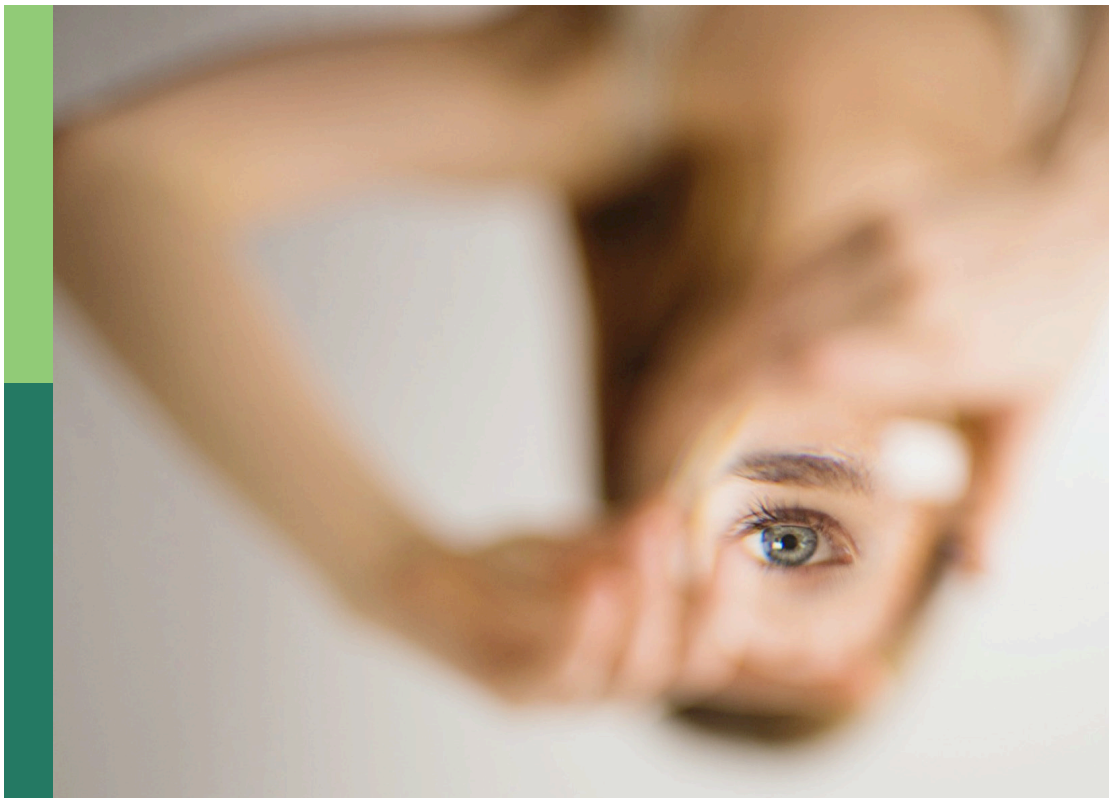
frontiers | Research Topics

Eye-tracking while reading for psycholinguistic and computational models of language comprehension

Edited by
Nora Hollenstein, Marijan Palmovic and Lena Ann Jäger

Published in
Frontiers in Psychology
Frontiers in Communication



ISSN 1664-8714
ISBN 978-2-8325-4101-2
DOI 10.3389/978-2-8325-4101-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Eye-tracking while reading for psycholinguistic and computational models of language comprehension

Topic editors

Nora Hollenstein — University of Copenhagen, Denmark

Marijan Palmovic — University of Zagreb, Croatia

Lena Ann Jäger — University of Potsdam, Germany

Citation

Hollenstein, N., Palmovic, M., Jäger, L. A., eds. (2023). *Eye-tracking while reading for psycholinguistic and computational models of language comprehension*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4101-2

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 28 November 2023
DOI 10.3389/fpsyg.2023.1326408

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Marijan Palmović
✉ marijan.palmovic@erf.unizg.hr

RECEIVED 23 October 2023
ACCEPTED 13 November 2023
PUBLISHED 28 November 2023

CITATION
Palmović M, Jäger LA and Hollenstein N (2023) Editorial: Eye-tracking while reading for psycholinguistic and computational models of language comprehension.
Front. Psychol. 14:1326408.
doi: 10.3389/fpsyg.2023.1326408

Editorial: Eye-tracking while reading for psycholinguistic and computational models of language comprehension

Marijan Palmović[1]*, Lena A. Jäger[2] and Nora Hollenstein[3]

[1]Laboratory for Psycholinguistic Research, University of Zagreb, Zagreb, Croatia, [2]Department of Computational Linguistics, University of Zurich, Zürich, Switzerland, [3]Center for Language Technology, University of Copenhagen, Copenhagen, Denmark

KEYWORDS

eye-tracking, reading, computational modeling, language comprehension, natural reading

Editorial on the Research Topic
Eye-tracking while reading for psycholinguistic and computational models of language comprehension

1 Aim of this Research Topic

Eye-tracking is a powerful technology for studying language processing. In recent years, it has been employed increasingly for reading studies based on collecting and analyzing reading corpora obtained in a natural setting, i.e., based on texts not experimentally manipulated as stimuli in minimal pairs. Typically, these texts are tagged with fixation and saccades data and some linguistic or psycholinguistic parameters. The large amount of the corpus data allows for new analytical techniques, resulting in new insights into psycholinguistic accounts of reading and, more generally, in psycholinguistic and computational models of language comprehension. Finally, collecting and sharing such corpora in various languages facilitates cross-linguistic studies of psycholinguistic phenomena, bilingual or multilingual studies, and research into individual differences among readers.

The creation of reading corpora allows for new directions in psycholinguistics research. It is the aim of this Research Topics to provide a platform for a discussion on this development in several directions:

1. The theoretical implications of large eye-tracking reading data in psycholinguistics;
2. Opportunities for comparative (cross-linguistic) and bilingual (multilingual) studies of psycholinguistic phenomena relevant to the formulation or evaluation of psycholinguistic or computational models of language comprehension;
3. The inclusion of languages other than English in order to alleviate the English language bias in psycholinguistic research.

In addition, methodological considerations within the eye-tracking research (e.g., corpus vs. experimental data) and between eye-tracking and similar methods (e.g., self-paced reading) reflect many issues in contemporary psycholinguistic modeling of language comprehension such as the interpretation of the processes captured by some dependent

variable obtained by eye-tracking or discussion about the arguments corroborating or refuting a particular psycholinguistic model. Finally, one expects that the reading corpora would allow for a more comprehensive study of the individual differences among readers, an issue that has recently attracted considerable attention in eye-tracking research.

2 Statistics of this Research Topic

The Research Topic was open from 27/06/2022 and the extended deadline concluded on 16/01/2023. Seventeen articles were submitted within this period, of which 11 were accepted after careful peer-reviewing.

3 Summary of this Research Topic

This Research Topic includes studies of a wide range of languages, including multiple language families (Sino-Tibetan, Indo-European, and Nortwest-Caucasian), various scripts (Latin and Cyrillic alphabets, Chinese characters), and modalities (written and signed languages). From a methodological perspective, the studies accepted in this Research Topic can be split into works concerning the computational modeling of reading and psycholinguistics investigations of language comprehension.

The computational articles focus on a diverse range of topics from predicting reading tasks (Hollenstein et al.), predicting metrics extracted from eye-tracking data (Salicchi et al.), to the acceptability of machine translation technology (Kasperė et al.), and using machine learning to extract neural components during rapid automatized naming (RAN) tests (Christoforou et al.).

The psycholinguistic articles in this Research Topic study a number of factors relevant to improving our understanding of reading comprehension, including lexical access (Chang et al.), grammatical errors (Søby et al.), individual differences (Gong and Shuai), typological differences (Zdorova et al.), text formatting (Medved et al.), and exposure to sign language (Ziubanova et al.). In the following, we briefly describe the contributions of each article.

Firstly, on the computational side, Christoforou et al. propose a novel machine-learning-based algorithm that extracts neural components from EEG and eye-tracking recordings of children with and without dyslexia during serial rapid automatized naming (RAN) tests. The authors show that these components capture the neural activity of cognitive processes associated with naming speed and are informative of group differences.

The ZuCo corpus contains eye-tracking and EEG data during normal reading and information-searching reading in English. The benchmark provides a new hidden testset for machine learning models trained to distinguish these two tasks (Hollenstein et al.). Improving the performance of reading task classification will be useful in identifying the relevant features and can advance models of reading.

Previous research in computational linguistics has investigated whether distributional language models can predict metrics extracted from eye-tracking data. In their study, Salicchi et al. propose a regression experiment for estimating different eye-tracking metrics on two English corpora, contrasting the quality of the predictions with and without the surprisal and the relatedness components. Their results suggest that both components play a role in the prediction, with semantic relatedness surprisingly contributing also to the prediction of function words.

Between the realms of computational language processing and psycholinguistics, Kasperė et al. leverage eye-tracking to investigate the acceptability of machine translation technology between professional translators and non-professionals. In a study in which participants read an English text machine-translated into Lithuanian, the authors analyze whether raw machine translation output is processed in the same way by both groups. In terms of acceptability overall, professional translators critically assess machine translation on all components, which confirms the findings of previous similar research. However, the current study draws attention to the lower awareness of non-professionals regarding machine translation quality.

On the side of psycholinguistic reading research, Chang et al. employ an eye-tracking experiment to corroborate the "graded pre-activation" account of lexical access in explaining the predictions of the coming words in a sentence.

Grammar errors are a natural part of everyday written communication and come in different forms, e.g., syntactic errors, morphological agreement errors, and orthographic errors. Søby et al. examine whether some types of naturally occurring errors attract more attention than others during the reading of Danish texts, measured by detection rates. While this study did not measure eye movements, the differences in error detection patterns point to shortcomings of existing eye-tracking models.

Furthermore, in a sentence reading eye-tracking study, Gong and Shuai assess participants' reading skills on a number of language and cognitive measures while manipulating the lexical properties of the words in the stimulus sentences. The interactions between text properties and reading skills proved to be significant on early and late eye-tracking measures.

Tywoniw analyzes reading strategies in English as L2 with participants of varying native language backgrounds. Their individual differences and the differences across experimental conditions (close reading, multiple-choice, and reading-to-summarize) are studied to identify the predictors of reading behavior.

Zdorova et al. study how typological differences impact reading behavior of Adyghe-Russian bilinguals. A robust frequency effect was found in Adyghe, while the words of the same length in Adyghe and Russian were read slower in Adyghe due to their complex morphological structure.

Not only linguistic characteristics but also visual aspects influence reading comprehension. More and more educational material is delivered to students through digital screens. Therefore, the text format in which these materials are presented is an important aspect to consider. Medved et al. investigate the effect of letter shape on readers' feelings of pleasantness during reading, reading fluency, and text comprehension and memorization of Slovenian texts. They find that softer typefaces of rounder shapes should be used in educational materials for a more pleasant reading experience and improved learning process.

Finally, Ziubanova et al. study the benefits of early exposure to spoken and sign language for deaf adults and adults with severe hearing impairments in an eye-tracking sentence reading experiment. The benefits of early exposure were confirmed for adults with severe hearing impairments.

Author contributions

NH: Writing - original draft, Writing—review & editing. MP: Writing—original draft, Writing—review & editing. LJ: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This Research Topic was conducted around the activities funded by the MultiplEYE COST Action (CA21131) and the SNSF-HRZZ MeRID project (IPCH-2022-04-3316).

Acknowledgments

We thank all contributors and reviewers for their efforts.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 11 January 2023
DOI 10.3389/fpsyg.2022.1006662

OPEN ACCESS

EDITED BY
Marijan Palmovic,
University of Zagreb, Croatia

REVIEWED BY
Xiaolu Wang,
Zhejiang University City College, China
Yuxia Wang,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Tao Gong
gtojty@gmail.com

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 29 July 2022
ACCEPTED 19 December 2022
PUBLISHED 11 January 2023

CITATION
Gong T and Shuai L (2023) Segmented relations between online reading behaviors, text properties, and reader–text interactions: An eye-movement experiment. *Front. Psychol.* 13:1006662. doi: 10.3389/fpsyg.2022.1006662

Segmented relations between online reading behaviors, text properties, and reader–text interactions: An eye-movement experiment

Tao Gong[1,2,3,4]* and Lan Shuai[1,3]

[1]Haskins Laboratories, New Haven, CT, United States, [2]School of Foreign Languages, Zhejiang University of Finance and Economics, Hangzhou, Zhejiang, China, [3]Educational Testing Service, Princeton, NJ, United States, [4]Google, New York, NY, United States

Purpose: To investigate relations between abilities of readers and properties of words during online sentence reading, we conducted a sentence reading eye-movements study on young adults of English monolinguals from the US, who exhibited a wide scope of individual differences in standard measures of language and literacy skills.

