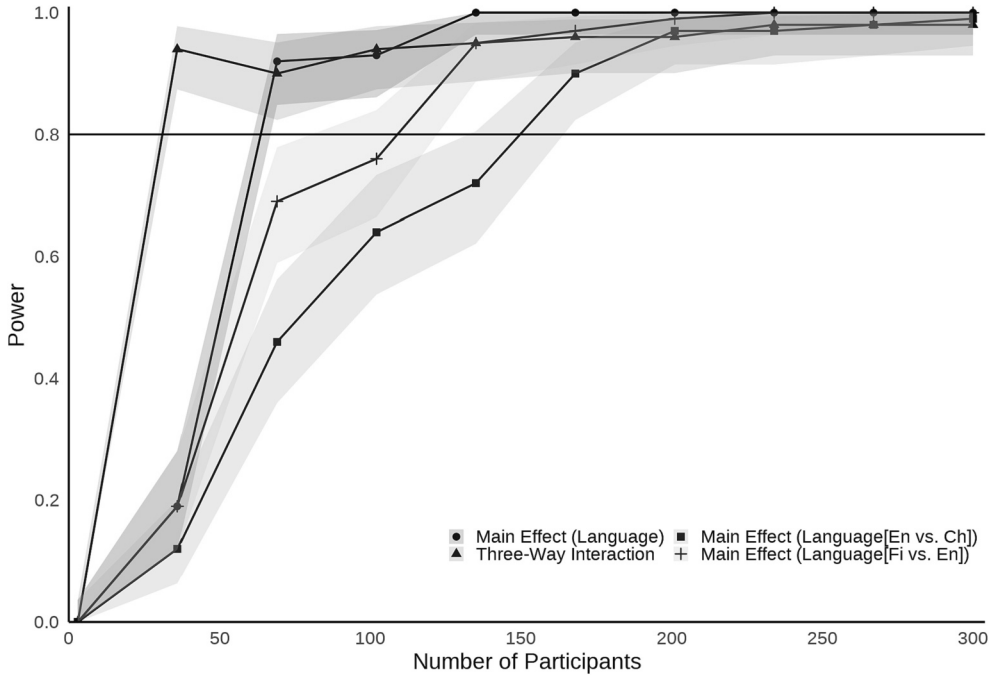

Fig. 1. Outcome of the powerCurve command (simr package) for different simulations. Shaded areas indicate 95% Confidence Intervals.

of the Liversedge et al. study led us to the following three conclusions: First, in principle, the original study was sufficiently (though admittedly, minimally) powered to test its primary hypothesis (and, of course, an alternative analytic approach to the basic question would have done this much more effectively); second, to detect relationships between languages, a minimum of 50 subjects per language would be necessary; finally, to meet stipulations of power from Brysbaert (2019), it would be necessary to test 100 subjects per language. On this basis, in the current study, we decided to use the original stimuli and experimental set up to test a total of 80 additional subjects per language. We did this to ensure that we met the most stringent requirements relating to considerations of power in respect of Liversedge et al. (2016).

We then undertook two sets of analyses. The first was based on the 80 new data sets from each language and these analyses represent a larger scale, independent sample, replication of the original study. We report these analyses in the online repository https://osf.io/efqnx/. In addition to these analyses, and given our firm, a priori, intention to test additional subjects up to 100 in total (to meet the most stringent suggestions regarding necessary power), we combined the data from the subjects in the original study with the data from the 240 new subjects to achieve our goal of testing 100 subjects in each language. We report these analyses in full below.

2. Method

2.1. Subjects

Eighty native Chinese speakers (undergraduate students of Tianjin Normal University, the age and gender were not available unfortunately, though the age range and the proportion of females to males was approximately comparable to the Finnish and English subjects), 80 native English speakers (undergraduates of University of Central Lancashire, 62 females, mean age = 23 years, SD = 6 years), and 84 native Finnish speakers (undergraduates of University of Turku, 77 females, mean age = 24 years, SD = 7 years) took part in the experiment. All subjects had normal or corrected-to-normal vision.

The data from a previous experiment (Liversedge et al., 2016) using the same texts and instructions were pooled with the new data. The original data were collected from 21 native Chinese speakers (undergraduate students of Tianjin Normal University), 19 native English speakers (undergraduates of University of Southampton), and 19 native Finnish speakers (undergraduates of University of Turku). This yielded a total of 101 Chinese subjects, 99 English subjects, and 103 Finnish subjects.

2.2. Apparatus

Eye movements were recorded monocularly using an EyeLink 1000 eye-tracker (SR Research Ltd., Ontario, Canada) at a 1000 Hz sampling frequency. Technical specifications related to the recordings at each research site are presented in Table 1.

2.3. Materials

The subjects read the same eight experimental texts as used in the Liversedge et al. (2016) study. They were short expository texts on a variety of topics (i.e., sheep, car race, football, oil, sugar, restaurant tipping behavior, walking as exercise, and wind energy). One of the texts (i.e., sheep) was used as a practice text. The texts were initially written in English and then translated to Chinese and Finnish, and subsequently back-translated to English (for further details, see Liversedge et al., 2016). Descriptive statistics of the texts are presented in Table 2.

As is apparent from Table 2, in Finnish the same contents are expressed using fewer words than in English or Chinese. Moreover, the average frequency of words in a sentence is considerably higher in English compared to Chinese and Finnish due extremely frequent articles and prepositions featuring as separate words. Finnish does not have articles and expressions appearing in English as prepositions are often expressed as inflectional endings. Likewise, Chinese lacks articles. It also lacks morphological marking, such as plural inflections; moreover, the same word can appear as either a noun or a verb. Thus, for the analyses the frequency measure was standardized separately for each language. Strokes per word was used as a measure of visual complexity mimicking prior studies on Chinese reading (e.g., Zang, Fu, Bai, Yan, & Liversedge, 2018). In Chinese, we counted the number of strokes included in the characters making up each word. In English and Finnish, we counted the number of strokes needed to represent the letters contained in each word. Thus, for example, letters "l" and s contain one stroke, letters "h" and "i" contain two strokes and letters "k" and "m" three strokes. In alphabetic languages, this measure correlates highly with average word length, $r_{English} = 0.96$, 95% CI [0.96, 0.96], $r_{Finnish} = 0.96$, 95% CI [0.95, 0.96], but only moderately in Chinese, $r_{Chinese} = 0.53$, 95% CI [0.52, 0.54]. We used this measure instead of the number-of-letters/characters measure used in Liversedge et al. (2016) for the reason that in Chinese there is very little variance in the number of characters in words; about 70% of words contain two characters, and there are few words that contain more than three characters.

The content of each text was split down into 2–6 ($M = 4.1$) pages so that each page contained 1–8 sentences. Pages were presented one at a time on the computer screen to allow comfortable reading whilst eye movements were recorded.

Table 1
Technical specifications of the experimental setup at the three research sites.

Specification	Experiment	Language		
		Chinese	English	Finnish
Monitor	Original	19" Dell	20" ViewSonic P227F	20" ViewSonic G225F
	New	24" Asus VG248QE	24" BenQ Zowie XL2540	24" BenQ Zowie XL2411
Eye-to-monitor distance (cm)	Original	70	70	70
	New	68	68	69
Resolution (pixels)	Original	1024 × 768	1024 × 768	1024 × 768
	New	1024 × 768	1024 × 768	1024 × 768
Font	Original	Song	Courier New	Courier New
	New	Song	Courier New	Courier New
Character size (pixels)	Original	25	14	14
	New	21	14	14
Character size (visual angle)	Original	0.87	0.42	0.46
	New	0.69	0.46	0.45

Original refers to the study reported in Liversedge et al. (2016).

Table 2
Descriptive statistics of the text characteristics in the three languages (Standard Deviations in Parentheses).

Text characteristic	Chinese	English	Finnish
Number of words in the text corpus	1774	1762	1301
Average number of words per sentence	14.73 (7.60)	14.61 (7.22)	10.63 (5.41)
Average word frequency per sentence	5.68 (0.87)	7.76 (0.73)	5.98 (1.03)
Average number of strokes per word	10.97 (1.53)	8.54 (1.29)	13.46 (2.72)
Average word length per sentence (in characters)	1.55 (0.20)	5.67 (0.75)	8.50 (1.40)

2.4. Procedure

The subjects were tested individually. They were informed that they

would be required to read and comprehend texts that would be presented passage by passage on the monitor. When they finished reading each page, they pressed a button from the keyboard to move to the next. After each text, two binary ('Yes' or 'No') questions were asked concerning the content mentioned in the text (e.g., *"Do windmills work well on islands?"*). The questions were answered by pressing designated *Yes* and *No* buttons on the keyboard; half of the questions required a Yes and half a No response. The response accuracy was high in the three groups of readers ($M_{Chinese} = 88\%$, $SD_{Chinese} = 21\%$; $M_{English} = 85\%$, $SD_{English} = 36\%$; $M_{Finnish} = 83\%$, $SD_{Finnish} = 38\%$) with no difference between the groups: English vs. Chinese, OR = 0.60, 95% CI [0.35, 1.04], Finnish vs. English, OR = 0.96, 95% CI [0.49, 1.87] (The full model on response accuracy is presented in Table A2 in the Appendix).

Prior to the reading task, a nine-point calibration procedure was completed. After a successful calibration (average calibration error $< 0.5°$) the text passages were presented one at a time. The eye-tracker was recalibrated after each text. The experimental session took approximately 40 min.

3. Results

3.1. Dependent variables

Similar to Liversedge et al. (2016), four sentence-level eye movement measures were computed: total fixation time, total number of fixations, average fixation duration, and rightward saccade size (see Table 3). *Total fixation time* is the sum of all the fixations made on the sentence. *Total number of fixations* is the number of fixations that landed on the sentence. *Average fixation duration* is the average duration of all fixations made on each sentence. *Rightward saccade length* is the average length (in number of characters and in visual angle) of progressive saccades (advancing from left to right) done in a sentence. As total fixation time and total number of fixations correlated very strongly with each other, $r = 0.96$, 95% CI [0.96, 0.96], total number of fixations was not analyzed further.

In addition, in order to assess the time-course of sentence-level reading in more detail, three additional eye fixation measures were calculated to tease apart the total fixation time measure (see Hyönä et al., 2003): first-pass forward fixation time, first-pass rereading time, and look-back time (see Table 3). *First-pass forward fixation time* is the summed duration of fixations that land on unread parts of the sentence during first-pass reading (i.e., before progressing to the next sentence). *First-pass rereading time* is the summed duration of fixations made when re-inspecting a sentence before moving on to the next sentence. *Look-back time* is the summed duration of fixations returning to a sentence from subsequent parts of the text after the first-pass reading. Observed means and standard deviations of the sentence-level eye movement measures are presented in Table 3. Moreover, we also computed two word-level eye movement characteristics (probability of word skipping and probability of refixating a word) as well as the overall probability of making a regression (a leftward saccade; see Table 3).

Table 3
Descriptive statistics for the eye movement measures for the three languages with all the subjects.