Method: We adopted mixed-effects regression models of gaze measures of early and late print processing stages from sentence onset to investigate possible associations between gaze measures, text properties, and skill measures. We also applied segmented linear regressions to detect the dynamics of identified associations.

Results: Our study reported significant associations between (a) gaze measures (first-pass reading time, total reading times, and first-pass regression probability) and (b) interactions of lexical properties (word length or position) and skill measures (vocabulary, oral reading fluency, decoding, and verbal working memory), and confirmed a segmented linear dynamics between gaze measures and lexical properties, which was influenced by skill measures.

Conclusion: This study extends the previous work on predictive effects of individual language and literacy skills on online reading behavior, enriches the existing methodology exploring the dynamics of associations between lexical properties and eye-movement measures, and stimulates future work investigating factors that shape such dynamics.

KEYWORDS

eye-movement, lexical properties, individual differences, mixed-effects regression, segmented linear regression

1. Introduction

Contemporary views of reading highlight connections among cognitive abilities of readers, properties of texts, reading comprehension, and online reading behavior. The simple view of reading (SVR) proposes that reading comprehension is a function of visual word recognition, decoding, and language comprehension, the first two of which are print-specific aspects of reading skill (Gough and Tunmer, 1986), and the latter is construed as an *amodal* (not limited to a particular module like reading or listening) aspect of language. However, how language and literacy skills relate to lexical properties (e.g., word frequency, length, predictability, and position in sentence) and online reading behavior remains *implied*, at best, in SVR. In addition, the self-teaching hypothesis (STH) (Share, 1995) proposes that decoding allows developing readers to transform unfamiliar printed letter strings into recognizable sounds from their spoken language. This process helps readers to internalize the orthographic features of new words. Although highlighting that decoding skill predicts development of reading comprehension, thus being necessary for a reader to learn all words, orthographically regular or not, STH does not state explicitly how decoding helps comprehension during online sentence reading. Furthermore, the lexical quality hypothesis (LQH) (Perfetti, 2007) and the verbal efficiency theory (VET) (Perfetti, 1985) advocate that what distinguishes good and poor readers is the ability to efficiently map orthographic forms to phonological representations, and ultimately to semantics. However, it is unclear how different aspects or levels of language and literacy skills influence reading processes.

Existing studies of individual differences in reading often focus on offline outcomes (e.g., reading comprehension), and these outcomes are in fact cumulative end products of various processes involved in meaning construction (Snow, 2002). Recent studies have begun to shift their attentions from reading outcomes to reading processes, as in moment-to-moment measures (e.g., eye-movements) of reading behavior (e.g., Traxler, 2007; Rayner, 2009b; Kuperman and Van Dyke, 2011; Radach and Kennedy, 2012; Rayner et al., 2012, 2015; Kuperman et al., 2018). Eye-movement patterns during reading are found to vary with lexical properties (Rayner and Duffy, 1986; Rayner, 1998; Joseph et al., 2013). In addition, eye-movement patterns also rely on cognitive capacities that support reading. The dynamics of information processing during reading is governed not only by lexical properties of the text (Radach and Kennedy, 2012), but also by knowledge and cognitive resources of the reader (Beck et al., 1982; Gough and Tunmer, 1986; Hoover and Gough, 1990; Catts et al., 2006). Reading comprehension emerges as a juxtaposition of the lexical properties and the skills, knowledge, and experience of the reader (Perfetti and Lesgold, 1977; Nelson Taylor and Perfetti, 2016; Kuperman et al., 2018).

Many previous studies have reported the "direct" roles of language and literacy skills in reading outcomes or predicting online reading behavior, but there lack enough investigations on whether those skills could also "indirectly" influence reading behavior through interactions with lexical properties, given that lexical properties are central to regulation of gaze behavior during connected text reading and influence of effortful lexical, syntactic, semantic, and pragmatic processing (Duffy et al., 1988; Rayner, 2009a). In addition, many studies have focused selectively on university students (there are exceptions though, e.g., one on participants of similar age and skill to those in our study (Kuperman and Van Dyke, 2011), and two on readers younger than those in our study (Joseph et al., 2013; Valle et al., 2013). University students often have a narrow range of language and literacy skills centered above average, which makes them insufficient to reveal the potentially much wider scope of individual differences in those skills and the general effects of such differences on online reading behavior (Henrich et al., 2010). Furthermore, through simple regression analyses, many existing studies only reported whether or not an online reading behavior is correlated with certain lexical properties and/or individual skills, yet there lack investigations on the dynamics of identified relations, e.g., does an identified correlation follow a simple linear relation, a nonlinear relation, or else? Given that there has been accumulated evidence informing us about what lexical properties or individual skills may or may not influence or be correlated with online reading behavior, it is time to further examine the dynamics of such causal or correlational relations concerning lexical properties, individual skills, and online reading behavior.

Noting these and given the dearth of research on how differences in basic reading skills and vocabulary exert influence on online reading at a sentence level, this study was designed to yield empirical data and inform relevant theories. Based on eye-movement measures and online reading process, this study aimed to investigate two research questions:

(a) Can interactions between language and literacy skills and lexical properties influence online reading behavior?
(b) What is the dynamics of the correlation between online reading behavior and lexical properties?

Answers to these questions will bring an intimate view of reading process at the levels of words, phrases, and larger units, and contribute to the research on how lexical properties and individual differences in language and literacy skills jointly affect online reading behavior.

Following a data-driven approach, this study centered on the skills concerned with reading process as gauged by online reading behaviors, and investigated how these skills interact with lexical properties during online reading. In view of the existing theories (e.g., SVR, STH, LQH, and VET), we focused on four skills: decoding, reading fluency, word knowledge (vocabulary), and working memory (see next section for details). Some of them were omitted in early studies [e.g.,

working memory was not included by Kuperman and Van Dyke (2011)]. In our study, all these skills were assessed by a battery of standard tests. In addition, as part of a general research program aimed at developing profiles of adolescent and young adult readers (aged from 16 to 25 years), especially those whose educational and occupational prospects might be constrained by their limited language and literacy skills, our study targeted on non-university students, who possess a much wider range of individual differences in these skills than would typically be found in university students (Braze et al., 2007). This enables a more detailed examination of the role of individual differences in reading behavior than would be possible with a more restricted range of differences (Peterson, 2001). Furthermore, after identifying significant interactions between lexical properties and individual skills on online reading behavior gauged by eye-movement measures, we further quantified the dynamics between the aspects involved in the significant interactions, i.e., the lexical properties and the eye-movement measures in participants with high and low levels of the skills. This data-driven analysis helps reveal how the skills influence online reading behavior via interactions with lexical properties.

In terms of methodology, we applied mixed-effects regression models on rich observations, and carefully controlled the family-wise errors, collinearity, and overfitting. This type of models can simultaneously address the main effects of the skill measures and their interactions with lexical properties in one model, and collectively reveal the key interactions with lexical properties. In addition, we designed a way to visualize the correlations between eye-movement measures and lexical properties under different levels of skills, and selected among popular regression models the "best" one to reflect the dynamics. A similar method was practiced to detect quantitative relations between decoding skills and comprehension scores in reading assessment (Wang et al., 2019).

Our study did not find significant effects attributable to individual skill differences, due primarily to the wider spans of the abilities in our participants than those of university students recruited in previous studies. Nonetheless, we identified significant interactions between lexical properties and individual skills, including: interactions between word length and verbal working memory and oral comprehension plus vocabulary in regulating first-pass reading time, interactions between word position and oral reading fluency and verbal working memory in shaping total reading times, and interaction between word position and decoding in adjusting first-pass regression probability. Our analysis revealed a segmented linear dynamics between lexical properties and eye-movement measures, which could be further manipulated by individual skills. All these findings reveal important predictability of those skills on online reading behavior, and contribute to theoretical discussions on how those skills regulate reading behavior at a sentence level through interactions with lexical properties. Note that more research is needed to better understand what factors shape the pivot points in the segmented linear curves.

2. Target skills and recent studies on them

Among various language and literacy skills, we focused on four of them.

Decoding is the ability to apply the orthography-to-phonology correspondence rules to pronounce written words. It is essential to translating print to spoken language, and includes, at least, the knowledge of letter patterns and letter-sound relationships, upon which all other reading skills are built (Share, 1995). SVH claims that decoding, together with listening comprehension, makes substantial contributions to variation in reading comprehension. Studies have revealed that reading comprehension differences are associated with decoding skill differences in children and adolescent readers (Shankweiler et al., 1999) and that the ability to retrieve phonological cues can predict individual differences in reading fluency (Barth et al., 2009). Studies of online reading processes have discovered that a high decoding skill enables a rapid access to a word's orthographic form and its meaning, thus accelerating word naming speed (Manis and Freedman, 2001) and reflecting high text-level reading fluency and word-level recognition during connected text reading (Wolf et al., 2002).

Reading fluency is the ability to read connected text quickly, accurately, and with expression. Conventional measures like the Gray Oral Reading Test (Wiederholt and Bryant, 2001) assess oral reading fluency. Recent tests measure this skill through silent reading, e.g., the Silent Reading Efficiency and Comprehension Test (Wagner et al., 2010). Regardless of modality, reading fluency measures draw on important capacities to lexical access (Perfetti, 1985) and mediate reading comprehension (e.g., Tilstra et al., 2009; Macaruso and Shankweiler, 2010; Silverman et al., 2012). Longitudinal and corpus-based studies have shown that reading fluency is a reliable index of reading comprehension in students (Fuchs et al., 2001; Miller and Schwanenflugel, 2008; Reschly et al., 2009; Petscher and Kim, 2011) and it performs as well as or better than other reading comprehension tests as a predictor for higher stakes comprehension tasks (Baker et al., 2008; Marcotte and Hintze, 2009). Eye-movement studies have also revealed that phonemic awareness, a known predictor for early word recognition and decoding, contributes to reading fluency (Ashby et al., 2013).

Vocabulary is another key component of reading skills. Orally assessed vocabulary knowledge captures variance in reading comprehension, even if comprehension and decoding skill are accounted for (Braze et al., 2007; Tunmer and Chapman, 2012). Vocabulary breadth and depth, as well as semantic relatedness can predict individual differences in

reading comprehension of fourth-grade students (Swart et al., 2016). Oral vocabulary makes an independent contribution to reading comprehension in grade school children (Ouellette and Beers, 2010) and young adult readers (Braze et al., 2007), and serves as a strong predictor for reading comprehension in typically developing Grades 1–3 students and dyslexic readers of Grades 4–5 (Chik et al., 2010). During sentence reading, high-vocabulary readers are found more likely to make online elaborative inferences than low-vocabulary ones (Calvo et al., 2003). Nelson Taylor and Perfetti (2016) report that: readers with greater knowledge of less common words tend to read faster and with greater accuracy in paragraph reading, and the amount of exposure to phonological and semantic constituents of words during training modulates re-reading behavior in this process.

Verbal working memory enables readers to hold on to verbal cues to comprehend lengthy or complex sentences, and thus facilitates readers' abilities to derive compositional meanings of sentences. High working memory capacity can accelerate the time course of predictive inferences during sentence reading (Estevez and Calvo, 2000). Compared to readers with higher working memory capacity, those with lower capacity exhibit more difficulties (in terms of longer regression and total fixation time) in associating relative clauses with preceding fragments (Traxler, 2007), and spend more time re-reading ambiguous regions of texts (Clifton et al., 2003). Higher working memory capacity is also associated with higher reading fluency (with lower gaze durations and fewer look-backs from the final word of a sentence) (Calvo, 2004).

Motivated by previous studies on those skills, our study attempted to explore how reader-text interaction predicts reading patterns between good and poor readers differing in those skills.

In this line of research, existing studies often focus on identifying (by mixed-effects or generalized regression models, or machine learning models) the language and literacy skills that directly or indirectly (via interaction with lexical properties) cast important effects on reading process, but rarely touch upon the *dynamics* of any identified correlations between lexical or individual properties and reading process, e.g., whether and how the correlation between target skills and lexical properties change alongside the levels of the skills. For example, a recent study (Kuperman and Van Dyke, 2011) has shown that individual scores in rapid automatized letter naming (RAN) and word identification tests can supersede the effects of word length and frequency at early processing stages, and serve as stronger predictors than word frequency across eye-movement measures. However, family-wise Type-I error was not carefully controlled in the analyses (e.g., the same critical p value of 0.05 was used over 150 models involving multiple predictors that are correlated with each other), which weakens the claims that those skill measures are reliable predictors for online reading behavior.

Another study from the same group (Kuperman et al., 2018) incorporated more cognitive and linguistic skills, used sentence stimuli with increasing lexical, syntactic, and discourse complexity, and adopted random forest models to detect key predictors for eye-movement measures. This work analyzed the effects of lexical properties, individual skills, interactions between word length and those skills, and sentence complexity on eye-movements around words inside sentences, at the end of sentences, and whole passages. The analyses reported reading habit, vocabulary size, reading efficiency, vocabulary IQ, and rapid naming scores as key predictors on eye-movement patterns during online reading.

This data-driven approach fails to identify multiple factors having dominant and comparatively small yet still important effects. In a random forest model, extremely-high relative importance score of a predictor could mask the roles of other predictors. Since importance scores are relative to predictors, one random forest model cannot address all possible interactions between lexical properties and skill measures. In addition, the work *indirectly* examined the effects of interactions with word length: word length was segmented into long and short groups, and two random forest models were fitted respectively on the two groups to detect important skill measures whose effects exhibited different tendencies between the two models. The arbitrary, binary segmentation of word length groups presumes that if a skill measure has an influence on word length, the tendencies of the effect should be different on short and long words. This is not always the case; some factors may take effect on very long words, and others may trigger different reading patterns on very short words. A question on whether reading processes differ between individuals with high and low levels of skills is more meaningful than whether such processes differ between long and short words; in this sense, segmentation on skill levels is more informative than segmentation on lexical properties.

3. Materials and methods

The data in this study consisted of: (a) participants' skill measure data obtained from a battery of standard psycho-educational tests; and (b) their eye-movement data gathered in a sentence reading experiment. The data were collected by trained research assistants. Informed consent was obtained from the participants of at least 18 years old; for those under 18, the participants provided assent and their parents or guardians signed written permissions. All participants were paid a proper remuneration for completing the protocols reported here together with the fMRI protocols reported elsewhere (Shankweiler et al., 2008; Braze et al., 2011). The procedures described here took ∼3.5 h; breaks were provided as needed.

3.1. Participants

Forty-five participants (age in 16–25 years, 27 females) were recruited from adult education centers, community college, and neighborhood-gathering places. Some participants had their secondary schooling interrupted but were then seeking a high school equivalency certificate or resuming work toward a regular high school diploma. At the time of experiment, most participants were enrolled in education programs (e.g., high school, adult school, or community college) (Braze et al., 2007, 2016). All participants were English monolinguals, and had normal or corrected-to-normal vision. They were prescreened to ensure the ability of reading simple sentences with comprehension. Data from one participant were excluded due to not completing all study components.

According to the power analysis in mixed-effects models (Brysbaert and Stevens, 2018), this number of sample size, together with the rich amount of eye-movement observations obtained during reading of multiple (72) sentences containing numerous (358) word types (see section 3.3 Materials and design), is sufficient to detect reliable significant factors.

3.2. Skill measures

Each participant was assessed in six domains of language and literacy skills, which served as the bases for analysis. **Table 1** shows the raw (and normative wherever available) scores of each measure and a key to the labels of them. The domains and the tests used to measure them were:

(1) *Vocabulary*, assessed by the Peabody Picture Vocabulary Test-Revised (ppvt) (Dunn and Dunn, 1997) and the Wechsler Abbreviated Scales of Intelligence Expressive Vocabulary Test (wasi.v) (Psychological Corporation, 1999). **Table 1** shows both raw and standard scores (normative sample mean = 100, SD = 15) of ppvt and both raw and *t*-scores (normative sample mean = 50, SD = 10) of wasi.v. Differences in word knowledge stem from (a) variations in language experience (in speech or print) and (b) differences in the ability to profit from it. Vocabulary is a good proxy for general, amodal, language ability of the community sample recruited in our study (Braze et al., 2016).

(2) *Listening comprehension*, assessed by the even-numbered items from the Reading Comprehension subtest of the Peabody Individual Achievement Test-Revised (piat.l) (Markwardt, 1998). Using the odd numbered items from this test for reading comprehension and the even numbered items for listening comprehension gives us a pair of tests well matched in task demand for both input modalities. **Table 1** shows both raw and grade equivalent scores, the latter of which were calculated following Markwardt (1998) (see Braze et al., 2007 for details). Knowledge of vocabulary, compositional semantics, and syntax constitute the bases of oral language comprehension (Birch and Rayner, 1997; Frisson and McElree, 2008). The ability to understand language presented to the ear is a good indicator of general, amodal, language comprehension ability.

(3) *Decoding*, assessed by the Woodcock-Johnson-III Word Identification subtest (wid) (Woodcock et al., 2001) and the Woodcock-Johnson-III Test of Achievement Word Attack subtest (watt) (Woodcock et al., 2001). These are untimed tests for the ability to accurately pronounce printed words and non-words. **Table 1** contains both raw and grade equivalent scores of the two measures.

(4) *Reading comprehension*, assessed by the odd numbered items from the Reading Comprehension subtest of the Peabody Individual Achievement Test-Revised (piat.r) (Markwardt, 1998) and the accuracies of the Passages 5, 7, and 9 from the Gray Oral Reading Test (gort.comp) (Wiederholt and Bryant, 2001). Calculation of grade equivalent scores of piat.r followed Braze et al. (2007). There were no standard scores of gort.comp, due to using only a subset of passages. Reading comprehension has been usefully thought of as the product of an individual's facility with language and decoding skill (Gough and Tunmer, 1986).

(5) *Oral reading fluency*, assessed as the reading speed (words per minute) for Passages 5, 7, and 9 from the Gray Oral Reading Test (gort.wpm); the total number of words in these passages is 361 (Wiederholt and Bryant, 2001). There were no standard scores, since the measure was based on an abbreviated form of the Gray Oral Reading Test. Oral reading fluency consists of visual scanning, decoding, and high level language processing (Silverman et al., 2012).

(6) *Verbal working memory*, assessed by a listening version of the Sentence Span task (sspan.corr) (Daneman and Carpenter, 1980). This ensures non-confoundness with reading skills. Verbal working memory has been shown to account for differences in vocabulary growth independent of language exposure (Gathercole and Baddeley, 1989; Gathercole et al., 1999; Gupta, 2006).

These individual difference measures can be grouped into two sets: those explicitly linked to reading ability (reading comprehension, decoding skill, and oral reading fluency), and those not (listening comprehension, vocabulary, and verbal working memory) (Gough and Tunmer, 1986; Hoover and Gough, 1990). They tap into abilities equally important to comprehension, no matter whether the language input arrives by ear or by eye.

In addition to these domains, we also assessed *print experience* by a magazine title recognition checklist (MRT) and an author recognition checklist (ART) (cf.,

TABLE 1 Raw scores and keys of the skill measures over 44 participants.

Name	Label	Mean	SD	Min.	25%	50%	75%	Max.	Skew	Kurtosis	Lambda
Age	–	20.61	2.27	16.6	18.73	20.16	22.41	25.49	0.30	−0.96	
Vocabulary	ppvt	172.41	17.64	132.00	160.50	176.50	187.00	196.00	−0.60	−0.71	1.18
	std.-score	103.39	14.65	78.00	92.00	102.00	115.00	132.00	0.12	−0.99	
	wasi.v	57.36	8.33	39.00	49.00	57.50	62.50	76.00	0.22	−0.73	
	t-score	53.25	9.73	36.00	44.00	52.50	60.00	74.00	0.38	−0.77	
Listening comprehension	piat.l	93.95	6.16	76.00	92.00	96.00	98.50	100.00	−1.34	0.95	2.08
	grade equiv.	12.00	1.84	6.90	11.60	13.00	13.00	13.00	−1.78	1.82	3.50
Decoding	wid	67.61	5.38	56.00	63.00	68.00	72.00	76.00	−0.19	−1.12	
	grade equiv.	13.18	4.81	5.60	8.50	12.70	19.00	19.00	0.05	−1.62	
	watt	27.23	3.06	20.00	25.50	28.00	30.00	32.00	−0.55	−0.66	1.38
	grade equiv.	10.83	4.67	4.30	7.10	10.20	15.40	19.00	0.40	−1.17	
Reading Comprehension	piat.r	89.48	10.25	68.00	83.50	95.00	97.50	99.00	−0.92	−0.57	1.34
	grade equiv.	10.79	2.86	5.00	8.50	13.00	13.00	13.00	−0.83	−0.92	1.62
	gort.comp	11.75	2.47	4.00	10.00	12.00	14.00	15.00	−0.68	0.38	1.57
Oral Reading Fluency	gort.wpm	177.01	39.04	87.34	149.91	176.82	197.81	288.80	0.39	0.31	
Verbal Working Memory	sspan.corr	31.88	5.78	20.00	27.00	33.00	36.50	42.00	−0.30	−1.04	

"Lambda" is for Box-Cox transformation for highly skewed scores; ppvt, Peabody Picture Vocabulary Test-Revised; wasi.v, Wechsler Abbreviated Scales of Intelligence Expressive Vocabulary Test; wid, Woodcock-Johnson-III Word Identification subtest; piat.l, even-numbered items in the Reading Comprehension subtest of the Peabody Individual Achievement Test-Revised; watt, Woodcock-Johnson-III Test of Achievement Word Attack subtest; piat.r, Peabody Individual Achievement Test-Revised; gort.comp, Gray Oral Reading Tests; gort.wpm, reading speed (words per minute) for Passages 5, 7, and 9 from the Gray Oral Reading Test; sspan.corr, listening version of the Sentence Span task.

10.3389/fpsyg.2022.1006662

Stanovich and Cunningham, 1992) to gauge a person's experience with language in printed form, which for literate individuals may well be a substantial part of their overall language experience, and *visual working memory* based on a computerized version of the Corsi Blocks task (corsi) (Corkin, 1974) implemented in Psyscope (Cohen et al., 1993). Given the fact ART and MRT only show high validity and reliability in proficient readers (e.g., university students) in their dominant language (McCarron and Kuperman, 2021), whereas our study is based upon participants having a wide span of reading skills, we excluded print experience in the regression analyses. In addition, compared to visual working memory, verbal working memory is more relevant to our sentence reading experiment, so we also excluded visual working memory in the regression analyses.

Prior to regression modeling, we examined the distributions of raw scores for deviations from normality. Several scores showed high skewness (absolute values over .5). To them, we applied Box-Cox transformations (Box and Cox, 1964) using the bcpower function in the R package *car* (Fox and Weisberg, 2011). All variables, transformed or not, were standardized by converting to *Z*-scores. **Table 2** is the correlation table of the transformed and standardized measures (cf., Braze et al., 2007, 2016).

To reduce collinearity and the total number of predictors in the regression models, we combined measures tapping into common latent constructs. This was done by (a) taking the average of the transformed and standardized scores, and then (b) converting the average scores back to *Z*-scores. Measures of vocabulary and listening comprehension were combined into a composite measure of oral comprehension plus vocabulary (*oral.comp*) (Tunmer and Chapman, 2012; Braze et al., 2016; Kukona et al., 2016). Composites were also derived for decoding (*decod.comp*) and reading comprehension (*readcomp.comp*). **Table 3** shows the correlation table of the centered and transformed skill measures.

It is not surprising that the correlation between reading comprehension and oral comprehension plus vocabulary is high, since oral knowledge is an important indicator of reading comprehension (see section 2. Target skills and recent studies on them). **Table 4** shows the statistics of the regression models between reading comprehension and oral comprehension plus vocabulary, decoding, and both, respectively. Consistent with early findings (Braze et al., 2007), a combination of both skills largely explains the variation of reading comprehension: R^2 of the model using decoding is .370, R^2 of the model using oral comprehension plus vocabulary is .712, and multiple R^2 of the regression model using both decoding and oral comprehension plus vocabulary is .738. Notably, we exclude reading comprehension from the list of predictors in the regression models.