	Chinese		English		Finnish	
Measure	M	SD	M	SD	M	SD
Total fixation time	2989	2371	3398	2223	3222	2379
Total number of fixations	12.96	9.67	16.79	10.62	15.95	11.14
First-pass forward fixation time	1853	1330	2547	1481	2329	1449
First-pass rereading time	1085	1267	912	1012	890	1155
Look-back fixation time	1071	1360	754	1078	1045	1294
Average fixation duration	226	45	203	37	199	36
Rightward saccade length (visual angle)	3.06	1.48	4.29	1.23	4.09	1.12
Rightward saccade length (characters)	4.28	2.18	9.47	2.62	8.99	2.46
Word skipping rate	0.57	0.50	0.35	0.48	0.23	0.42
Word refixation rate	0.09	0.29	0.18	0.39	0.27	0.44
Regression rate	0.17	0.38	0.16	0.37	0.15	0.36

3.2. Data analysis

The data were analyzed using linear mixed-effects models (LMM; e.g., Baayen, Davidson, & Bates, 2008) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in the *R* statistical software (Version 4.0.3; R Core Team, 2020). Separate models were built for each eye movement measure. It has been recommended that only minimal data filtering is conducted when analyzing data with mixed-effects models appended with model criticism (Baayen & Milin, 2010). In other words, we compared models using non-filtered reading times to those using filtered data, where reading times $> |2.5|$ *SD* were filtered out. It turned out that R^2 values were either better with the non-filtered data or virtually identical between the filtered and non-filtered models, thus favoring the use of non-filtered data. The reading time measures were skewed and consequently transformed. Box-Cox Power Transform was used to identify appropriate transformations. First-pass forward fixation time and rightward saccade size were square-root transformed and the other measures (i.e., total fixation time, first-pass rereading time, look-back time, and average fixation duration) were logarithmically transformed prior to the analyses.

As for the fixed effects variables, *language* was fitted in the models as a repeated contrast-coded fixed effects variable (Chinese was compared to English, and English was compared to Finnish; for more details, see Schad, Vasishth, Hohenstein, & Kliegl, 2020), as in Liversedge et al. (2016). *Number of words*, *average log frequency of words,* and *visual complexity of words* in a sentence were fitted in the models as centered fixed effects variables. The effect of these three variables to total sentence reading time is illustrated in Fig. 2. In general, one unit increase in average log frequency decreased the total sentence reading time by −36 ms, 95% CI [−54, −17], and one unit increase in visual complexity increased total sentence reading time + 96 ms, 95% CI [88, 104], and one word increase +211 ms 95% CI [208, 214].[1] Following Liversedge et al. (2016), all the possible three-way interaction combinations including language were also added in the models.

Subjects and sentences were entered in the models as random intercepts (Baayen et al., 2008). The maximal random structure was fitted in the models (Barr et al., 2013). If the model failed to converge with the full random structure, it was trimmed top-down starting with correlations between factors (see Brauer & Curtin, 2018). For the sake of brevity, only significant effects are reported in the text. The final models are reported in the Appendix (Tables A3–A9). The data and the R-scripts are made available via Open Science Framework https://osf.io/efqnx/.

The models with the new 243 subjects and the models with the all 303 subjects produced the same pattern of results with only minor differences. As the differences were only minor and the effects were always in the same direction, we only report models with the all 303 subjects here. All the final models with the new 243 subjects and their respective descriptive statistics are reported as an online Appendix in the Open Science Framework.

3.3. Total fixation time

First, *a simple model* was fitted for total sentence fixation time to investigaite differences between the languages (see Table A3). The model showed that total fixation times were longer for English than

[1] We also created figures similar to Fig. 2 for the first pass reading time and the average fixation duration. The effects in these figures were extremely similar to the effects shown in Fig. 2, and therefore, to avoid redundancy, they are not included.

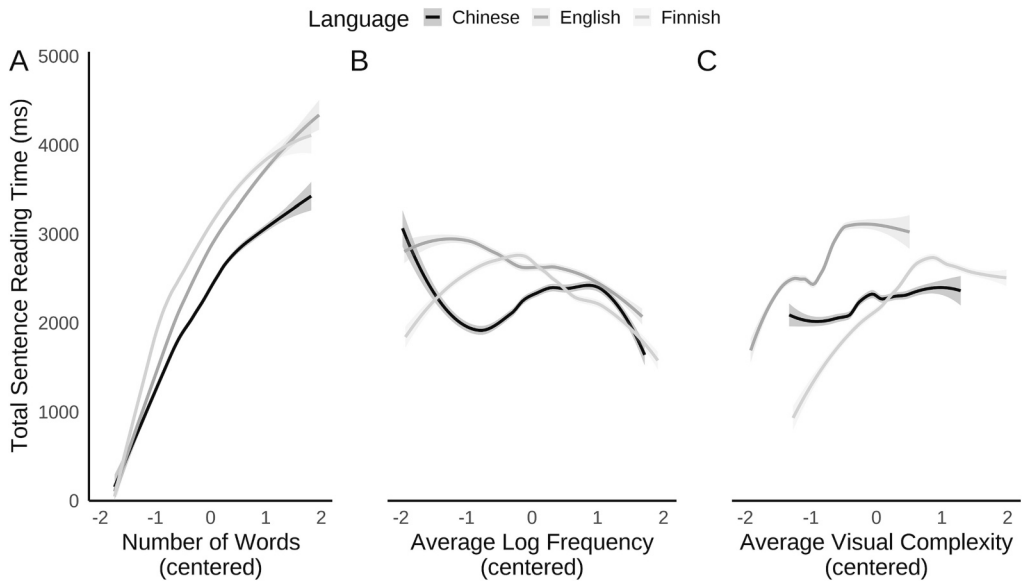

Fig. 2. Total sentence reading times (ms) as a function of number of words (A), average log frequency (B), and average visual complexity (C). Shaded areas in the panels represent 95% Confidence Intervals.

Chinese and Finnish. To explore further these findings, Bayes Factor (BF) was calculated for both comparisons. Bayesian LMM was built using the *rstanarm* package (Goodrich, Gabry, Ali, & Brilleman, 2020) in *R*. Normal prior (i.e., following normal distribution) was used, and standard deviation for the prior was extracted from Liversedge et al. (2016; $SD = 0.71$). The Savage-Dickey method (Wagenmakers, Lodewyckx, Kuriyal and Grasman, 2010) was applied for estimating the Bayes Factor using the *bayestestR* package (Makowski, Ben-Shachar and Lüdecke, 2019). The results showed strong evidence against the null hypothesis for the comparison between Chinese and English, BF = 4453.71, but evidence favoring the null hypothesis for the comparison between Finnish and English, BF = 0.77 (see Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019, for guidelines in interpreting BFs).

It was noticable that the observed total sentence reading times differed between the original study and the new dataset, with the 838 ms decrease in Chinese total fixation times between the original and new dataset being the most dramatic one (English showed a 139 ms and Finnish a 432 ms increase in total fixation times). Hence, we decided to conduct an explotarory analysis to compare whether a statistical difference emerged between the original and new data set in total fixation time. To test this, we built a LMM with *language* and *dataset* (old vs. new; deviation coded) fitted into the model as the fixed effects. The model revealed an interaction between *language* (English vs. Chinese) and *dataset* (see Table A10). This interaction indicates that for Chinese, total fixation times were indeed faster in the new than the original dataset, whereas for English and Finnish the difference between the original and new dataset was non-significant. We will return to this finding in the Discussion.

The complex model for total fixation time revealed four main effects (Table A4). First, there was an effect of *language*; Chinese subjects read the sentences with shorter total fixation times than English subjects. Second, an effect of *average word frequency* was obtained; total fixation time decreased when average word frequency in sentences increased. Third, an increase in *number of words* in the sentence was associated with an increase in total fixation time. Finally, total fixation time also increased as a function of *visual complexity of words*.

The main effects were modified by two-way interactions. The interaction between *average word frequency* and *number of words* suggests that word frequency effects are more robust for sentences containing many words. On the other hand, the interaction between *language* (Finnish vs. English) and *number of words* suggests that the number-of-words effect is more robust for Finnish than English (see Fig. 3). The same is true between Finnish and Chinese. Here English and Chinese patterned together, as the Chinese-English comparison did not interact with number of words. This effect is likely to reflect differences in synthetic (Finnish) and analytic (Chinese and English) languages briefly described in the Introduction. Words in Finnish are generally longer and contain more information than words in Chinese and English – hence a greater number-of-words effect in Finnish.

In addition, a three-way interaction emerged between *language* (English vs. Chinese), *average word frequency*, and *number of words* (see Fig. 3). The interaction suggests that Chinese readers demonstrate a number-of-words effect of similar magnitude for sentences containing frequent or infrequent words. On the other hand, for English readers, the number-of-words effect is more robust for sentences containing infrequent than frequent words. The same is true for Finnish, which patterned together with English (i.e., the English-Finnish contrast was not involved in the interaction). The finding that Chinese readers demonstrated an effect of the number-of-words of similar magnitude regardless of word frequency may reflect their tendency to skip over many words (on average 57% of words were skipped). The condensed format of the Chinese script makes frequent word skipping possible, including also words that are relatively infrequent.

In sum, the results for total fixation time, long sentences, and sentences containing many infrequent or visually complex words were read with longer total fixation times than short sentences and sentences containing frequent or visually less complex sentences. Moreover, the number-of-words effect was more robust for sentences containing infrequent than frequent words. Regarding language differences, two effects emerged. First, Finnish displayed a stronger number-of-words effect than Chinese or English. Second, English and Finnish differed from Chinese by displaying a stronger number-of-words effect for sentences containing infrequent than frequent words. In what follows, we examined the time course of the effects by decomposing total sentence fixation time to first-pass and second-pass (called look-back time) fixation time. Moreover, the first-pass reading of sentences was further decomposed into forward and rereading fixation time.

3.4. First-pass forward fixation time

The model for first-pass forward fixation time revealed main effects of *language* (English vs. Chinese), *average word frequency*, *number of*

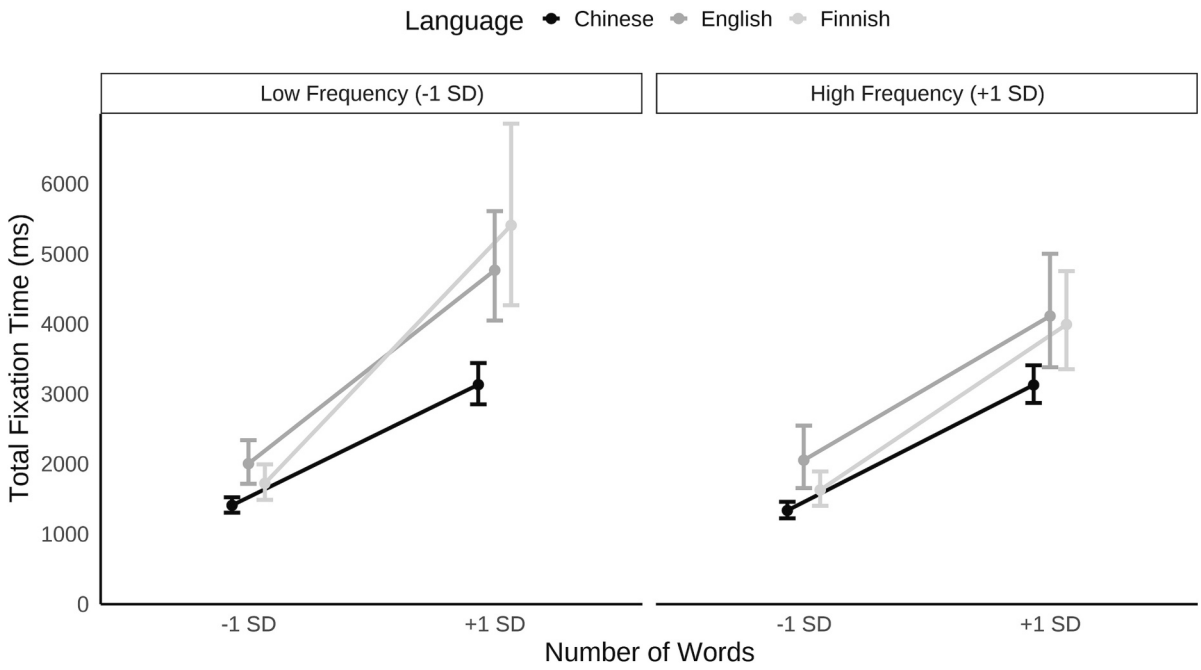

Fig. 3. Model estimates for total fixation time back-transformed to ms from log-values. Number of words and average word frequency are divided into ±1 *SD* for illustrative purposes. Error bars represent 95% confidence intervals.