After these preprocessing stapes, the skill measures used in our regression analyses are: (a) oral comprehension plus vocabulary (oral.comp); (b) decoding (decod.comp); (c) oral reading fluency (gort.wpm); and (d) verbal working memory (sspan.corr).

3.3. Materials and design

Participants were asked to read 72 individual sentences while their eye-movements were recorded. Presentation order was pseudo-random across participants. These sentences were filler items in a study of comprehension process in young adults with limited literacy skills (Braze et al., 2006). All of the sentences were grammatical and transparent in meaning. The word types in them were carefully selected among high frequent words, and common names for persons, states, or holidays. The linguistic aspects of these sentences, such as part of speech or syntactic complexity, were carefully controlled. **Supplementary Table 1** shows the complete list of the sentences. Many of these sentences were simple in terms of structure; forty-six stimuli sentences (over 79%) had no embedding structures, e.g., "Most of the students will be going to the class picnic next month."; and the other 26 had one dependent clause, e.g., "The waiter had told the customer that the pies were fresh." There were 503 unique word types (819 word tokens) in these sentences, an average of 11.375 word tokens per sentence (range = 11–16). Note that previous studies on university students involved sentences with increasing complexity in semantics and syntax (e.g., Kuperman and Van Dyke, 2011; Kuperman et al., 2018)), we leave the investigation of the relations between sentence complexity, reading skills, and online reading behavior for future work.

Before the experiment, we asked some individuals to evaluate the understandability of these sentences, based on a scale of 5, from "easy to understand" to "hard understand". These individuals were recruited similarly as the experiment participants, but did not participate the experiment. All of them marked the filler sentences as "easy to understand".

For each word in a sentence, we recorded its ordinal position in the sentence (note that the sentence initial and final words were excluded), its length in characters (Len_W), and its frequency of occurrence per million words (Freq_W). Word position is a context-dependent property, but word length and frequency are independent of sentence. Lexical frequencies were obtained from the Corpus of Contemporary American English (COCA).[1] Frequency summaries for our materials exclude contractions (n = 2) and proper nouns (n = 23), both having no COCA frequencies. Possessive forms (n = 6) used the COCA frequencies of their uninflected forms. Analyses otherwise included all the remaining words found in the sentences. Most of the type frequencies showed skewed distributions, and thus log-transformed

1 http://corpus.byu.edu/coca/

(base *e*). Following Kuperman and Van Dyke (2011, 2013) and other standard practice, we excluded words with a high likelihood of being skipped (i.e., highly-frequent and very short words).

Table 5 shows the lexical properties of the words contained in the sentences. Regression models targeting online reading indicators (gaze measures) at a word also included parameters for length and frequency of the previous and subsequent words. Differences between Len_W and Len_{W-1} (or Len_{W+1}) are due to the exclusion of sentence initial and final words in the current word set (see Eye-movement measures), so are differences between Freq_W and Freq_{W-1} (or Freq_{W+1}).

TABLE 2 Correlations between the age and the 9 skill measures, after Box-Cox transformation (for ppvt, watt, piat.r, and gort.comp) and standardization.

Measures	1	2	3	4	5	6	7	8	9
1. Age									
2. ppvt	0.591								
3. wasi.v	0.378	0.829							
4. piat.l	0.408	0.638	0.541						
5. wid	0.487	0.816	0.716	0.441					
6. watt	0.075	0.376	0.354	−0.025	0.613				
7. piat.r	0.550	0.798	0.718	0.634	0.714	0.317			
8. gort.comp	0.351	0.673	0.610	0.625	0.586	0.367	0.648		
9. gort.wpm	0.374	0.577	0.577	0.177	0.617	0.347	0.481	0.348	
10. sspan.corr	0.319	0.626	0.669	0.380	0.601	0.392	0.573	0.557	0.474

$n = 44$, $|r| \geq 0.24$ corresponds to $p < 0.05$; $|r| \geq 0.31$ to $p < 0.01$; $|r| \geq 0.39$ to $p < 0.001$. ppvt, Peabody Picture Vocabulary Test-Revised; wasi.v, Wechsler Abbreviated Scales of Intelligence Expressive Vocabulary Test; piat.l, even-numbered items in the Reading Comprehension subtest of the Peabody Individual Achievement Test-Revised; wid, Woodcock-Johnson-III Word Identification subtest; watt, Woodcock-Johnson-III Test of Achievement Word Attack subtest; piat.r, Peabody Individual Achievement Test-Revised; gort.comp, Gray Oral Reading Tests; gort.wpm, reading speed (words per minute) for Passages 5, 7, and 9 from the Gray Oral Reading Test; sspan.corr, listening version of the Sentence Span task.

TABLE 3 Correlations between the age and the 5 composite or independent measures.

Measures	1	2	3	4	5
1. Age					
2. oral.comp	0.520				
3. decod.comp	0.313	0.231			
4. readcomp.comp	0.497	0.844	0.608		
5. gort.wpm	0.374	0.503	0.536	0.457	
6. sspan.corr	0.319	0.632	0.553	0.622	0.474

$n = 44$, $|r| \geq 0.24$ corresponds to $p < 0.05$; $|r| \geq 0.31$ to $p < 0.01$; $|r| \geq 0.39$ to $p < 0.001$. oral.comp, oral comprehension plus vocabulary, a composite variable of ppvt (Peabody Picture Vocabulary Test-Revised), wasi.v (Wechsler Abbreviated Scales of Intelligence Expressive Vocabulary Test) and piat.l (even-numbered items in the Reading Comprehension subtest of the Peabody Individual Achievement Test-Revised); decod.comp, decoding skill, a composite variable of wid (Woodcock-Johnson-III Word Identification subtest) and watt (Woodcock-Johnson-III Test of Achievement Word Attack subtest); readcomp.comp, reading comprehension skill, a composite variable of piat.r (Peabody Individual Achievement Test-Revised) and gort.comp (Gray Oral Reading Tests); gort.wpm, reading speed (words per minute) for Passages 5, 7, 9 from the Gray Oral Reading Test; sspan.corr, listening version of the Sentence Span task.

TABLE 4 Regression models targeting reading comprehension.

Model A:	Est.	SE	*t*	*p*	R^2
Decoding	0.608	0.123	4.964	0.00001	0.370
Model B:					
Oral comprehension plus vocabulary	0.844	0.083	10.190	<0.00001	0.712
Model C:					
Decoding	0.195	0.097	2.016	0.0503	0.738
Oral comprehension plus vocabulary	0.743	0.097	7.593	<0.00001	

Model A: Using decoding to predict reading comprehension; Model B: Using oral comprehension plus vocabulary to predict reading comprehension; Model C: Using both decoding and oral comprehension plus vocabulary to predict reading comprehension. R^2 is the proportion of variance captured by a given variable after considering all other predictors in the model.

Prior to the analyses, we mean-centered lexical properties. Word length was measured in terms of number of characters. Log-transformed word frequency was standardized. Word frequencies were highly correlated with lengths of respective words: Pearson's r between current word length and current word frequency was −0.731 ($p < 0.001$), −0.775 ($p < 0.001$) between previous word length and previous word frequency, and −0.752 ($p < 0.001$) between next word length and next word frequency. Following Kuperman and Van Dyke (2011), we residualized word frequencies against lengths of respective words. This was done by fitting a regression model for each of the three properties (previous, current, and next words) in which the frequency of the relevant word was predicted by its length. We took the residuals (distances between the observed and fitted values) of these models as the values of word frequency. The residualized frequencies remained strongly correlated with the original frequencies but orthogonal to the lengths of respective words: Pearson's r between residualized and original frequencies was 0.697 ($p < 0.001$) for current word frequencies, 0.643 ($p < 0.001$) for previous word frequencies, and 0.665 ($p < 0.001$) for next word frequencies. The residualization (or orthogonalization) procedure does not change the result for the residualized variable, the overall explanatory power of the model, and any indices of model fit. Some scholars pointed out that such orthogonalization (Wurm and Fisicaro, 2014) could not be a useful remedy for collinearity; note that in our experiment, the significant factors reported by the regression analyses using the orthogonalized or unorthogonalized word frequency and word length values are the same.

3.4. Apparatus and procedure

During the test session, participants were instructed to read, one by one, a number of sentences, and to answer yes/no comprehension questions about the contents of the sentences just read (see **Supplementary Table 1**). Comprehension questions occurred immediately after some sentences on about a sixth of trials to ensure that participants stayed focused on the reading comprehension task throughout the session. The mean response accuracy to the comprehension questions was 0.913 (SD = 0.067).

Each sentence was presented on a single line vertically centered on a monitor, which was positioned approximately 64 centimeters from the participants' eyes. The sentences were displayed in a monospace font (Bitstream MonoSpace 821) in black with a light background, at a screen resolution 1,280 × 1,024 and a refresh rate 85 Hz. Font size was set such that each character subtended about 17 minutes of visual arc. Participants wore an EyeLink II head-mounted eye tracker (SR Research), the sampling rate of which was set to 250 Hz. Before the test session, the accuracy of the eye tracker was calibrated based on a 9-point full-screen calibration. Over the course of the session, measurement accuracy was monitored, and if needed, the device was re-calibrated (this was rarely necessary). Data were collected binocularly. Our analyses were based primarily on the right eye data. The right eye data of one participant was problematic, and therefore, the left eye data of the participant were used.

In each trial, a fixation point appeared first at the position of the second character of the first word of the sentence (vertically centered on the screen and about 1.5 inches from the left edge). After fixating on this point, participants pressed a button to bring up a sentence and started to read it. Sentences would not show up if participants were not fixating on this point. After reading the whole sentence, participants clicked the button again. This prompted either the next trial or the display of a comprehension question. Participants gave answers to the comprehension questions by pressing the buttons denoting "yes" and "no," respectively.

3.5. Eye-movement measures

We calculated the eye-movement measures using the in-house software (Braze, 2005), which served to tally gaze measures for each word. We removed fixations shorter than 50 ms, as well as blinks and instances of track-loss. We also excluded the sentence initial and final words from analysis, as a common practice (Kliegl et al., 2004). There remained a total of 15,733 eye-movement observations, covering 358 word types

TABLE 5 Lexical properties of the words in the sentence stimuli.

Name	Label	Mean	SD	Min.	25%	50%	75%	Max.	Skew	Kurtosis
Current word length	Len_W	6.03	2.03	2	4	6	7	14	0.66	0.37
Current word frequency	$Freq_W$	10.16	1.85	3.33	9.03	10.35	11.48	13.91	−0.75	1.03
Previous word length	Len_{W-1}	4.15	2.12	1	3	3	6	13	0.96	0.68
Previous word frequency	$Freq_{W-1}$	13.50	2.93	4.53	11.15	14.03	16.28	17.04	−0.52	−0.59
Next word length	Len_{W+1}	4.19	2.19	1	2	4	5	13	0.99	0.57
Next word frequency	$Freq_{W+1}$	13.09	2.78	3.66	11.00	13.56	15.28	17.04	−0.55	−0.48

Word positions in sentences are excluded here. Word lengths are raw values before mean-centered. Word frequencies are log-transformed type frequencies from the COCA database.

in 72 sentences. The volume of the data is comparable to other eye-tracking studies of individual differences. We focused on five informative, widely-used eye-movement measures (Rayner, 1998):

(1) *First fixation duration*, the duration of the initial fixation a reader makes on a region (word) during first-pass reading. It is typically considered to reflect early stage processes during lexical access (Inhoff, 1984).
(2) *First-pass reading time* (a.k.a. *gaze duration*), the summed duration of all fixations a reader makes on a word before fixating any subsequent word, and before gaze leaves the word for the first time, whether advancing to the next word or regressing to an earlier word. It is often considered to reflect sentence structure, parsing decisions (Rayner et al., 1983; Ferreira and Clifton, 1986), or predictability of words in context (Boston et al., 2008). First fixation duration and first-pass reading time are conditional upon a word receiving a first-pass reading. If a word was initially skipped and thus nominally accrued a zero value for these measures, then that data point was omitted from the following analyses, because we do not wish to infer from word-skipping that a word is not processed at all, or that its processing load is zero (Rayner and Pollatsek, 1989).
(3) *Total reading time*, the sum of all fixations falling again into the current word region. It reflects the integrative effect of both early and late stage processes during lexical access.
(4) *Incidence of first-pass regression*, coding for whether the eye-movement at the end of first pass reading moved back to a previous part of the sentence (= 1), or advanced to a subsequent word (= 0).
(5) *Refixation incidence*, being 1 if a word is refixated after the first-pass, or 0 otherwise.