words, and *visual complexity of words* (Table A5). The direction of these effects was the same as with total fixation time. These main effects were qualified by four two-way interactions. First, there was an interaction between *average word frequency* and *number of words*. The number-of-words effect was slightly stronger for sentences containing frequent than infrequent words when the number of words was relatively low, but this effect wore off as the number of words increased (Fig. 4A). Second, there were an interaction between *number of words* and *visual complexity*, indicating that the number-of-words effect was notably stronger for sentences containing many visually complex words (Fig. 4B). Third, the interaction between *language* (Finnish vs. English) and *average word frequency* indicates a stronger frequency effect for Finnish than English readers (Fig. 4C). The model did not show a similar interaction for the Chinese-English comparison, suggesting that average frequency affected Chinese and English similarly. Fourth, the interaction between *language* (Chinese vs. English) and *number of words* reflects a stronger number-of-words effect for English than Chinese readers (Fig. 4D). The model did not show a similar interaction for the Finnish-English comparison, suggesting that number of words affected Finnish and English similarly.

To sum up, the results for first-pass forward fixation time were similar to those for total fixation time. The model revealed effects of number of words, average word frequency and visual complexity, suggesting that the examined sentence features exerted an immediate effect in sentence processing. Also, the interaction between number of words and average word frequency observed in total fixation time emerged in first-pass forward fixation time. On the other hand, unlike in total fixation time, the number-of-words effect was greater for sentences containing visually complex words. With regard to language differences, two effects emerged. First, Finnish displayed a stronger word frequency effect than Chinese or English. An analogous effect was observed in total fixation time and we interpret it to reflect differences associated with synthetic and analytic languages. Second, Finnish and English displayed a stronger number-of-words effect than Chinese. This is likely to reflect a significantly greater skipping rate in Chinese (0.57) than in Finnish (0.23) or English (0.35). The skipping rate in turn reflects language differences in the horizontal extent of words in printed text.

3.5. First-pass rereading time

The model for first-pass rereading time revealed two main effects (see Table A6). First, there was an effect of *number of words*, indicating that readers tended to make more reinspective fixations when reading long than short sentences. This could just be a probabilistic effect: longer sentences offer more possibilities for making reinspective fixations simply based on their larger surface area. Second, the effect of *visual complexity* indicates that readers spent more time rereading sentences when the visual complexity of the words within sentences increased. The model showed no effects of *language* or *average word frequency*.

3.6. Look-back time

The model for look-back time revealed main effects of *language* (English vs. Chinese) and *average word frequency* (see Table A7*)*. These main effects were qualified by two two-way interactions between *average word frequency* and *number of words*, and *language* (English vs. Chinese) and *number of words*. First, the interaction between average word frequency and number of words indicates a steeper increase in look-back time as a function of number of words when reading sentences containing low-frequency than high-frequency words (Fig. 5A). Second, the interaction between language (Chinese vs. English) and number of words indicates that Chinese readers demonstrated a number-of-words effect in look-back time, whereas English readers did not (Fig. 3.

5B). On the other hand, the model did not show an interaction for the Finnish-English comparison. One possibility for the Chinese readers' increased look-backs is that it may be a compensation for frequent word skipping during first-pass reading (more than half of the words are left unfixated). In other words, Chinese readers may need to go back to long sentences to confirm their exact meaning. As mentioned in the Discussion, frequent changes in education policy may encourage university students in China to read quickly. Finally, the model showed no effects related to visual complexity.

3.7. Rightward saccade length

The model for rightward saccade length revealed main effects of

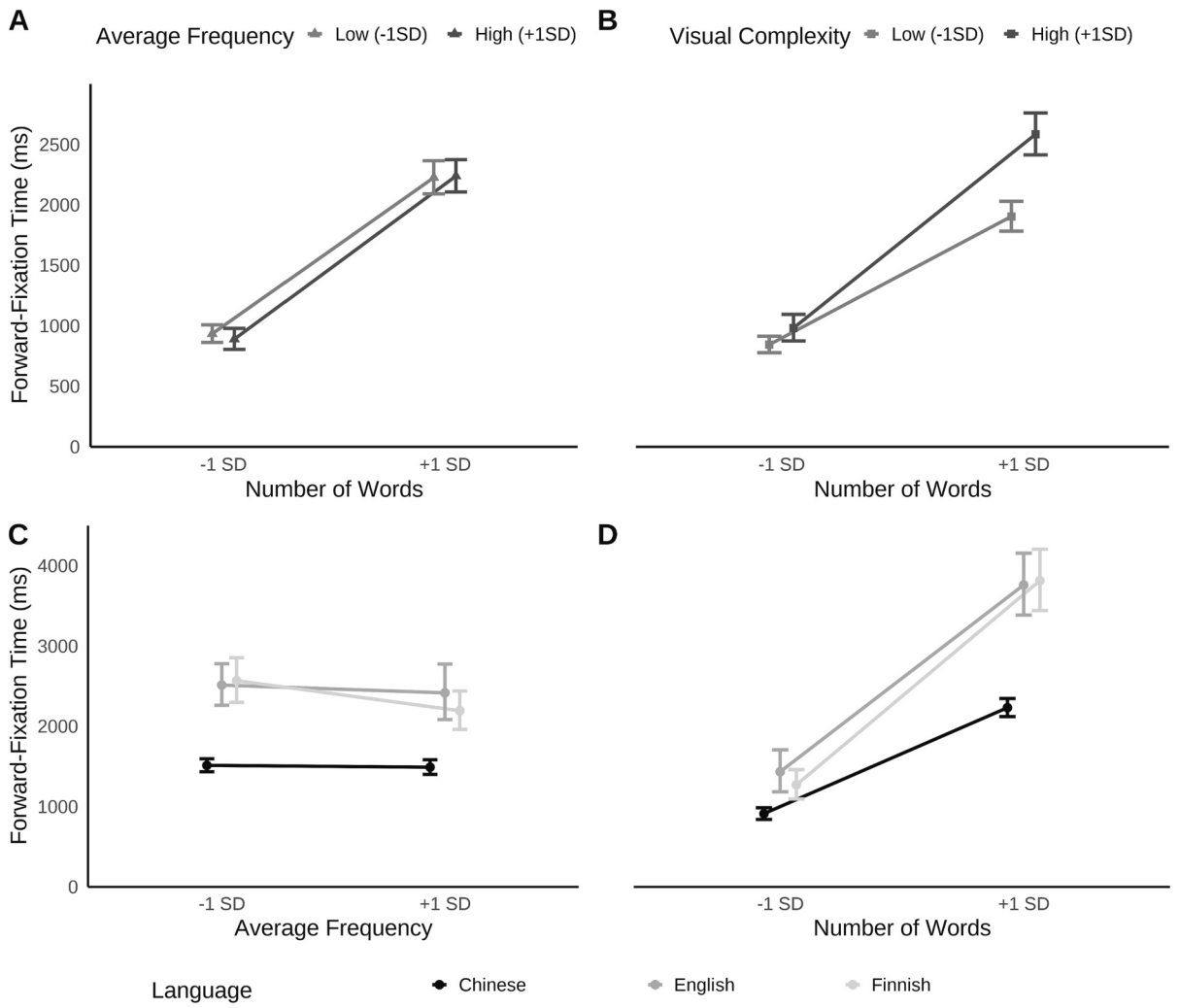

Fig. 4. Model estimates for first-pass forward-fixation time. Panel A represents an interaction between *number of words* and *average word frequency*. Panel B represents an interaction between *number of words* and *visual complexity*. Panel C represents an interaction between *language* (Finnish vs. English) and *average word frequency*. Panel D represents an interaction between *language* (English vs. Chinese) and *number of words*. First-pass forward-fixation times were back-transformed to ms from square-root values. Number of words, average word frequency, and visual complexity are divided into ±1 *SD* for illustrative purposes. Error bars represent 95% confidence intervals.

language (Chinese vs. English, and Finnish vs. English), *number of words*, and *visual complexity* (see Table A8). Chinese readers made the shortest saccades and Finnish readers the longest saccades. This effect reflects differences in average word length between the languages. The number-of-words effect suggest that readers made longer saccades in reading sentences containing many than few words. Finally, the main effect of visual complexity is due to readers making longer saccades in sentences containing many visually complex words. It should be born in mind that visual complexity means different things for Finnish and English than for Chinese. In alphabetic languages, there is a practically one-to-one relationship between visual complexity and average word length, whereas in Chinese it reflects the complexity of the characters in the words.

Rightward saccade length also revealed four two-way interactions between *number of words* and *visual complexity*, *language* (Finnish vs. English) and *average frequency*, *language* (English vs. Chinese and Finnish vs. English) and *number of words,* and *language* (English vs. Chinese and Finnish vs. English) and *visual complexity*. These effects were modified by two three-way interactions. First, there was an interaction between *language* (Finnish vs. English and Chinese vs. English), *average word frequency*, and *visual complexity*. In Chinese, sentences containing frequent words are read with longer rightward saccades when the words were of low visual complexity (see Fig. 6A). This may take the form of more frequent skipping of words. On the other hand, no frequency effect was observed in reading sentences containing visually complex words. In Finnish, rightward saccade length was not affected by either frequency or visual complexity (i.e., average word length). This makes sense considering the fact that words in Finnish are generally long resulting in a relatively few words being skipped. Finally, English displayed a pattern, where a frequency effect was apparent only for sentences containing visually complex words (i.e., long words). These are sentences containing few short words such as articles and prepositions but more likely containing a bit longer content words, although shorter than in Finnish.

Second, there was a three-way interaction between *language* (Finnish vs. English), *number of words* and *visual complexity*. Both Finnish and English readers tend to increase the length of progressive saccades when reading sentences containing many words, with one exception (see Fig. 6B). When reading visually complex sentences (i.e., sentences containing long words), English readers demonstrated an opposite pattern: They lengthened their progressive saccades when reading sentences containing few long words in comparison to sentences containing many long words. The model did not reveal a difference in the Chinese-English comparison.

The results of rightward saccade length may be summarized as two

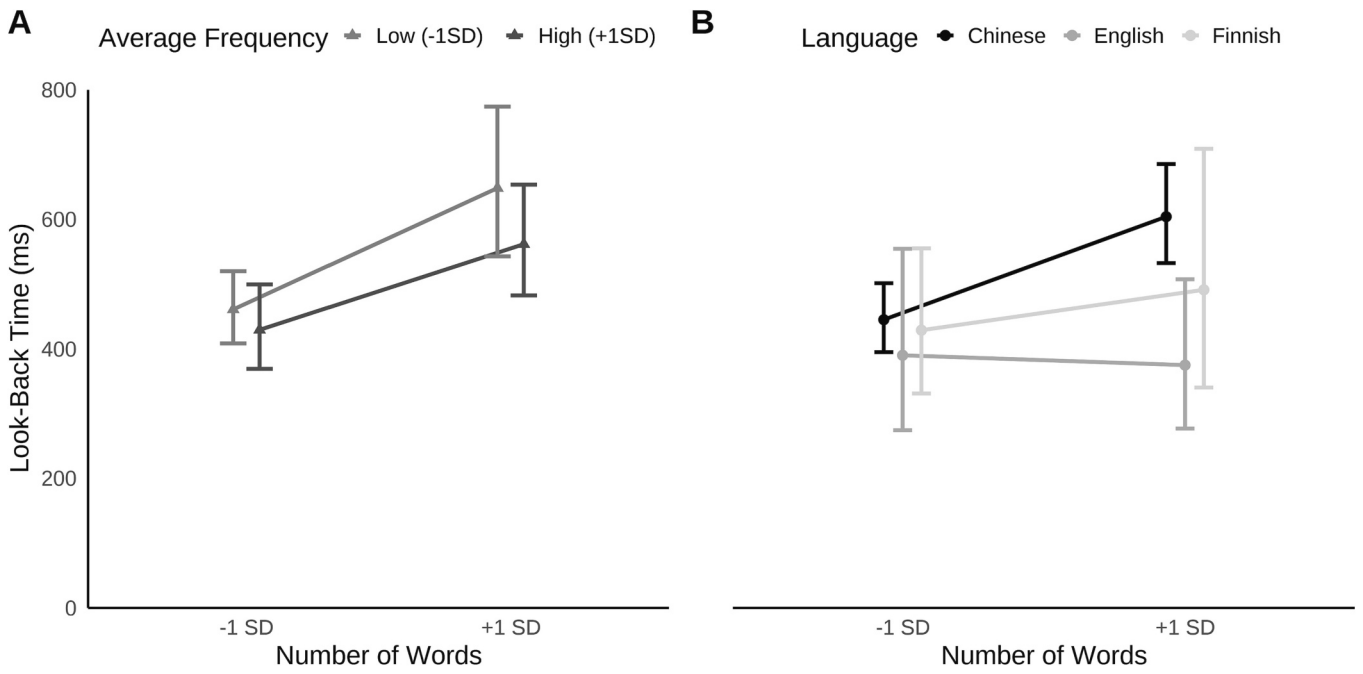

Fig. 5. Model estimates for look-back time. Panel A represents an interaction between *number of words* and *average word frequency*. Panel B represents an interaction between language (English vs. Chinese) and *number of words*. Look-back times are back-transformed to ms from log-values. Number of words and average word frequency are divided into ±1 *SD* for illustrative purposes. Error bars represent 95% confidence intervals.

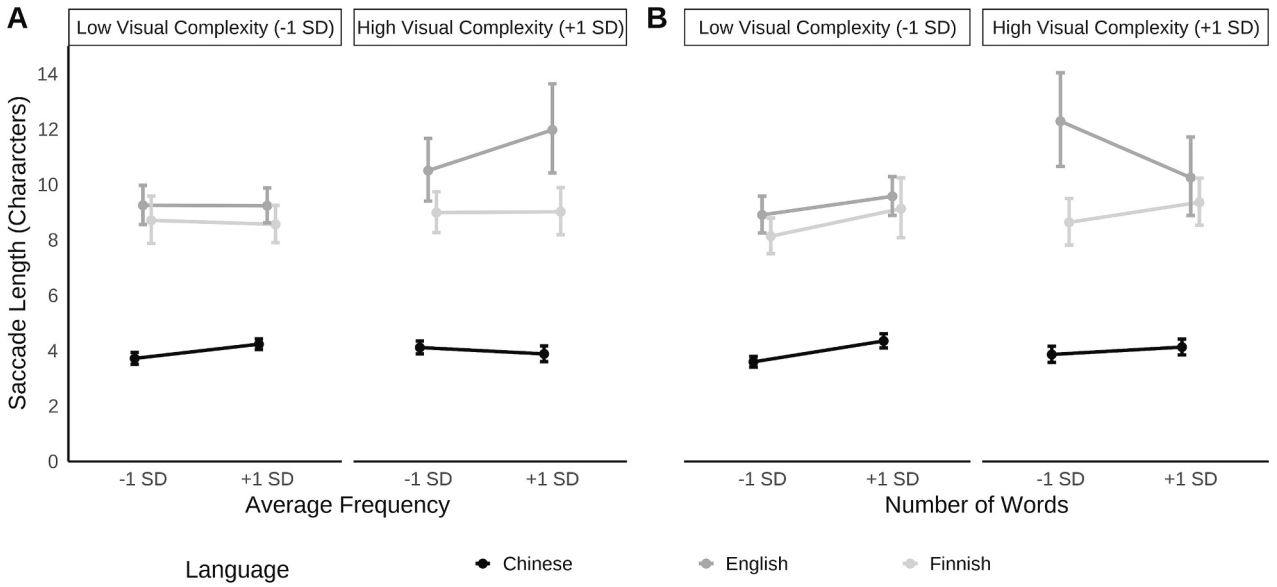

Fig. 6. Model estimates for rightward saccade length. Panel A represents an interaction between *language* (Chinese vs. English, and Finnish vs. English), *average word frequency* and *visual complexity*. Panel B represents an interaction between *language* (Finnish vs. English), *number of words*, and *visual complexity*. Rightward saccade length is back-transformed to characters from square-root values. Number of words, average frequency, and visual complexity are divided into ±1 *SD* for illustrative purposes. Error bars represent 95% confidence intervals.

main findings. First, rightward saccade length was affected by language with Chinese readers making the shortest forward saccades and English readers the longest forward saccades, even longer than Finnish readers. It was of no surprise that the average forward saccade length was half the size for Chinese readers compared to readers of the two alphabetic language. It merely reflects the visual density of the Chinese script. However, more surprising is the finding that English readers made longer saccades than Finnish readers. One might have expected an opposite pattern, as words in Finnish are longer than in English. Presumably, Finnish readers are more inclined to engage in refixations (see Table 3) because there are so many long words, whilst in English the tendency to refixate is less strong. Second, two interactions modified this overall difference. English readers lengthened their saccades when reading sentences containing long and frequent words and also when reading sentences containing few but long words. These effects likely reflect the fact that English texts contain lots of articles and prepositions. When there are fewer of them, forward saccades become longer provided the words are relatively frequent.

3.8. Average fixation duration

The model for average fixation duration (see Table A9) revealed main effects of *language* (English vs. Chinese) and *average word frequency*. Chinese readers made overall longer fixations than English and Finnish readers, who did not differ from each other. Sentences containing frequent words were read with shorter fixations than sentences containing infrequent words. Moreover, the model revealed two three-way interactions. First, there was an interaction between *language* (Finnish vs. English), *average word frequency* and *number of words*. Readers of all three languages increased their fixations when reading sentences containing many words (presumably responding to an increase in sentence complexity), with one exception. Finnish readers did not show this effect when reading long sentences containing many frequent words (Fig. 7A). It is not clear, why the effect is confined to Finnish readers. Second, there was an interaction between *language* (English vs. Chinese, and Finnish vs. English), *number of words*, and *visual complexity*. This interaction is primarily due to English readers increasing their average fixation duration in response to an increase in number of words, but only for sentences of high visual complexity (Fig. 7B). As speculated above, this may be due to English readers displaying a tendency to increase their fixation times on longer words rather than making a refixation on them.

In sum, as expected, Chinese readers made longer fixations than English or Finnish readers. The effect here very likely reflects differences in the visual density between logographic and alphabetic scripts. For the Chinese character-based orthography wherein most words are one or two characters long, density is high, whereas for English and Finnish that are alphabetic and have longer average word lengths, density is comparatively low. Average fixation duration was also longer when reading sentences containing infrequent than frequent words. This demonstrates that infrequent words take longer to recognize than frequent words. The result also demonstrates the robust cross-linguistic generality of this influence. The observed interactions suggest that English readers lengthen their fixations when reading sentences containing many long words. Similar to the effects in forward saccade length, this may reflect a relative absence of articles and prepositions in those sentences and possibly English readers' tendency to increase fixation duration instead of making a refixation. Finnish readers tended to differ from Chinese and English readers by shortening their fixations when reading sentences containing many frequent words, though the effects are small in size, and it is therefore sensible to treat them with some caution.

4. Discussion

The present study was a follow-up study of Liversedge et al. (2016). One of the main motivations for conducting it was the criticism regarding the statistical power in the original study (Brysbaert, 2019). In order to remedy it we increased the sample size to 303 subjects to test whether the effects obtained by Liversedge et al. replicated in a more representative sample. We were particularly interested in examining whether the null effect of language would replicate in total fixation time. For this purpose we made use of recent advances in statistical modeling, including Bayesian models. Moreover, we built common statistical models for all tested languages, English, Chinese and Finnish, to investigate how fundamental sentence charasteristics determine readers' eye movement patterns across languages. Here we departed from the analysis plan of Liversedge et al., who built separate models for Chinese and the two alphabetic languages. This was motivated by the fact that word length has limited variability in Chinese (1–4 characters), whereas it varies widely in alphabetic languages. Liversdge et al. pursued this route in their analyses because Chinese characters do not correspond to letters in alphabetic languages. Thus, it was not feasible to use average word length in sentences as a common measure across languages. In the present study, we circumvented this apples-and-pears issue with the measure of visual complexity (number of strokes needed to write a word regardless of whether the word was alphabetic or character based). Finally, to obtain a more detailed picture of the time course of sentence processing we decomposed the total fixation time measure to three different components.