Measures (1) – (3) are continuous, and (4) and (5) are binary (0/1) to capture possible effects on late stage of processing. Measures (4) and (5) are generally treated as indices of processing load associated with integration difficulty (Rayner et al., 1983, 1989).

Table 6 summarizes the gaze measures, which reflect different, but perhaps overlapping stages of word recognition, text comprehension, and integration during online sentence reading. First fixation duration and first-pass reading time reflect the early stages of print processing involving first encounter of a word by the reader following the default reading direction (left to right in English). By contrast, incidence of first-pass regression and refixation incidence reflect the later stages of print processing involving integration of word information with syntactic and/or discourse context or resolution of ambiguity whenever necessary (Vasishth et al., 2013). Total reading time is a cumulative index of "early" and "late" stages of processing. Individual differences in several components of skilled reading (e.g., decoding, oral reading fluency, vocabulary knowledge, working memory) may have different effects as gauged by these eye-movement measures (Kuperman and Van Dyke, 2011; Nelson Taylor and Perfetti, 2016).

In our dataset, 11,965 out of the total 15,733 eye-movement observations (76.050%) were first-pass eye-movements, and only 3,768 had distinct first fixation durations and first-pass reading times. This indicates that during first-pass reading, most words were fixated exactly once (many words in our simple stimuli sentences were short; see **Table 5**, over half of the words are shorter than 6 characters). Therefore, it is expected that if any factors can exert significant effects during first-pass reading, they might be captured mainly by first-pass reading time, not by first fixation duration. In addition, our stimuli sentences were simple in structure, which might not trigger many regressive eye-movements or second-pass reading in our participants. Therefore, incidence of first-pass regression and refixation incidence might not capture many significant effects, unlike previous studies involving more complex sentence stimuli (Kuperman and Van Dyke, 2011; Kuperman et al., 2018)).

3.6. Analytic approach

We conducted two types of statistical analysis.

First, we used linear and logistic mixed-effects regression models (Baayen, 2008; Quené and van den Bergh, 2008) with crossed random effects to analyze respectively the continuous and categorical eye-movement measures and identify interactions between lexical properties and reading related skills. Mixed-effects models allow for simultaneous

TABLE 6 Summary of the eye-movement measures.

Name	Mean	SD	Min.	25%	50%	75%	Max.	Skew	Kurtosis
First fixation duration	236.98	96.79	52	176	216	272	996	1.92	6.54
First-pass reading time	293.54	145.47	52	192	252	360	1000	1.50	2.73
Total reading time	378.07	237.39	52	212	312	464	3080	2.28	9.39
Incidence of first-pass regression	0.16	0.37	0				1	1.80	1.25
Refixation incidence	0.25	0.43	0				1	1.14	−0.70

The first three measures are continuous, and the other two are binary (0/1).

consideration of multiple covariates, while keeping the between-participants and between-items variance under statistical control (Pinheiro and Bates, 2000; Baayen et al., 2008). Unlike the random forest models used in Kuperman et al. (2018), mixed-effects models can simultaneously address multiple factors having different scales of effect sizes and directly report significance of main effects and/or interactions.

We fit five mixed-effects regression models (Quené and van den Bergh, 2008) targeting the five eye-movement measures, respectively. To reflect the collinearity of a model, we reported the condition number kappa of the model and the maximum variance inflation factor (VIF) of all predictors in the model. A condition number kappa smaller than 10 and a VIF smaller than 5 typically indicate a low degree of collinearity (Kutner et al., 2004).

Each of the five models included 23 fixed effects, consisting of seven lexical properties, four composite and single skill measures, and 12 interactions between each of the skill measures and each of the lexical properties, namely word position in a sentence, word frequency and word length. This approach provides an integrative picture of the effects of multiple skill measures on eye-movement patterns. We controlled the family-wise Type I error probability by setting the critical p value for identifying significance as 0.05/23 ≈0.00217. Given this extremely strict setting of critical p value, we focused on both the significant ($p < 0.00217$) and marginally significant (p is close to 0.00217) factors.

Each model included the same random effect structure, consisting of two intercepts respectively for subject and for word nested under sentence, and one slope of word frequency for subject. In principle, the slope of word length for subject should also be added in each model. However, as shown above, word length was negatively correlated with word frequency, and post-hoc analyses revealed that the separate contributions of word length to the variation in the dependent variables was <1%. Therefore, we excluded this slope in the regression models. In addition, maximal random effect structures involving other types of slopes are theoretically desirable (Barr et al., 2013) and have been applied in recent individual difference studies (e.g., Protopapas and Kapnoula, 2016). However, we did not pursue such complicated models in consideration of practical constraints on model convergence (Bates et al., 2015a).

All the mixed-effects models were implemented using the R packages *lme4* (Bates et al., 2015b) and *lmerTest* (Kuznetsova et al., 2017).

Second, after identifying significant interactions, we continued examining the dynamics of lexical properties and eye-movement measures in individuals having different levels of target skills. Very few existing studies have investigated such dynamics. Our approach proceeded as follows. Given a two-way interaction between a lexical property and a skill measure, we first divided the participants into a high and a low group based on the medium value of the skill measure to ensure the same number of participants in each group. Then, we plotted the eye-movement measure in each group against the lexical property. A cross-group comparation of the correlations between lexical properties and eye-movement measures could reveal the effects of individual skill on online reading behavior. Instead of binary groups, quartile or quintile groups were used in some studies (e.g., Protopapas and Kapnoula, 2016), given enough participants in each group for statistical analysis. To identify correlation, we first fit a nonlinear polynomial regression (*loess*) between the lexical property and the eye-movement measure as the baseline, and then, used widely-adopted regression models in psychological and educational research to quantify the pattern of the correlation. For simplicity, the current study only compared simple linear regression (or logistic regression) and segmented linear regression. For each model, lexical property was treated as an independent variable, and eye-movement measure a dependent one.

Models were compared based on Akaike information criterion (AIC) and mean squared error (MSE). AIC deals with the trade-off between the simplicity and goodness of fit of a model (Akaike, 1974), but AIC alone is less informative when multiple models have similarly high or low AICs (Burnham and Anderson, 2002). In this situation, MSE is referred to, which compromises variance and bias to minimize both (see Equation 1, where obs_i is the observed essay score, pre_i is the predicted score from a model, and n is the number of data points). The *best* model that appropriately reflects the correlation between lexical property and eye-movement measure is the one having smaller AIC and MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(obs_i - pre_i)^2 \quad (1)$$

A recent study examining the correlation between typing speed and writing essay score has used a similar method to identify the dynamics of such correlation (Gong et al., 2022). In that study, additional models like logistic regression and ordinal categorical regression were used for model fitting, but the segmented linear regression remained the best fitting model.

In our study, the segmented regression was implemented using the R package *segmented* (Muggeo, 2008).

4. Results

The analyses were carried out in R 3.2.4 (R Core Team, 2013). The raw data, R codes, and the results can be found at: https://github.com/gtojty/IndDiff_EM.

All the regression models showed a low degree of collinearity; the kappas of these models were all below 10 and the VIFs of the independent factors in these models were all below 5. The significant main effects of lexical properties reported in these models are shown in **Supplementary Table 2** and discussed in **Supplementary material**. No skill measures

showed significant main effects on any eye-movement measures (their p values were all above. 00217), due primarily to the wide spans of the skill measures in our study.

Our study focuses on the interactions whose p values are smaller than (significant) or close to (marginally significant) the threshold. 00217. For the sake of completeness, Tables 7–10 list all the interactions between lexical properties and skill measures having p values below 0.05/5 = 0.01. Effect size (Cohen's d) of each interaction was measured using the lme.dscore function in the R package *EMAtools*.[2] Significant (and marginally significant) interactions are visualized in Figures 1–3. For each interaction, the correlation between the involved lexical property and eye-movement measure in the participants having high and low levels of the involved skill measure can be best described as a segmented linear relation. Below, we discuss these interactions identified in the regression models.

4.1. First fixation duration

Table 7 lists one interaction between word frequency and decoding skill in determining first fixation duration. Its p value is over .00217, so it is not marked as a significant interaction.

2 https://cran.r-project.org/web/packages/EMAtools/index.html

4.2. First-pass reading time

Table 8 shows two interactions on first-pass reading time whose p values are below .01. Given their p values are smaller than (or close to) .00217, they are marked significant (or marginally significant). Figure 1 illustrates these interactions by showing that the correlation between word length and first-pass reading time is contingent on oral comprehension plus vocabulary and verbal working memory.

Figure 1 shows that the sensitivity of first-pass reading time to word length is better described as a segmented linear relation than a simple linear relation: the segmented linear curves well match the baseline loess curve and have smaller AIC and MSE than the linear curve (see Supplementary Table 2). In each panel, the segmented linear curve shows a pivot value of word length, below which the slop of the fitting curve remains small, whereas above which the slope increases, indicating that the participants showed longer first-pass reading time when reading longer words. Between the two panels in each figure, the sensitivity of first-pass reading time to word length exhibits different tendencies.

In Figure 1A, compared to the poor readers having low levels of oral comprehension plus vocabulary (the right panel), for words of the same length, the good readers having high levels of that skill (the left panel) had shorter first-pass reading time. Also, the good readers showed smaller slopes in the segmented linear curve than the poor readers (i.e., 5.585 vs. 8.938 and 19.39

TABLE 7 Interaction on first fixation duration.

Factor	Est.	SE	t	p	d
Decoding × word frequency	3.995	1.480	2.699	0.009	0.744

Its p value is below 0.01 but over 0.00217.

TABLE 8 Interactions on first-pass reading time.

Factor	Est.	SE	t	p	d
Oral comprehension plus vocabulary × word length	**−2.513**	**0.846**	**−2.970**	**0.002**	**−0.045**
Verbal working memory × word length	**2.890**	**0.831**	**3.480**	**0.001**	**0.058**

All listed interactions have p values below 0.01. Interactions having p values below or close to 0.00217 are bolded.

TABLE 9 Interactions on total reading time.

Factor	Est.	SE	t	p	d
Oral reading fluency × word position	**2.013**	**0.661**	**3.040**	**0.002**	**0.051**
Verbal working memory × word position	**−2.196**	**0.732**	**−3.000**	**0.002**	**−0.050**
Verbal working memory × word length	3.885	1.353	2.870	0.004	0.048

All these interactions have p values below 0.01. Interactions having p values below or close to 0.00217 are bolded.

TABLE 10 Interaction on incidence of first-pass regression.

Factor	Est.	SE	z	p	d
Decoding × word position	**0.031**	**0.009**	**3.276**	**0.001**	**0.055**

Its p value is below 0.00217. Statistically significant factors are shown in bold.