We obtained a number of important results in our study. First and foremost, the total sentence reading time results for Chinese, Finnish and English were not the same and this finding differed from that reported by Liversedge et al. In the current results, total sentence reading times were shorter for Chinese sentences compared with Finnish and

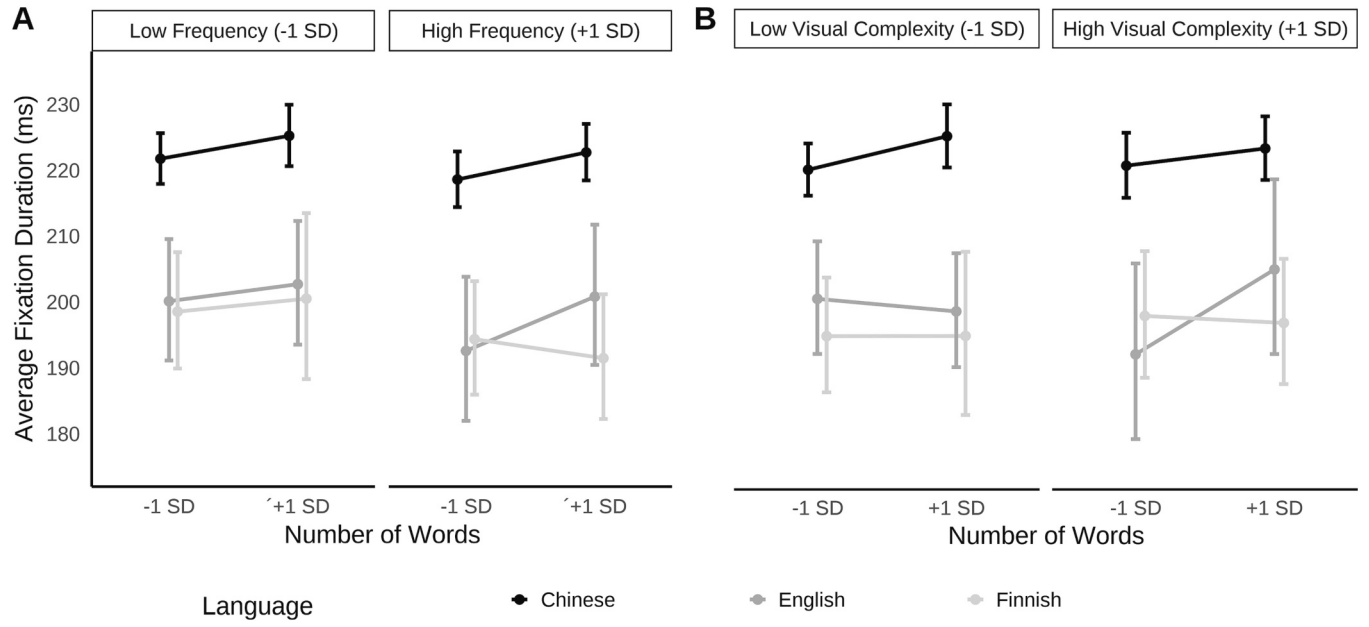

Fig. 7. Model estimates for average fixation duration. Panel A represents an interaction between *language* (Finnish vs. English), *average word frequency* and *number of words*. Panel B represents an interaction between *language* (English vs. Chinese and Finnish vs. English), *number of words*, and *visual complexity*. Average fixation duration was back-transformed to ms from log-values. Number of words, average word frequency and visual complexity are divided into ±1 *SD* for illustrative purposes. Error bars represent 95% confidence intervals.

English sentences which themselves did not differ. We consider this inconsistent result in detail below. Importantly, however, the majority of remaining aspects of the results were consistent with those reported by Liversedge et al. A variety of sentence level reading metrics were influenced by basic lexical characteristics of words (frequency, number of words or visual complexity) and most of these effects replicated the effects (or, in respect of visual complexity, complemented the effects) reported by Liversedge et al. To this extent, the current results extend the existing findings and provide additional support for the suggestion that these factors exert influence on eye movements in reading across languages. The current results also reflect a balance in the processing loads associated with the visual and linguistic processing systems such that the relationships between saccade extent and processing time indices were related to the orthographic density of the particular language under consideration. Finally, there were a number of other complex interactive effects across languages that we also consider in our detailed discussion below.

4.1. The Universality of Reading

Does the present study provide evidence for the universality of reading? This was the central theoretical question we set out to address in both the Liversedge et al. and the present study. Recall in the Introduction, we considered universality in relation to many aspects of visual and cognitive processes underlying reading. One of the aspects that was important was to examine total fixation time in reading sentences in expository texts in logographic (Chinese) and alphabetic (English and Finnish) scripts as a proxy of the ease of reading. And note that total sentence reading time is a natural unit of analysis in relation to this question, as in the present study each sentence expressed the same meaning across the three scripts. As Liversedge et al. (2016) observed no overall difference in total sentence reading time between the three languages, they put forth the notion of universality of processing. It should be borne in mind that the universality principle does not deny the fact that at a micro-level reading behavior is bound to demonstrate script-specific differences (for a review of universal and script-specific reading mechanisms, see Li, Huang, Yao, & Hyönä, 2022). To illustrate, average fixation duration was longer and average saccade length shorter when reading a more visually dense logographic script than a less visually dense alphabetic script. The universality principle holds that despite these differences at the micro level, the overall time to extract and represent the basic meaning of a comparable sentence across languages should be similar. One of the main findings of the present study is that when we used a larger sample size our results partly failed to replicate the lack of difference in overall sentence reading time across languages. Chinese was read with shorter sentence reading times than the two alphabetic languages. When converting total sentence fixation times to words-per-minute (wpm) values, Chinese subjects read the texts with a rate of 296 wpm, while English subjects produced a reading rate of 258 wpm and Finnish subjects a reading rate of 198 wpm. The overall reading rate observed for English readers compares favorably with the one estimated by Brysbaert (2019) in his meta-analysis of 190 reading studies. Brysbaert estimated the silent reading rate of English to be 260 wpm for reading fiction and 238 wpm for reading non-fiction.[2] The finding that the overall rate observed in the present study is more comparable to the rate estimated for fiction may be explained by the fact that the popular science texts were purposefully selected to be relatively easy to comprehend.

Bayesian modeling revealed no difference in total sentence reading time between English and Finnish. Also the difference in reading rate between English and Finnish may be considered more apparent than real. It reflects the fact that Finnish lacks articles and many prepositions are expressed as suffixes resulting in Finnish texts containing fewer words than English texts (see Table 2). This difference can be taken into consideration using an expansion/contraction index (Brysbaert, 2019), which indicates how many words are needed to translate a 1000-word English text to a given language. When the observed reading rate for Finnish is adjusted by this index, the reading rate of Finnish is practically identical (259 wpm) to that of English. On the other hand, the adjustment changes the observed reading rate of Chinese only minimally (302 wpm). Yet, it is considerably higher than that estimated by Brysbaert for Chinese based on 18 studies (260 wpm vs. 296 wpm observed in the present study).

The relatively faster reading rate observed for Chinese readers in the present study compared to the original study might be due to several reasons. First, note that the Chinese subjects originally tested in the Liversedge et al. (2016) study were tested in October 2008, whereas the current Chinese subjects were tested in April and November 2019. That is to say, the second set of subjects were tested over a decade later than were the original subjects. It is also important to note that since 2017, the Ministry of Education of China implemented a series of innovations to the National College Entrance Examination, known as Xin Gao Kao. Any individual wishing to become a student in higher education in China is required to take the National College Entrance Examination. A very significant change to this examination was made in relation to the Chinese language component of it, with the change being aimed at supporting the development of Chinese language proficiency (http://www.moe.gov.cn/jyb_xwfb/moe_2082/zl_2016n/2016_zl50/201610/t20161014_284877.html). Specifically, in the examination, the number of characters that prospective students were required to read increased from 7000 to 9000, and the number of examination questions also increased by 5–8%. Very importantly, there was no corresponding increase in the period of time allowed for the examination to be taken. Consequently, after 2017, students wishing to pursue higher education in China were required to read more words in the same time period, that is, were required to read faster when undertaking their assessment. In this way, the change in policy and assessment might have promoted an increased amount and increased speed of reading in students wishing to pursue a university degree in China. It is important to note two further points; first, the subjects tested in the original Liversedge et al. study were university students studying at Tianjin Normal University (TNU), and had therefore taken the National College Entrance Examination; second, these subjects were tested prior to 2017 when the changes to the National College Entrance Examination were introduced. Furthermore, the new additional Chinese subjects tested in current study were tested after the changes to the National College Entrance Examination. In addition to this important national change, in the period between Chinese data acquisition in the original Liversedge et al. study and the experimental testing for the current study, the degree programmes in Psychology at Tianjin Normal University have attained “First-Tier” status, that is, the Psychology degree programmes have become recognized as key disciplines in the Higher Education Admission System, which means the current psychology students admitted to these degree programmes at TNU very likely will read faster and perform better than those recruited to Psychology in previous years prior to “First-Tier” recognition of the degree programmes. Thus, on the basis of these two differences, one to entrance examination requirements and the other to the educational attainment of the students selected for the degree programmes from which experimental subjects were recruited, it seems likely that students' overall reading speed, and their efficiency in computing basic sentential meaning might have been increased. If this account is correct, then it appears that social factors rather than differences in the nature of orthographies between languages contributed to the pattern of effects reported in the current study.

[2] One further point of note in respect of the English total sentence reading time results is the degree of similarity between the current results and those of Liversedge et al. (2016). Here we obtained mean total reading times of 3232, whilst Liversedge et al. reported total reading times of 3093 (mean diference = 139 ms). These results suggest a degree of reliability, at least in respect of total sentence reading times, across independent samples (c.f., Staub, 2021).

In following this line of explanation, the next question we must now address concerns whether these results undermine the account we developed previously regarding universality in reading across different languages with different orthographies. From our perspective, the answer to this question is "Perhaps, but not completely". Recall that Liversedge et al. in the original paper considered three aspects of their results to evidence universality of process across languages. As we have described above, one piece of evidence was the finding that total sentence reading times did not differ between Chinese, Finnish and English. Clearly, the present results undermine this suggestion, and if our explanation of the faster Chinese reading times is correct, then it appears that rate at which individuals in a particular culture process written text can be modulated by social factors. To this extent, then, the comparability of total sentence reading times does not necessarily reflect universality of process. However, Liversedge et al. also considered universality in relation to the basic lexical characteristics of words (frequency, number of words or visual complexity) and how these variables exert influences that are comparable across languages. Completely in line with the findings of Liversedge et al., the present results once again showed that reading times were affected by such lexical characteristics. Total sentence reading times increased as the number of words and their visual complexity increased, whereas it decreased as average word frequency in sentences increased. To reiterate, the findings for frequency and number of words replicate perfectly those reported by Liversedge et al. and reinforce the claim that both these factors are primary lexical characteristics that affect eye movements in reading similarly across languages. Our new index of visual complexity patterns similarly, suggesting that it too represents a basic visual characteristic that captures variance in eye movement behavior similarly across languages with very different orthographies. To this extent, the present findings like those in the orginal study do offer some support for claims of universality in respect of word representations that are central to reading processing across languages.

Finally, as we have argued in the Introduction, in order to read efficiently the visual and linguistic processing systems have to operate with synchrony such that the visual system delivers visual information in a timely way to the linguistic processing system, with the linguistic processing system itself incrementally interpreting text. Again, as Liversedge et al. originally argued, the view that informational exchange between the visual encoding system and the linguistic processing system is something of a balance may itself reflect universality in the reading process. That is, the rate of information transfer from one system to the other may be a critical determinant of the rate of processing. Clearly, there may exist differences in the rate of such exchange between visual and linguistic processing systems across languages. However, the informational exchange process itself does appear to be universal and does appear to operate as a stricture with respect to eye movements during reading in different languages. Specifically, Chinese readers generally make longer fixations, along with shorter rightward saccades, because Chinese is a densely packed orthography with limited horizontal spatial layout. Consequently, shorter saccadic eye movements are sufficient to efficiently deliver information from the visual encoding system to the linguistic processing system. However, during each fixation, due to the linguistic density of Chinese, a substantial amount of linguistic information is delivered, and necessarily, the linguistic processing system has to proceed relatively slowly in order for sentential meaning to be computed. In contrast, for a relatively sparse alphabetic orthography such as Finnish, where spoken language is mapped onto written language with perfect phoneme-grapheme correspondence, readers generally make shorter fixations, along with longer rightward saccades. They do this because the visual encoding system must work harder to recruit and deliver orthographic information in order to keep up with a very "hungry" linguistic processing system. As with the original findings of Liversedge et al., the current results, to us, compellingly demonstrate that the linguistic density of an orthography modulates the balance in information exchange between the visual and linguistic processing systems. And, as we originally suggested, we consider that this situation represents a universal aspect of processing, namely, information representation and exchange during reading.

4.2. The Time Course and Nature of Reading

As noted earlier, one aim of the present study was to analyse in more detail the time course of processing sentences during text reading. In order to do so, we decomposed the total fixation time measure into first-pass forward fixation time, first-pass rereading time and look-back time. The first-pass forward fixation time indexes early effects, whereas look-back time reflects delayed effects. In this section, we examine how language differences emerged in the component measures. The pattern of results is summarized in Table 4.

As can be seen from Table 4, the Chinese subjects read the texts with shorter total fixation times than the English and Finnish subjects. This language difference was noticeable even in the earliest measure (first-pass forward fixation time) indexing the first encounter with sentence content. On the other hand, readers of different languages did not differ in the amount of first-pass rereading they carried out. However, very interestingly, Chinese readers spent more time looking back to previous sentences than either the English or Finnish readers (we return to this effect below). Even so, this difference did not completely remove the overall difference in total fixation time between Chinese readers and readers of the two alphabetic languages. It appears that the Chinese readers push through the text to initially read the text quickly. We suggest that this might be due to the encouragement that they have received to read rapidly (due to policy stipulations). However, in striving to read quickly, it is possible that the Chinese readers might have not initially read quite so carefully meaning that they subsequently need to spend time re-reading text to ensure that their interpretation is entirely appropriate.

The measured sentence-level characteristics of word frequency, number of words and visual complexity also affected the time course of sentence processing across the three languages. All these main effects were observed in first-pass forward fixation time, suggesting that their effects were immediate. Number of words and visual complexity similarly also affected first-pass rereading time with the effect of average word frequency re-emerging in look-back time. Finally, number of words and visual complexity ceased to affect delayed sentence processing, as indexed by look-back time. As all the measured sentence features index variation in sentence complexity (i.e., number of words

Table 4
Summary of the effects obtained for the different sentence fixation time measures. (an x in the table signifies that an effect was robust).

Effect	Total fixation time	First-pass forward	First-pass rereading	Look-back time
Number of words	x	x	x	
Visual complexity	x	x	x	
Average word frequency	x	x		x
Number of words x word frequency	x	x		x
Number of words x visual complexity		x		
Language	CHI < UK = FIN	CHI < UK = FIN		CHI > UK = FIN
Language x number of words	FIN > CHI = UK	CHI < UK = FIN		CHI > UK = FIN
Language x word frequency		FIN > CHI = UK		
Language x number of words x word frequency	CHI < UK = FIN			

and their frequency and visual complexity all contribute to overall sentence complexity), it is perhaps not surprising that their effects emerged during first-pass reading. On the other hand, only the word frequency effect lingered beyond first-pass reading, whereas effects of sentence length and words' visual complexity (i.e., approximating average word length in English and Finnish and number of strokes in words in Chinese) ceased to affect later processing. This pattern may be interpreted to suggest that readers' look-back behavior is affected by lexical and conceptual difficulty (i.e., linguistic processing demands), indexed by average word frequency in sentences, but not by mere visual complexity (i.e., visual processing demands). On the other hand, number of words and visual complexity caused subjects to go back in text and at least partly reread the sentences as they were working through them for the first time. The conclusion that look-back behavior is affected by lexical and conceptual difficulty is further corroborated by the interaction between average word frequency and number of words in sentences that we obtained for look-back time. It suggests that subjects are especially prone to look back and re-read sentences containing many infrequent words. It may be noted that this interaction was also observed as a more immediate effect indexed by first-pass forward fixation time.

Next, we discuss how the basic lexical characteristics of the words in the sentences (average word frequency, number of words and words' average visual complexity) interacted with language. Table 4 summarizes these effects. First, total fixation times demonstrated that Finnish readers were more strongly affected by the number of words in sentences than English or Chinese readers. This makes sense considering the fact that words in Finnish are somewhat longer than words in English or Chinese (see Table 2). Thus, with each additional word, the change to the physical extent of the sentences was more profound in Finnish than Chinese or English, resulting in more robust effects of the number-of-words in a sentence. As argued above, it likely reflects typological differences between synthetic and analytic languages. In synthetic languages like Finnish words carry more information due to words often being morphologically complex. Thus, an increase in number of words is likely to lengthen reading times more than in analytic languages where words are morphologically less complex.

Second, an interaction between language and number of words was also observed in first-pass forward fixation time. Here the effect was such that Chinese readers were less affected by the number of words than readers of alphabetic scripts. The effect reflects the fact in Chinese words typically contain one or two characters, making the sentences visually much shorter than in alphabetic languages. This is also why Chinese readers skipped over more than half of the words during first-pass reading, whereas the word skipping rate was much smaller for English and particularly for Finnish (see Table 3).[3] The effect was reversed in look-back time: in this measure, Chinese readers were more affected by the number of words than that of readers of alphabetic scripts. To sum up, with Chinese words occupying less space, Chinese readers are able to progress through the sentences faster than readers of the two alphabetic languages. Yet, they need to pay a small price for that later, as revealed by their increased look-backs especially to sentences containing many words. As we mentioned earlier, this pattern of results might tie into our consideration of why in the present study Chinese readers, overall, had shorter sentence reading times than Finnish or English readers. Again, perhaps Chinese readers are initially pushing through sentences quite quickly, as suggested by the short first-pass times, though they also spend time reinspecting text later to ensure that they have formed an appropriate interpretation.

Third, Finnish readers tend to show in first-pass forward fixation time a larger word frequency effect than English or Chinese readers. One possible explanation is that as words in Finnish are overall longer, they often need two fixations to be recognized, which may be less likely the case when they are frequent. This may result in a stronger word frequency effect in reading Finnish. As is apparent from Fig. 2c, the reduction in first-pass forward fixation time for sentences containing many frequent words was particularly noticeable in Finnish. Finally, total fixation time also yielded a three-way interaction between language, number of words and average word frequency. The pattern of effects suggests that Chinese readers increased their total fixation times in response to an increase in number of words to a similar degree when the words were infrequent and frequent. On the other hand, English and Finnish subjects demonstrated a more robust number-of-words effect when the sentences contained infrequent than frequent words. This effect could not be reliably located in any component measure of total fixation measure, but a trend was observable in all of them. A possible explanation may again be ascribed to the visual density of the Chinese script: As words occupy little space, word processing is facilitated (many words are recognized without fixating them), also including more infrequent words.

4.3. The implications of the Visual complexity measure

The present analysis of total fixation time differed from that of Liversedge et al. (2016) in one important way. Instead of using average word length as one sentence feature, we replaced it with a measure of words' visual complexity as indexed by the number of strokes needed to write words. Both measures have their shortcomings. Average word length works fine with English and Finnish, where there is considerable variability in word length, but it is less optimal for Chinese where variability is significantly reduced, as words vary from 1 to 4 characters, the majority of words comprising only one or two characters. On the other hand, the number of strokes measure works fine with Chinese; the number of strokes in a character varies from 1 to 36 for the 7000 frequently used Chinese characters (Zang, Liversedge, Bai, & Yan, 2011). Yet, it has the problem of meaning somewhat different things for Chinese versus English and Finnish (Liversedge et al. referred to it as the dilemma of comparing apples with pears). In Chinese, it indexes the visual density of strokes in characters, whereas in English and Finnish it has a practically one-to-one correspondence with the number of letters (the more strokes, the longer the word). In other words, in alphabetic languages it indexes the sentences' spatial extent, whereas in Chinese it does so to a much smaller extent.

Despite the shortcoming of the visual complexity measure, it has the significant advantage of allowing us to draw direct comaprisons and construct formal statistical models of data from all three languages. Indeed, through the direct comparison of the three languages, we are better able to understand comprehensively how sentence-level characteristics of word frequency, number of words and visual complexity affect the time course of processing across the three languages, and how these different variables might exert different influences across languages. For example, it is through such direct comparison that we can see that compared with readers of alphabetic scripts, Chinese readers were less affected by the number of words in first-pass forward fixation time, but more affected in look-back time. And in turn, this leads us to suggest that Chinese readers appear to undertake initial fast reading followed by an amount of verification processing. Such insight would not be possible without direct comparison between Chinese and

[3] We are grateful to Victor Kuperman for drawing our attention to the fact that the skipping rates in the present study for Chinese and the skipping rates for Korean reported in Siegelman et al., (2022) (MECO study) allow us to compare the prevalence of such behavior in a character based vs. an alphabetic language with a quite similar orthographic form. Skipping rates were 0.57 for Chinese and 0.32 for Korean. The difference between these skipping rates is quite considerable. Furthermore, in this context, skipping rates for Finnish (present study = 0.23, MECO = 0.17) and for English (present study = 0.35, MECO = 0.3) between the two studies were quite comparable. Together these findings suggest that whether a written language is character based or alphabetic might itself modulate the degree to which word skipping occurs during reading. However, without a comprehensive analysis of directly comparable stimuli across these two languages, it is hard to be certain of the cause of this difference. Further research is required to better understand this matter.

alphabetic reading. This reinforces our decision to adopt this approach here. It is also noteworthy that the selected sentence features entered in the statistical models did an excellent job in modeling the Chinese data, as indicated by the much smaller confidence intervals in many of the measures for Chinese than for the two alphabetic languages (see Figs. 1, 2 and 4). In contrast, when average word length was used in the models instead of visual complexity (not reported here), confidence intervals for Chinese were substantially larger than for the two other languages. This is a further reason why we considered it justified to use the visual complexity measure in our modeling work. To reiterate, its advantage was that it made possible direct comparison of the three languages in the same statistical models.

4.4. Increased power and Bayesian evaluation of the null

As discussed earlier, the central theoretical claim put forward by Liversedge et al. relied on a null effect in total sentence reading times across languages and the absence of such an effect was considered to reflect universality associated with the computation of basic sentential meaning. However, the weakness of the original study was that it reported a single ANOVA analysis of the total sentence reading time. Liversedge et al. did not undertake LMM analyses for this critical assessment. LMM analyses are acknowledged to be more powerful than ANOVA analyses, assessing every observation of each dependent variable and capturing both subject and item variance. Furthermore, Liversedge et al. did not calculate power analyses, nor did they undertake Bayseian analyses to evaluate evidence in favor of the null effect. The present study has addressed these empirical weakenesses directly. The simulation-based power analyses showed that to attain 80% power, 21–50 subjects per language were required to detect an effect of language, and 13 subjects per language were required to detect a three-way interaction between language, the number of words in a sentence and average word frequency. Clearly this number is not far from the number of subjects that were tested in the original study. The question now arises as to whether power concerns really were a serious concern in respect of the original study. In fact, we agree that the study was slightly under-powered, and we acknowledge this concern; however, we also feel that it should not be overstated. One conclusion that we feel we are able to form from the current study is that it is important to carry our power analyses that are based on reasonable assumptions in respect of effect sizes and item numbers (as well as subjects numbers) when assessing levels of power. Failure to consider the contribution of both subjects and items leads to an overestimation of the number of subjects that should be tested to attain adequate levels of statistical power.