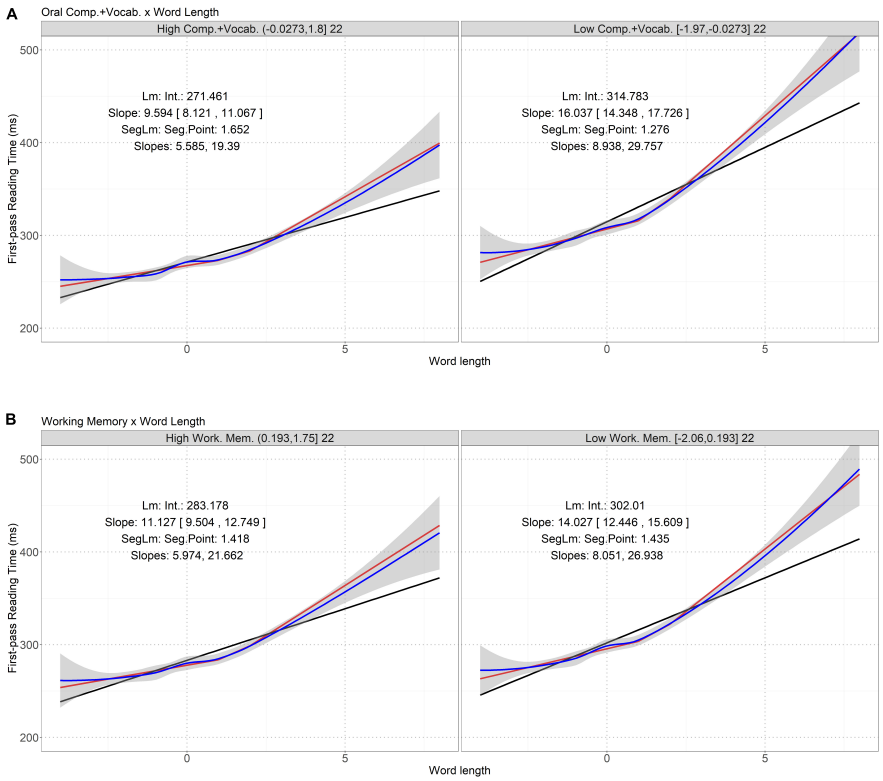


FIGURE 1
Interactions between word length and oral comprehension plus vocabulary **(A)** and verbal working memory **(B)** on first-pass reading time. Word length is mean-centered. The two panels in each figure represent the high and low skill groups. The titles of the panels show the level ranges (within round or square brackets) of the skill measure in the two groups and the numbers of participants in these groups. In each panel, the blue line is the loess fitting curve and the shaded area is standard error. The black line is the linear regression fitting curve ("Lm"). "Int." shows the interception (β_0), and "Slope" the slope (β_1). Numbers in square brackets are 95% confidence interval of the slope. The red line is the segmented linear regression fitting curve ("SegLm"). "Seg.Point" shows the pivot point at word length, below and above which the slopes of the curve are distinct (see "Slopes"). See **Supplementary Table 3A** for AIC and MSE of these models. The segmented linear models have the smallest AIC and MSE closest to that of the loess regressions.

vs. 29.757), indicating that the good readers were less sensitive to word length. Finally, the pivot points of word length were similar in the poor (1.652) and good (1.276) readers.

In Figure 1B, similarly, compared to the good readers having high levels of verbal working memory, the poor readers having low levels of that skill spent relatively more time in reading long words, and for both long and short words, their first-pass reading times remained more sensitive to word length (shown by the slopes of the segmented linear curves, 26.938 vs. 21.662 and 8.051 vs. 5.974). Nonetheless, the pivot points of word length in the poor and good readers were similar (1.435 vs. 1.418).

4.3. Total reading time

Table 9 shows three interactions on total reading time whose *p* values are below .01, two of which are marked as marginally significant and visualized in Figure 2.

Figure 2 illustrates a segmented linear relation between total reading time and word position in a sentence. Total reading time drops when the participants read the first few words in a sentence, and then, increases when they read the latter words in a sentence. The negative and positive slopes of the segmented linear fitting curves clearly reflect this bifurcating tendency.

In Figure 2A, compared to the good readers having high levels of oral reading fluency, the total reading time of the poor readers having low levels of that skill is generally longer, and it is more sensitive to the beginning words in a sentence, as shown by the more negative slopes (−27.098 vs. −11.428) below the pivot points of word position. However, the smaller positive slopes (3.008 vs. 17.01) above the pivot points suggest that the total reading time of the poor readers is less sensitive to the latter words in a sentence. In addition, the pivot points of word position increases from −0.715 in the poor readers to 1.375 in the good readers.

In Figure 2B, compared to the good readers having high levels of verbal working memory, the total reading time of the

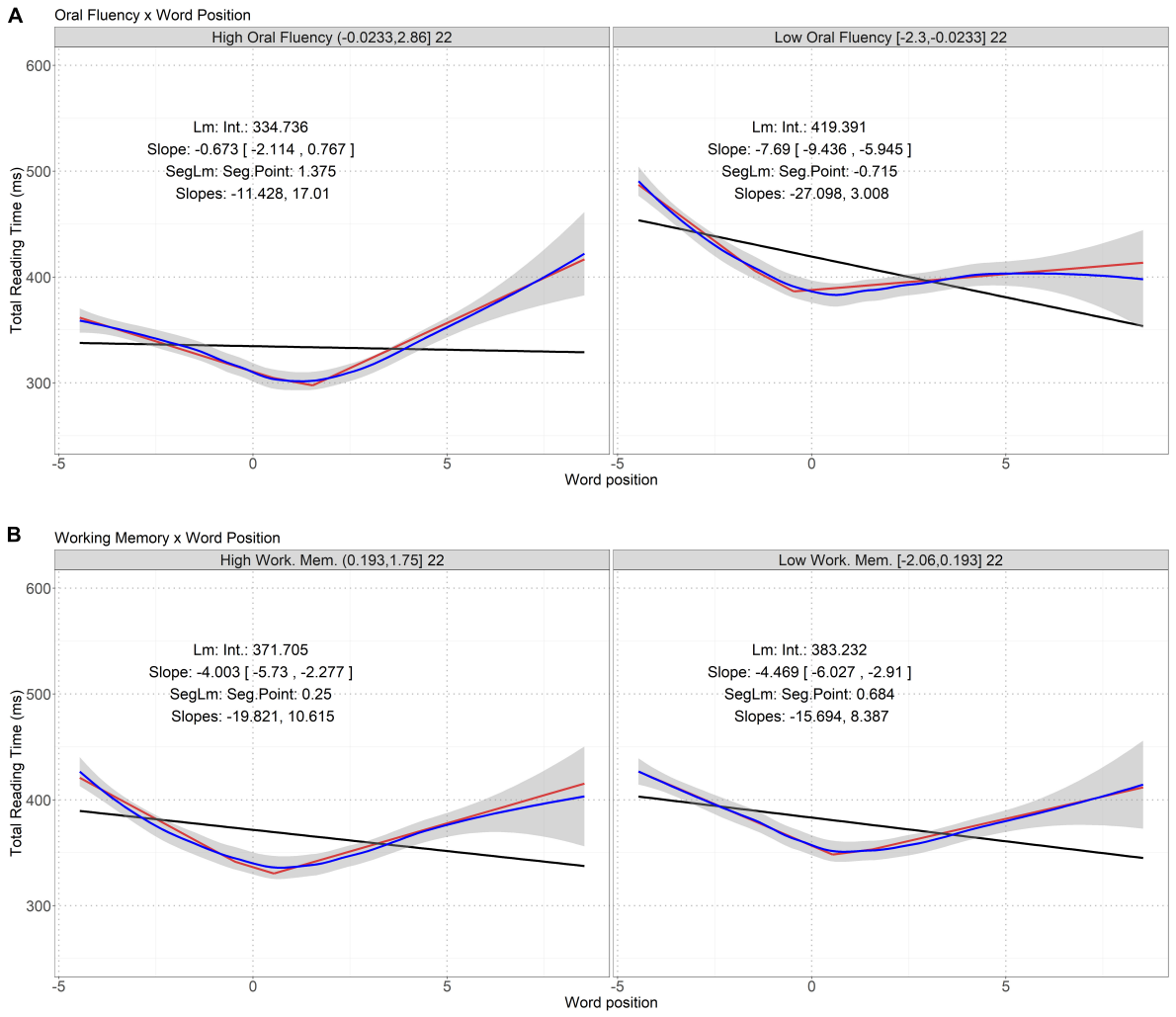


FIGURE 2
Interactions between word position and oral fluency **(A)** and verbal working memory **(B)** on total reading time. Word position is mean-centered. The two panels in each figure represent the high and low skill groups. See **Supplementary Table 3B** for AIC and MSE of different models, which shows the segmented linear models have the smallest AIC and MSE closest to that of the loess regressions.

poor readers having low levels of that skill is less sensitive to word position in a sentence, as shown by the smaller absolute slopes both below (−15.694 vs. −19.821) and above (8.387 vs. 10.615) the pivot points of word position. In addition, the pivot points in the two panels drop from 0.684 in the poor readers to 0.250 in the good readers.

A comparison of Figures 1, 2 reveals that verbal working memory casts its influence on first-pass reading time via interaction with word length and total reading time via interaction with word position. To be specific, compared to the poor readers having low levels of verbal working memory, the first-pass reading time of the good readers is less sensitive to word length, but their total reading time is more sensitive to word position.

4.4. Incidence of first-pass regression

Table 10 shows that the interaction between decoding and word position had a p value below .00217. Figure 3 visualizes this significant interaction.

Figure 3 shows a segmented linear relation between first-pass regression and word position in a sentence. The probability of regression during the first-pass reading starts to increase when the participants read the latter words in a sentence. Compared to the poor readers having low levels of decoding, the probability of regression during the first-pass reading of the good readers increases a lot on the latter words in a sentence, as shown by bigger slopes (.042 vs. .016) above the pivot points of word position. The pivot points of word position are similar in the poor (2.108) and good (2.537) readers.

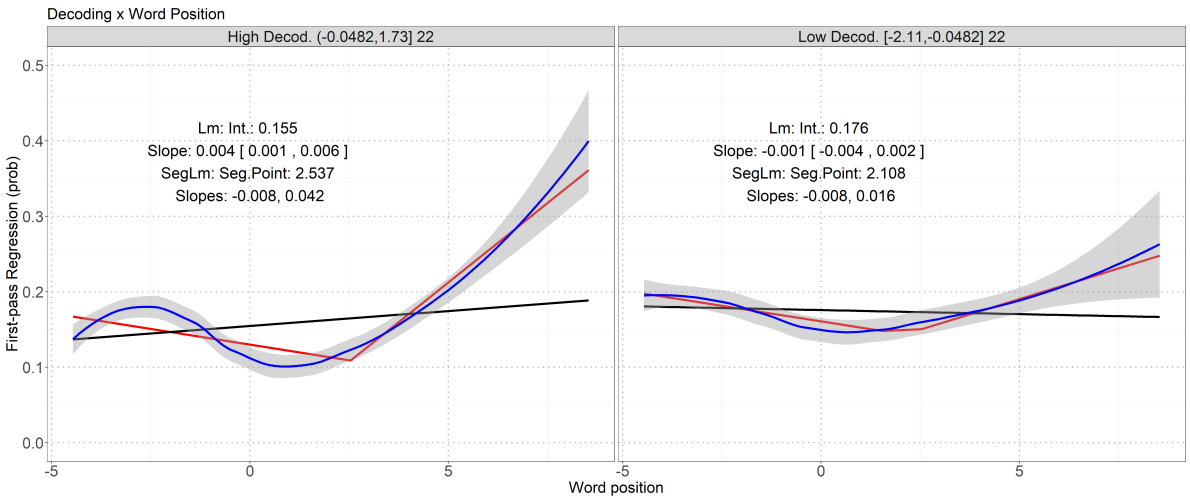


FIGURE 3
Interaction between word position and decoding on incidence of first-pass regression. Word position is mean-centered. The two panels represent the high and low decoding groups. See **Supplementary Table 3C** for AIC and MSE of different models, which shows the segmented linear models have the smallest AIC and MSE closest to that of the loess regressions. "Lm" here is logistic regression. Note that in the left panel, it seems that the loess regression fitting curve also has a pivot point near the lower bound of word position. Since it is much closer to the boundary, there are insufficient data points for the segmented linear model to identify it as a pivot point.

4.5. Refixation incidence

No interactions on refixation incidence have p values below .01.

5. Discussion

5.1. Effects of interactions between language and literacy skills and lexical properties on online reading behavior

Previous studies have reported significant main effects of some of the language and literacy skills discussed in this paper, or bigger effect sizes of these skills than those of lexical properties (e.g., Kuperman and Van Dyke, 2011). However, in our analyses, main effects of skill measures never reach statistical significance, though those of lexical properties often do. The effect sizes of the skill measures are also smaller than those of lexical properties. This is because that our study focused on individuals with a much wider range of language and literacy skills; only those having the highest scores of the skill measures were comparable to university students (cf. Braze et al., 2007). Such wide range of individual differences in the skill measures could result in insignificance and low effect sizes of the measures on online reading behavior. These findings can enrich existing evidence and trigger revisits on the theoretical discussions of individual differences and their roles in reading process and outcome (comprehension) (Bennink and Spoelstra, 1979; Bleckley et al., 2003).