5. Conclusion

The present study is an effort to deliver a more compelling version of the study reported by Liversedge et al. We have tested a substantially larger number of subjects in each language and we have engaged seriously with criticisms levelled at the original paper in relation to issues of power. We have computed a more comprehensive set of reading time measures from our eye movement data sets and we have conducted more sophisticated statistical analyses of our data. Finally, we have adopted a cross-linguistic comparability index (visual complexity) that permitted us to make direct statistical comparisons across all three languages. All of these aspects of the current study go beyond that reported originally by Liversedge et al. In relation to our results, it is certainly the case that one particular, quite central, aspect of our study was not replicated – total sentence reading time differences emerged across languages and we have provided a socio-cultural account for this effect. Importantly, it was also the case that many aspects of the results of the current study did replicate those of the original study, and in our view, the degree of replication leads us to maintain our view that there are aspects of reading that are universal across languages.

CRediT authorship contribution statement

Simon P. Liversedge: Conceptualization, Methodology, Writing – original draft. **Henri Olkoniemi:** Investigation, Formal analysis. **Chuanli Zang:** Methodology, Writing – review & editing. **Xin Li:** Investigation. **Guoli Yan:** Supervision. **Xuejun Bai:** Supervision. **Jukka Hyönä:** Funding acquisition, Supervision, Conceptualization, Writing – review & editing.

Data availability

All data sets and analysis scripts are publicly available at the online repository https://osf.io/efqnx/

Acknowledgement

We acknowledge support from ESRC Grant (ES/R003386/1) and the Academy of Finland (315963).

Appendix A. Final models

Table A1 Fixed effect values for the model of total fixation time with the original subjects from the Liversedge et al. (2016) study.

Fixed effects	β	95% CI	t
Intercept	7.86	[7.72, 7.94]	126.26***
Language (English vs. Chinese)	−0.08	[0.13, 0.30]	−0.96
Language (Finnish vs. English)	−0.18	[−0.17, −0.01]	−2.13*
Random effects	*n*	*Variance*	*SD*
Sentence (Intercept)	119	0.32	0.57
Sentence (Language: English vs. Chinese)		0.02	0.16
Sentence (Language: Finnish vs. English)		0.01	0.12
Subject (Intercept)	59	0.07	0.26
Residual		0.10	0.32
Observed Total Fixation Times			
Language	*M*	*SD*	
Chinese	3657	2622	
English	3286	2225	
Finnish	2863	2139	

* = $p < .05$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A2 Fixed effect values of the model on response accuracy of the text comprehension questions.

Fixed effects	Odds Ratios	95% CI	z
Intercept	12.80	[6.74, 24.30]	7.79***
Language (English vs Chinese)	0.60	[0.35, 1.04]	−1.83
Language (Finnish vs English)	0.96	[0.49, 1.87]	−0.12
Random effects	*n*	*Variance*	*SD*
Text (Intercept)	22	1.45	1.21
Text (Chinese)		0.03	0.18
Text (English)		0.28	0.53
Text (Finnish)		0.67	0.82
Subjects (Intercept)	284	0.48	0.69

*** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method. The data are missing for response accuracy of the Chinese subjects participating in the Liversedge et al. (2016) study (here $N_{Chinese} = 80$).

Table A3 Fixed effect values for the simple model of total fixation time.

Fixed effects	β	95% CI	t
Intercept	7.83	[7.72, 7.94]	142.73***
Language (English vs. Chinese)	0.22	[0.13, 0.30]	5.28***
Language (Finnish vs. English)	−0.09	[−0.17, −0.01]	−2.27*
Random effects	*n*	*Variance*	*SD*
Sentence (Intercept)	119	0.32	0.57
Subject (Intercept)	303	0.08	0.29
Residual		0.12	0.35

* = $p < .05$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A4 Complex model for total fixation time.

Fixed effects	β	95% CI	t
Intercept	7.86	[7.80, 7.92]	250.85***
Language (English vs. Chinese)	0.37	[0.26, 0.48]	6.81***
Language (Finnish vs. English)	−0.07	[−0.18, 0.04]	−1.28
Average Frequency	−0.04	[−0.07, −0.02]	−3.04**
Number of Words	0.44	[0.39, 0.49]	17.70***
Visual Complexity	0.15	[0.12, 0.19]	8.26***
Average Frequency × Number of Words	−0.03	[−0.06, −0.01]	−2.45*
Average Frequency × Visual Complexity	0.02	[−0.01, 0.04]	1.27
Number of Words × Visual Complexity	0.01	[−0.03, 0.05]	0.39
Language (English vs. Chinese) × Average Frequency	−0.02	[−0.08, 0.05]	−0.53
Language (Finnish vs. English) × Average Frequency	−0.06	[−0.12, 0.01]	1.80
Language (English vs. Chinese) × Number of Words	−0.02	[−0.10, 0.05]	−0.60
Language (Finnish vs. English) × Number of Words	0.12	[0.03, 0.21]	2.72**
Language (English vs. Chinese) × Visual Complexity	0.01	[−0.08, 0.09]	0.14
Language (Finnish vs. English) × Visual Complexity	−0.02	[−0.09, 0.04]	−0.71
Language (English vs. Chinese) × Average Frequency × Number of Words	−0.05	[−0.10, −0.01]	−2.49*
Language (Finnish vs. English) × Average Frequency × Number of Words	−0.02	[−0.07, 0.03]	−0.77
Language (English vs. Chinese) × Average Frequency × Visual Complexity	0.003	[−0.07, 0.07]	0.07
Language (Finnish vs. English) × Average Frequency × Visual Complexity	0.003	[−0.05, 0.05]	0.10
Language (English vs. Chinese) × Number of Words × Visual Complexity	−0.04	[−0.14, 0.05]	−0.90
Language (Finnish vs. English) × Number of Words × Visual Complexity	0.06	[−0.01, 0.14]	1.67
Random effects	*n*	*Variance*	*SD*
Subjects (Intercept)	303	0.08	0.29
Subjects (Number of Words)		0.0004	0.02
Subjects (Visual Complexity)		0.0005	0.02
Sentence (Intercept)	119	0.07	0.26
Sentence (Language: English vs. Chinese)		0.04	0.20
Sentence (Language: Finnish vs. English)		0.02	0.12
Residual		0.11	0.34

* = $p < .05$, ** = $p < .01$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A5 Complex model for first-pass forward-fixation time.

Fixed effects	β	95% CI	t
Intercept	45.67	[44.70, 46.64]	92.03***
Language (*English* vs. *Chinese*)	10.87	[8.65, 13.09]	9.60***
Language (*Finnish* vs. *English*)	−0.89	[−3.19, 1.40]	−0.76
Average Frequency	−0.85	[−1.39, −0.31]	−3.10**
Number of Words	11.10	[10.40, 11.79]	31.30***
Visual Complexity	3.02	[2.39, 3.66]	9.32***
Average Frequency × Number of Words	−0.58	[−1.01, −0.14]	−2.61**
Average Frequency × Visual Complexity	0.47	[−0.005, 0.95]	1.96

(*continued on next page*)

(*continued*)

Fixed effects	β	95% CI	t
Number of Words × Visual Complexity	1.16	[0.45, 1.87]	3.19**
Language (*English* vs. *Chinese*) × Average Frequency	−0.34	[−1.60, 0.92]	−0.53
Language (*Finnish* vs. *English*) × Average Frequency	−1.42	[−2.76, −0.09]	−2.09*
Language (*English* vs. *Chinese*) × Number of Words	3.20	[1.69, 4.71]	4.16***
Language (*Finnish* vs. *English*) × Number of Words	1.33	[−0.41, 3.06]	1.49
Language (*English* vs. *Chinese*) × Visual Complexity	1.15	[−0.46, 2.76]	1.39
Language (*Finnish* vs. *English*) × Visual Complexity	−0.34	[−1.75, 1.07]	−0.47
Language (*English* vs. *Chinese*) × Average Frequency × Number of Words	−0.78	[−1.58, 0.03]	−1.89
Language (*Finnish* vs. *English*) × Average Frequency × Number of Words	−0.82	[−1.76, 0.13]	−1.70
Language (*English* vs. *Chinese*) × Average Frequency × Visual Complexity	0.39	[−0.95, 1.72]	0.57
Language (*Finnish* vs. *English*) × Average Frequency × Visual Complexity	−0.44	[−1.58, 0.70]	−0.76
Language (*English* vs. *Chinese*) × Number of Words × Visual Complexity	0.06	[−1.72, 1.84]	0.07
Language (*Finnish* vs. *English*) × Number of Words × Visual Complexity	−0.33	[−1.96, 1.30]	−0.40
Random effects	*n*	*Variance*	*SD*
Subjects (Intercept)	303	37.74	6.14
Sentence (Intercept)	119	7.88	2.81
Sentence (Language: *English* vs. *Chinese*)		11.67	3.42
Sentence (Language: *Finnish* vs. *English*)		6.78	2.60
Residual		34.49	5.87

* = $p < .05$, ** = $p < .01$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A6 Complex model for first-pass rereading time.

Fixed effects	β	95% CI	t
Intercept	6.23	[6.15, 6.31]	147.53***
Language (*English* vs. *Chinese*)	0.04	[−0.15, 0.22]	0.38
Language (*Finnish* vs. *English*)	−0.08	[−0.28, 0.11]	−0.83
Average Frequency	−0.05	[−0.11, 0.01]	−1.70
Number of Words	0.36	[0.29, 0.44]	9.32***
Visual Complexity	0.12	[0.05, 0.19]	3.30**
Average Frequency × Number of Words	−0.04	[−0.08, 0.01]	−1.51
Average Frequency × Visual Complexity	0.02	[−0.03, 0.08]	0.90
Number of Words × Visual Complexity	−0.02	[−0.10, 0.07]	−0.37
Language (*English* vs. *Chinese*) × Average Frequency	−0.03	[−0.17, 0.11]	−0.43
Language (*Finnish* vs. *English*) × Average Frequency	−0.07	[−0.23, 0.09]	−0.88
Language (*English* vs. *Chinese*) × Number of Words	−0.06	[−0.23, 0.11]	−0.71
Language (*Finnish* vs. *English*) × Number of Words	0.07	[−0.13, 0.27]	0.70
Language (*English* vs. *Chinese*) × Visual Complexity	−0.05	[−0.22, 0.13]	−0.55
Language (*Finnish* vs. *English*) × Visual Complexity	0.003	[−0.16, 0.17]	0.03
Language (*English* vs. *Chinese*) × Average Frequency × Number of Words	−0.06	[−0.15, 0.02]	−1.42
Language (*Finnish* vs. *English*) × Average Frequency × Number of Words	−0.03	[−0.15, 0.09]	−0.42
Language (*English* vs. *Chinese*) × Average Frequency × Visual Complexity	0.08	[−0.07, 0.23]	1.08
Language (*Finnish* vs. *English*) × Average Frequency × Visual Complexity	−0.04	[−0.17, 0.09]	−0.60
Language (*English* vs. *Chinese*) × Number of Words × Visual Complexity	−0.06	[−0.25, 0.13]	−0.60
Language (*Finnish* vs. *English*) × Number of Words × Visual Complexity	0.08	[−0.11, 0.27]	0.83
Random effects	*n*	*Variance*	*SD*
Subjects (Intercept)	303	0.12	0.35
Subjects (Average Frequency)		0.001	0.04
Subjects (Number of Words)		0.01	0.09
Subjects (Visual Complexity)		0.003	0.05
Sentence (Intercept)	119	0.08	0.27
Sentence (Language: *English* vs. *Chinese*)		0.11	0.34
Sentence (Language: *Finnish* vs. *English*)		0.11	0.33
Residual		0.58	0.76

* = $p < .05$, ** = $p < .01$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A7 Complex model for look-back time.

Fixed effects	β	95% CI	t
Intercept	6.11	[6.02, 6.19]	138.33***
Language (*English* vs. *Chinese*)	−0.30	[−0.48, −0.13]	−3.37***
Language (*Finnish* vs. *English*)	0.18	[−0.03, 0.39]	1.68
Average Frequency	−0.07	[−0.14, −0.01]	−2.11*
Number of Words	0.07	[−0.02, 0.16]	1.65
Visual Complexity	0.03	[−0.05, 0.10]	0.73
Average Frequency × Number of Words	−0.06	[−0.12, −0.002]	−2.02*
Average Frequency × Visual Complexity	0.03	[−0.02, 0.08]	1.07
Number of Words × Visual Complexity	−0.06	[−0.14, 0.03]	−1.34
Language (*English* vs. *Chinese*) × Average Frequency	0.01	[−0.12, 0.15]	0.21
Language (*Finnish* vs. *English*) × Average Frequency	−0.08	[−0.25, 0.10]	−0.83
Language (*English* vs. *Chinese*) × Number of Words	−0.17	[−0.34, −0.01]	−2.05*
Language (*Finnish* vs. *English*) × Number of Words	0.08	[−0.13, 0.30]	0.74

(*continued on next page*)

(*continued*)

Fixed effects	β	95% CI	t
Language (*English* vs. *Chinese*) × Visual Complexity	0.004	[−0.16, 0.17]	0.05
Language (*Finnish* vs. *English*) × Visual Complexity	0.005	[−0.18, 0.17]	−0.06
Language (*English* vs. *Chinese*) × Average Frequency × Number of Words	−0.07	[−0.15, 0.01]	−1.60
Language (*Finnish* vs. *English*) × Average Frequency × Number of Words	0.01	[−0.13, 0.16]	0.19
Language (*English* vs. *Chinese*) × Average Frequency × Visual Complexity	−0.01	[−0.15, 0.13]	−0.18
Language (*Finnish* vs. *English*) × Average Frequency × Visual Complexity	0.004	[−0.12, 0.13]	0.06
Language (*English* vs. *Chinese*) × Number of Words × Visual Complexity	−0.02	[−0.22, 0.17]	−0.24
Language (*Finnish* vs. *English*) × Number of Words × Visual Complexity	0.07	[−0.14, 0.27]	0.64
Random effects	*n*	*Variance*	*SD*
Subjects (Intercept)	303	0.12	0.34
Subjects (Average Frequency)		0.00000002	0.0001
Subjects (Number of Words)		0.01	0.09
Subjects (Visual Complexity)		0.01	0.10
Sentence (Intercept)	119	0.08	0.28
Sentence (Language: *English* vs. *Chinese*)		0.05	0.22
Sentence (Language: *Finnish* vs. *English*)		0.10	0.31
Residual		0.81	0.90

* = $p < .05$, ** = $p < .01$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A8 Complex model for rightward Saccade length.

Fixed effects	β	95% CI	t
Intercept	2.76	[2.72, 2.81]	128.70***
Language (*English* vs. *Chinese*)	1.17	[1.07, 1.27]	23.10***
Language (*Finnish* vs. *English*)	−0.28	[−0.38, −0.18]	−5.43***
Average Frequency	0.02	[−0.0002, 0.05]	1.94
Number of Words	0.03	[0.002, 0.06]	2.09*
Visual Complexity	0.06	[0.03, 0.09]	4.16***
Average Frequency × Number of Words	−0.01	[−0.03, 0.01]	−1.34
Average Frequency × Visual Complexity	0.002	[−0.02, 0.02]	0.18
Number of Words × Visual Complexity	−0.05	[−0.08, 0.02]	−2.91**
Language (*English* vs. *Chinese*) × Average Frequency	0.03	[−0.02, 0.09]	1.18
Language (*Finnish* vs. *English*) × Average Frequency	−0.06	[−0.11, −0.0004]	−1.98*
Language (*English* vs. *Chinese*) × Number of Words	−0.12	[−0.18, −0.06]	−3.72***
Language (*Finnish* vs. *English*) × Number of Words	0.12	[0.04, 0.19]	3.17***
Language (English vs. Chinese) × Visual Complexity	0.16	[0.08, 0.23]	4.02***
Language (*Finnish* vs. *English*) × Visual Complexity	−0.13	[−0.19, −0.07]	−4.38***
Language (*English* vs. *Chinese*) × Average Frequency × Number of Words	0.02	[−0.02, 0.06]	1.14
Language (*Finnish* vs. *English*) × Average Frequency × Number of Words	−0.003	[−0.04, 0.04]	−0.16
Language (*English* vs. *Chinese*) × Average Frequency × Visual Complexity	0.11	[0.04, 0.17]	3.34**
Language (*Finnish* vs. *English*) × Average Frequency × Visual Complexity	−0.05	[−0.09, −0.001]	−1.92
Language (*English* vs. *Chinese*) × Number of Words × Visual Complexity	−0.06	[−0.14, 0.02]	−1.56
Language (*Finnish* vs. *English*) × Number of Words × Visual Complexity	0.09	[0.03, 0.16]	2.74**
Random effects	*n*	*Variance*	*SD*
Pariticipants (Intercept)	244	0.07	0.26
Pariticipants (Visual Complexity)		0.001	0.03
Sentence (Intercept)	119	0.01	0.11
Sentence (Language: English vs. Chinese)		0.03	0.16
Sentence (Language: Finnish vs. English)		0.01	0.11
Residual		0.09	0.29

* = $p < .05$, ** = $p < .01$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A9 Complex model for average fixation duration.

Fixed effects	β	95% CI	t
Intercept	5.31	[5.29, 5.33]	620.21***
Language (*English* vs. *Chinese*)	−0.10	[−0.14, −0.05]	−4.62***
Language (*Finnish* vs. *English*)	−0.03	[−0.08, 0.01]	−1.63
Average Frequency	−0.01	[−0.02, −0.005]	−3.33**
Number of Words	0.01	[−0.002, 0.01]	1.52
Visual Complexity	0.001	[−0.01, 0.01]	0.36
Average Frequency × Number of Words	−0.0002	[−0.01, 0.01]	−0.09
Average Frequency × Visual Complexity	0.004	[−0.002, 0.01]	1.30
Number of Words × Visual Complexity	0.004	[−0.01, 0.01]	0.78
Language (*English* vs. *Chinese*) × Average Frequency	−0.002	[−0.02, 0.01]	−0.29
Language (*Finnish* vs. *English*) × Average Frequency	−0.01	[−0.02, 0.01]	−0.84
Language (*English* vs. *Chinese*) × Number of Words	0.002	[−0.01, 0.02]	0.26
Language (*Finnish* vs. *English*) × Number of Words	−0.01	[−0.03, 0.01]	−0.89
Language (*English* vs. *Chinese*) × Visual Complexity	0.003	[−0.02, 0.02]	0.33
Language (*Finnish* vs. *English*) × Visual Complexity	0.01	[−0.01, 0.02]	0.93
Language (*English* vs. *Chinese*) × Average Frequency × Number of Words	0.004	[−0.01, 0.01]	0.87
Language (*Finnish* vs. *English*) × Average Frequency × Number of Words	−0.01	[−0.03, −0.003]	−2.01*

(*continued on next page*)

(*continued*)

Fixed effects	β	95% CI	t
Language (*English* vs. *Chinese*) × Average Frequency × Visual Complexity	−0.002	[−0.02, 0.01]	−0.23
Language (*Finnish* vs. *English*) × Average Frequency × Visual Complexity	0.01	[−0.01, 0.02]	1.16
Language (*English* vs. *Chinese*) × Number of Words × Visual Complexity	0.01	[0.01, 0.03]	1.27
Language (*Finnish* vs. *English*) × Number of Words × Visual Complexity	−0.02	[−0.04, −0.002]	−1.72
Random effects	*n*	*Variance*	*SD*
Subjects (Intercept)	244	0.01	0.12
Subjects (Average Frequency)		0.00004	0.01
Subjects (Number of Words)		0.00003	0.01
Sentence (Intercept)	119	0.001	0.03
Sentence (Language: *English* vs. *Chinese*)		0.001	0.04
Sentence (Language: *Finnish* vs. *English*)		0.001	0.03
Residual		0.02	0.13

* = $p < .05$, ** = $p < .01$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

Table A10 Model for total fixation time comparing original and new datasets.

Fixed effects	β	95% CI	t
Intercept	7.84	[7.73, 7.95]	140.64***
Language (English vs. Chinese)	0.11	[0.003, 0.21]	2.01*
Language (Finnish vs. English)	−0.13	[−0.23, −0.03]	−2.48*
Dataset (Original vs. New)	0.04	[−0.05, 0.12]	0.89
Language (English vs. Chinese) × Dataset (Original vs. New)	−0.37	[−0.57, 0.18]	−3.79***
Language (Finnish vs. English) × Dataset (Original vs. New)	−0.11	[−0.31, 0.09]	−1.09
Random effects	*n*	*Variance*	*SD*
Subjects (Intercept)	303	0.08	0.28
Sentence (Intercept)	119	0.32	0.57
Sentence (Language: English vs. Chinese)		0.04	0.21
Sentence (Language: Finnish vs. English)		0.02	0.12
Residual		0.11	0.34

* = $p < .05$, *** = $p < .001$; p values are estimated using Satterthwaite's degrees of freedom method.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*, 2011–2084. https://doi.org/10.21500/20112084.807

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods, 23*, 389–411. https://doi.org/10.1037/met0000159

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language, 109*, 104047.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 12;1(1):9*. https://doi.org/10.5334/joc.10

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*, 263–279. https://doi.org/10.1017/S0140525X11001841

Goodrich B, Gabry J, Ali I, Brilleman S (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1. Retrieved from https://mc-stan.org/rstanarm.

Fu, Y., Liversedge, S. P., Bai, X., Moosa, M., & Zang, C. (2023). *Character representations in lexical identification during natural reading* (In submission).

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493–498. https://doi.org/10.1111/2041-210X.12504

Hyönä, J., Lorch, R. F., & Rinck, M. (2003). Eye movement measures to study global text processing. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 313–334). Amsterdam: Elsevier Science. https://doi.org/10.1016/B978-044451020-4/50018-9.

Li, X., Huang, L., Yao, P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology, 1*, 133–144. https://doi.org/10.1038/s44159-022-00022-6

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. https://doi.org/10.1016/j.cognition.2015.10.013

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*, 1541. https://doi.org/10.21105/joss.01541

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, Article 104092. https://doi.org/10.1016/j.jml.2020.104092

R Core Team. (2020). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110*. https://doi.org/10.1016/j.jml.2019.104038

Siegelman, N., Schroeder, S., Acartürk, C., et al. (2022). Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement Corpus (MECO). *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01772-6

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*, 1325–1346. https://doi.org/10.1037/bul0000169

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods, 51*, 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60*, 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Zang, C., Fu, Y., Bai, X., Yan, G., & Liversedge, S. P. (2018). Investigating word length effects in Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance, 44*, 1831–1841. https://doi.org/10.1037/xhp0000589

Zang, C., Liversedge, S. P., Bai, X., & Yan, G. (2011). Eye movements during Chinese reading. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 961–978). New York: Oxford University Press.