Although lacking direct influence on online reading behavior, some of the language and literacy related skills could significantly influence online reading behavior via interactions with lexical properties. Our study showed that oral comprehension, vocabulary, verbal working memory, oral reading fluency, and decoding could predict online reading patterns via interactions with word length or position in a sentence. We also compared the effects of the interactions involving these skills on online reading patterns between the good and poor readers with respect to these skills.

To be specific, oral comprehension and vocabulary interact with word length to predict first-pass reading time (see **Figure 1A**); readers with good oral comprehension skill and vocabulary knowledge could efficiently process words with various lengths, thus being less troubled by long words during first-pass reading. First-pass reading time arguably reflects the duration of lexical processing, including recognition of orthographic or phonological features of a word and retrieval of semantic information from memory once attention is allocated to the word (Inhoff, 1984). This finding contributes to recent discussions on whether vocabulary knowledge could influence reading comprehension over and above the effect of language comprehension including listening comprehension (Braze et al., 2007, 2016; Tunmer and Chapman, 2012; Protopapas et al., 2013). At the early stage of print processing vocabulary knowledge already helps good readers efficiently reduce first-pass reading time on words of various lengths.

Verbal working memory presumably affects the rate at which word information is assimilated during first-pass reading, especially on long words. As shown in **Figure 1B**, the first-pass reading time of the good readers with high levels of verbal working memory are less sensitive to word length than the poor readers. In addition, verbal working memory helps predict total reading time via interaction with word position (see **Figure 2B**). Total reading time reflects the integration of early and late processing during lexical access. Word position in a sentence is a context-dependent property. A general increase in total reading time on words toward the end of a sentence reflects so-called wrap-up effects (Rayner et al., 2000; Warren et al., 2009). In our study, such effects became more explicit in readers having high levels of verbal working memory; efficient verbal working memory reduces the processing time for the first few words of a sentence but induces more wrap-up effects towards the end of a sentence.

Oral reading fluency interacts with word position to predict total reading time (see **Figure 2A**); a high level of this skill is associated with a less sensitivity to the first few words in a sentence, but more sensitivity to latter words in a sentence, in line with the wrap-up effects. In addition, less fluent readers generally have more difficulty in processing individual words and integrating word semantics with context, and hence spend more time reading a few words of a sentence; by contrast, more fluent readers spend less time reading words in a sentence, especially those near the beginning or in the middle of a sentence. These findings are in line with and complement the existing theories on oral and/or silent reading fluency (Fuchs et al., 2001; Tilstra et al., 2009; Kim et al., 2011; Silverman et al., 2012; Ashby et al., 2013). Furthermore, as shown in **Figure 2**, there is no monotonic change of the correlation between word position and total reading time. This indicates that the effects of oral reading fluency and verbal working memory on regulating online reading patterns are complex, possibly also subject to other factors.

Decoding skill interacts with word position to predict probability of first-pass regression (see **Figure 3**); good decoders tended to have more regressive reading when reading words towards the end of a sentence, reflecting their sentence decoding processes. Early studies have reported the effects of decoding on early (first-pass reading time) and overall (total reading time) reading and re-reading probability (Kuperman and Van Dyke, 2011; Nash and Heath, 2011; Kuperman et al., 2018). In our study, the effect of decoding on re-reading probability was fulfilled via an interaction with word position. All these are in line with the claims that decoding skill is among the key factors in lexical access (Barth et al., 2009; Hulme and Snowling, 2012) and provide evidence for VET (Perfetti, 1985; Shankweiler and Crain, 1986) and LQH (Perfetti and Hart, 2002; Perfetti, 2007) by showing how decoding influences reading processes.

5.2. Segmented linear dynamics of the correlation between lexical properties and eye-movement measures

In addition to confirming that language and literacy skills can influence online reading behavior indirectly via interactions with lexical properties, our study further investigated the dynamics of the correlation between lexical properties and eye-movement measures regulated by particular individual skills. Our quantitative analyses revealed that such dynamics cannot be simply described as a linear relation; instead, many of the correlations follow a segmented linear relation, with at least two distinct slopes throughout the values of the relevant lexical properties. Some of the dynamics are monotonic (see **Figure 1**), with positive and increasing slopes around long words, whereas others are not (see **Figures 2**, **3**), with a transition from a negative to a positive slope. The observed segmented linear relations suggest a complex effect of key language and literacy skills on regulating reading patterns via interactions with word length or position. Between the good and poor readers based on some skills, the durations of reading time are different, so are the sensitivity of reading time or regression probability to word length or position. In addition, the pivot values of word length or position in the segmented linear correlations indicate a transition of the degree of correlation. Note that in many cases, the pivot points are not close to the mean value 0, so arbitrary binary segmentation based on word length or position (Kuperman et al., 2018) cannot clearly reveal such dynamics. This dynamics echoes the effects of interactions between lexical properties and skill measures on online reading behavior: due to individual skills, the unimodal associations between eye-movement patterns and lexical properties are broken, the degrees of associations become different when the values of lexical properties are below or above the pivot points, and the high and low levels of the skills further influence the pivot lexical property values and the degrees of associations below and above the pivot values.

The observed dynamics in all these aspects can lead to more comprehensive theories on the dynamic relations between individual skills, text properties, and reading process. For example, some theories of reading (Perfetti and Hart, 2002) and empirical studies (Johnston and Kirby, 2006; Savage, 2006) have challenged the linear assumption between decoding and reading outcomes like reading comprehension. For example, Johnston and Kirby (2006) showed that naming speed, a measure of decoding skill, had its primary effect on less able readers. A recent study of reading assessment has shown distinct relations between decoding skill and comprehension scores between good and poor decoders in Grades 5 to 10 (Wang et al., 2019). Some eye-movement studies have revealed close relations between components of decoding (e.g., phonemic awareness) and other skills (e.g., reading fluency) (Barth et al., 2009; Ashby et al., 2013). Our study enriched the findings in

this line of research by visualizing the segmented linear relations between lexical properties and online reading behavior, which are manipulated by individual differences in individual language and literacy skills. This study can also inspire more empirical studies to further investigate what factors help shape the slopes and pivot values in the segmented linear models.

6. Conclusion

This study investigated the eye-movement data of simple sentence reading from 44 young adults in high schools, adult education centers, community colleges, or neighborhood communities. A total of six domains of individual differences, plus age, were tested to assess their effects via themselves and interactions with lexical properties on online reading behavior. Three of these domains tap into components of reading ability: reading comprehension, decoding skill, and oral reading fluency. The other three tap into domains not reading specific: listening comprehension, vocabulary, and verbal working memory. By evaluating the effect of each domain while controlling for the others, we identified a series of interactions between properties of text (length and position) and skills of readers (oral comprehension, vocabulary, verbal working memory, oral reading fluency, and decoding), which manipulated both the early and late stages of online reading process as gauged by eye-movement measures (first-pass reading time, total reading time, and first-pass regression). We also visualize segmented linear dynamics of the effects of these interactions on online reading patterns. All these findings speak to the necessity of incorporating interactions between lexical properties and reading-related skills to enrich empirical evidence, extend and refine theories about reading outcomes and processes, and trigger new theories or hypotheses on how language and literacy skills interact with lexical properties to influence reading process.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/gtojty/IndDiff_EM.

Ethics statement

The studies involving human participants were reviewed and approved by Haskins Laboratories. Informed consent was obtained from the participants of at least 18 years old; for those under 18, the participants provided assent and their parents or guardians signed written permissions.

Author contributions

TG designed and implemented the experiment. Both author collected and analyzed the data and wrote and edited the manuscript.

Conflict of interest

TG was employed by Google.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1006662/full#supplementary-material

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Ashby, J., Dix, H., Bontrager, M., Dey, R., and Archer, A. (2013). Phonemic awareness contributes to text reading fluency: Evidence from eye movements. *Sch. Psychol. Rev.* 42, 157–170. doi: 10.1080/02796015.2013.12087482

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511801686

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

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., et al. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *Sch. Psychol. Rev.* 37, 18–37. doi: 10.1080/02796015.2008.12087905

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure in mixed effects models: Keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Barth, A., Catts, H., and Anthony, J. (2009). The component skills underlying reading fluency in adolescent readers: A latent variable analysis. *Read. Writ.* 22, 567–590. doi: 10.1007/s11145-008-9125-y

Bates, D., Kliegl, R., Vasishth, S., and Baayen, R. H. (2015a). Parsimonious mixed models. *arXiv* [Preprint]. arXiv:1506.04967v1.

Bates, D., Machler, M., Bolker, B., and Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Beck, I. L., Perfetti, C. A., and McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *J. Educ. Psychol.* 74, 506–521. doi: 10.1037/0022-0663.74.4.506

Bennink, C. D., and Spoelstra, T. (1979). Individual differences in field articulation as a factor in language comprehension. *J. Res. Pers.* 13, 480–489. doi: 10.1016/0092-6566(79)90010-2

Birch, S. L., and Rayner, K. (1997). Linguistic focus affects eye movements during reading. *Mem. Cogn.* 25, 653–660. doi: 10.3758/BF03211306

Bleckley, M., Durso, F. T., Crutchfield, J. M., Engle, R. W., and Khanna, M. M. (2003). Individual differences in working memory capacity predict visual attention allocation. *Psychon. Bull. Rev.* 10, 884–889. doi: 10.3758/BF03196548

Boston, M. F., Hale, J., Kliegl, R., and Patil, U. (2008). Parsing costs as predictors of reading dificulty: An evaluation using the potsdam sentence corpus. *J. Eye Mov. Res.* 2, 1–12. doi: 10.16910/jemr.2.1.1

Box, G. E. P., and Cox, D. R. (1964). An Analysis of Transformations. *J. R. Stat. Soc. Series B Methodol.* 26, 211–252. doi: 10.1111/j.2517-6161.1964.tb00553.x

Braze, D. (2005). *Fixation analysis (Version 3.01)*.

Braze, D., Katz, L., Magnuson, J. S., Mencl, W. E., Tabor, W., Van Dyke, J. A., et al. (2016). Vocabulary does not complicate the simple view of reading. *Read. Writ.* 29, 435–451. doi: 10.1007/s11145-015-9608-6

Braze, D., Mencl, W. E., Shankweiler, D. P., Tabor, W., and Schultz, A. (2006). "Skill-related differences in the online reading behavior of young adults: Evidence from eye-movements," in *Proceeding of the talk given at the 13th annual meeting of the society for the scientific study of reading*, Vancouver.

Braze, D., Mencl, W. E., Tabor, W., Pugh, K. R., Constable, R. T., Fulbright, R. K., et al. (2011). Unification of sentence processing via ear and eye: An fMRI study. *Cortex* 47, 416–431. doi: 10.1016/j.cortex.2009.11.005

Braze, D., Tabor, W., Shankweiler, D. P., and Mencl, W. E. (2007). Speaking up for vocabulary: Reading skill differences in young adults. *J. Learn. Disabil.* 40, 226–243. doi: 10.1177/00222194070400030401

Brysbaert, M., and Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *J. Cogn.* 1:9. doi: 10.5334/joc.10

Burnham, K. P., and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*, 2rd Edn. London: Springer-Verlag.

Calvo, M. G. (2004). Relative contribution of vocabulary knowledge and working memory span to elaborative inferences in reading. *Learn. Individ. Differ.* 15, 53–65. doi: 10.1016/j.lindif.2004.07.002

Calvo, M. G., Estevez, A., Dowens, M. G., and Calvo, M. G. (2003). Time course of elaborative inferences in reading as a function of prior vocabulary knowledge. *Learn. Instr.* 13, 611–631. doi: 10.1016/S0959-4752(02)00055-5

Catts, H. W., Adlof, S. M., and Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *J. Speech Lang. Hear. Res.* 49, 278–293. doi: 10.1044/1092-4388(2006/023)

Chik, P., Ho, C., Yeung, P.-s, Wong, Y.-k, Chan, D., Chung, K., et al. (2010). Contribution of discourse and morphosyntax skills to reading comprehension in Chinese dyslexic and typically developing children. *Ann. Dyslexia* 62, 19–21. doi: 10.1007/s11881-011-0062-0

Clifton, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., and Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *J. Mem. Lang.* 49, 317–334. doi: 10.1016/S0749-596X(03)00070-6

Cohen, J. D., MacWhinney, B., Flatt, M., and Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behav. Res. Methods Instr. Comput.* 25, 257–271. doi: 10.3758/BF03204507

Corkin, S. (1974). Serial-ordering deficits in inferior readers. *Neuropsychologia* 12, 347–354. doi: 10.1016/0028-3932(74)90050-5

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.* 19, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Duffy, S. A., Morris, R. K., and Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *J. Mem. Lang.* 27, 429–446. doi: 10.1016/0749-596X(88)90066-6

Dunn, L. M., and Dunn, L. M. (1997). *Peabody picture vocabulary test*, 3rd Edn. Circle Pines, MN: American Guidance Service, Inc. doi: 10.1037/t15145-000

Estevez, A., and Calvo, M. G. (2000). Working memory capacity and time course of predictive inferences. *Memory* 8, 51–61. doi: 10.1080/096582100387704

Ferreira, F., and Clifton, C. (1986). The independence of syntactic processing. *J. Mem. Lang.* 25, 348–368. doi: 10.1016/0749-596X(86)90006-9

Fox, J., and Weisberg, S. (2011). *An R companion to applied regression*, 2nd Edn. Thousand Oaks, CA: Sage.

Frisson, S., and McElree, B. (2008). Complement coercion is not modulated by competition: Evidence from eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 1–11. doi: 10.1037/0278-7393.34.1.1

Fuchs, L. S., Fuchs, D., Hosp, M. K., and Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Sci. Stud. Read.* 5, 239–256. doi: 10.1207/S1532799XSSR0503_3

Gathercole, S. E., and Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *J. Mem. Lang.* 28, 200–213. doi: 10.1016/0749-596X(89)90044-2

Gathercole, S. E., Service, E., Hitch, G. J., Adams, A. M., and Martin, A. J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Appl. Cogn. Psychol.* 13, 65–77. doi: 10.1002/(SICI)1099-0720(199902)13:1<65::AID-ACP548>3.0.CO;2-O

Gong, T., Zhang, M., and Li, C. (2022). Association of keyboarding fluency and writing performance in online-delivered assessment. *Assess. Writ.* 51:100575. doi: 10.1016/j.asw.2021.100575

Gough, P. B., and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial Spec. Educ.* 7, 6–10. doi: 10.1177/074193258600700104

Gupta, P. (2006). Nonword repetition, phonological storage, and multiple determinations. *Appl. Psycholinguist.* 27, 564–568. doi: 10.1017/S0142716406260399

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83. doi: 10.1017/S0140525X0999152X

Hoover, W. A., and Gough, P. B. (1990). The simple view of reading. *Read. Writ.* 2, 127–160. doi: 10.1007/BF00401799

Hulme, C., and Snowling, M. J. (2012). Learning to read: What we know and what we need to understand better. *Child Dev. Perspect.* 7, 1–5. doi: 10.1111/cdep.12005

Inhoff, A. W. (1984). Two Stages of word processing during eye fixations in the reading of prose. *J. Verbal Learn. Verbal Behav.* 23, 612–624. doi: 10.1016/S0022-5371(84)90382-7

Johnston, T. C., and Kirby, J. R. (2006). The contribution of naming speed to the simple view of reading. *Read. Writ.* 19, 339–361. doi: 10.1007/s11145-005-4644-2

Joseph, H. S. S. L., Nation, K., and Liversedge, S. P. (2013). Using eye movements to investigate word frequency effects in children's sentence reading. *Sch. Psychol. Rev.* 42, 207–222. doi: 10.1080/02796015.2013.12087485

Kim, Y.-S., Wagner, R. K., and Foster, E. (2011). Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Sci. Stud. Read.* 15, 338–362. doi: 10.1080/10888438.2010.493964

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* 16, 262–284. doi: 10.1080/09541440340000213

Kukona, A., Braze, D., Johns, C. L., Mencl, W. E., Van Dyke, J. A., Magnuson, J. S., et al. (2016). The real-time prediction and inhibition of linguistic outcomes: Effects of language and literacy skill. *Acta Psychol.* 171, 72–84. doi: 10.1016/j.actpsy.2016.09.009

Kuperman, V., and Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *J. Mem. Lang.* 65, 42–73. doi: 10.1016/j.jml.2011.03.002

Kuperman, V., and Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 802–823. doi: 10.1037/a0030859

Kuperman, V., Matsuki, K., and Van Dyke, J. A. (2018). Contributions of reader- and text-level characteristics to eye-movement patterns during passage reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 1687–1713. doi: 10.1037/xlm0000547

Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). *Applied linear regression models*, 4th Edn. Burr Ridge: McGraw-Hill Irwin.

Kuznetsova, A., Brockhoff, P. B., and Christiansen, R. H. B. (2017). lmerTest package: Tests in linear mixed-effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Macaruso, P., and Shankweiler, D. P. (2010). Expanding the simple view of reading in accounting for reading skills in community college students. *Read. Psychol.* 31, 454–471. doi: 10.1080/02702710903241363

Manis, F., and Freedman, L. (2001). "The relatiosnhip of naming to multiple reading measures in disabled and non-disabled normal readers," in *Dyslexia, fluency and the brain*, ed. M. Wolf (Walmgate: York Press), 65–92.

Marcotte, A. M., and Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *J. Sch. Psychol.* 47, 315–335. doi: 10.1016/j.jsp.2009.04.003

Markwardt, F. C. (1998). *Peabody individual achievement test–revised*. Circle Pines, MN: American Guidance Service, Inc. doi: 10.1037/t15139-000

McCarron, S. P., and Kuperman, V. (2021). Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behav. Res. Methods* 53, 2226–2237.

Miller, J., and Schwanenflugel, P. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Read. Res. Q.* 43, 336–354. doi: 10.1598/RRQ.43.4.2

Muggeo, V. (2008). Segmented: An r package to fit regression models with broken-line relationships. *R. News* 8, 20–25.

Nash, H., and Heath, J. (2011). The role of vocabulary, working memory and inference making ability in reading comprehension in down syndrome. *Res. Dev. Disabil.* 32, 1782–1791. doi: 10.1016/j.ridd.2011.03.007

Nelson Taylor, J., and Perfetti, C. A. (2016). Eye movements reveal readers' lexical quality and reading experience. *Read. Writ.* 29, 1069–1103. doi: 10.1007/s11145-015-9616-6

Ouellette, G., and Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Read. Writ.* 23, 189–208. doi: 10.1007/s11145-008-9159-1

Perfetti, C. A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Sci. Stud. Read.* 11, 357–383. doi: 10.1080/10888430701530730

Perfetti, C. A., and Hart, L. (2002). "The lexical quality hypothesis," in *Precursors of functional literacy*, ed. L. Verhoeven (Philadelphia, PA: John Benjamins), 189–213. doi: 10.1075/swll.11.14per

Perfetti, C. A., and Lesgold, A. M. (1977). "Discourse comprehension and sources of individual differences," in *Cognitive processes in comprehension*, eds M. A. Just and P. A. Carpenter (Hillsdale, NJ: Lawrence Erlbaum Assoc).

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *J. Consum. Res.* 28, 450–461. doi: 10.1086/323732

Petscher, Y., and Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *J. Sch. Psychol.* 49, 107–129. doi: 10.1016/j.jsp.2010.09.004

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer-Verlag. doi: 10.1007/978-1-4419-0318-1

Protopapas, A., and Kapnoula, E. C. (2016). Short-term and long-term effects on visual word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 42, 542–565. doi: 10.1037/xlm0000191

Protopapas, A., Mouzaki, A., Sideridis, G. D., Kotsolakou, A., and Simos, P. G. (2013). The role of vocabulary in the context of the simple view of reading. *Read. Writ. Q.* 29, 168–202. doi: 10.1080/10573569.2013.758569

Psychological Corporation (1999). *Wechsler abbreviated scale of intelligence*. San Antonio: Harcourt Brace & Co.

Quené, H., and van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *J. Mem. Lang.* 59, 413–425. doi: 10.1016/j.jml.2008.02.002

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Radach, R., and Kennedy, A. (2012). Eye movements in reading: Some theoretical context. *Q. J. Exp. Psychol.* 66, 429–452. doi: 10.1080/17470218.2012.750676

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

Rayner, K. (2009b). Eye movements in reading: Models and data. *J. Eye Mov. Res.* 2, 1–10. doi: 10.16910/jemr.2.5.2

Rayner, K. (2009a). Eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol.* 62, 1457–1506. doi: 10.1080/17470210902816461

Rayner, K., Abbott, M. J., and Plummer, P. (2015). "Individual differences in perceptual processing and eye movements in reading," in *Handbook of individual differences in reading: Reader, text and context*, ed. P. Afflerbach (New York, NY: Informa UK Limited), 348–363.

Rayner, K., and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Mem. Cogn.* 14, 191–201. doi: 10.3758/BF03197692

Rayner, K., and Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.

Rayner, K., Carlson, M., and Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *J. Verbal Learn. Verbal Behav.* 22, 358–374. doi: 10.1016/S0022-5371(83)90236-0

Rayner, K., Kambe, G., and Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *Q. J. Exp. Psychol. Hum. Exp. Psychol.* 53a, 1061–1080. doi: 10.1080/713755934

Rayner, K., Pollatsek, A., Ashby, J., and Clifton, C. (2012). *The psychology of reading*. New York, NY: Psychology Press. doi: 10.4324/9780203155158

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., and Clifton, C. Jr. (1989). Eye movements and on-line language comprehension processes. *Lang. Cogn. Process.* 4, 21–49. doi: 10.1080/01690968908406362

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., and Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *J. Sch. Psychol.* 47, 427–469. doi: 10.1016/j.jsp.2009.07.001

Savage, R. (2006). Reading comprehension is not always the product of nonsense word decoding and linguistic comprehension: Evidence from teenagers who are extremely poor readers. *Sci. Stud. Read.* 10, 143–164. doi: 10.1207/s1532799xssr1002_2

Shankweiler, D. P., and Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition* 24, 139–168. doi: 10.1016/0010-0277(86)90008-9

Shankweiler, D. P., Lundquist, E., Katz, L., Stuebing, K. K., Fletcher, J. M., Brady, S., et al. (1999). Comprehension and decoding: Patterns of association in children with reading difficulties. *Sci. Stud. Read.* 3, 69–94. doi: 10.1207/s1532799xssr0301_4

Shankweiler, D. P., Mencl, W. E., Braze, D., Tabor, W., Pugh, K. R., and Fulbright, R. K. (2008). Reading differences and brain: Cortical integration of speech and print in sentence processing varies with reader skill. *Dev. Neuropsychol.* 33, 745–776. doi: 10.1080/87565640802418688

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition* 55, 151–218. doi: 10.1016/0010-0277(94)00645-2

Silverman, R. D., Speece, D. L., Harring, J. R., and Ritchey, K. D. (2012). Fluency has a role in the simple view of reading. *Sci. Stud. Read.* 17, 108–133. doi: 10.1080/10888438.2011.618153

Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. Santa Monica, CA: RAND.

Stanovich, K. E., and Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Mem. Cogn.* 20, 51–68. doi: 10.3758/BF03208254

Swart, N. M., Muijselaar, M., Steenbeek-Planting, E. G., Droop, M., de Jong, P. F., and Verhoeven, L. (2016). Differential lexical predictors of reading comprehension in fourth graders. *Read. Writ.* 30, 489–507. doi: 10.1007/s11145-016-9686-0

Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., and Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *J. Res. Read.* 32, 383–401. doi: 10.1111/j.1467-9817.2009.01401.x

Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. *Mem. Cognit.* 35, 1107–1121. doi: 10.3758/BF03193482

Tunmer, W. E., and Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *J. Learn. Disabil.* 45, 453–466. doi: 10.1177/0022219411432685

Valle, A., Binder, K. S., Walsh, C. B., Nemier, C., and Bangs, K. E. (2013). Eye movements, prosody, and word frequency among average-and high-skilled second-grade readers. *Sch. Psychol. Rev.* 42, 171–190. doi: 10.1080/02796015.2013.12087483

Vasishth, S., von, der Malsburg, T., and Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdiscip. Rev. Cogn. Sci.* 4, 125–134. doi: 10.1002/wcs.1209

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., and Pearson, N. A. (2010). *Test of silent reading efficiency and comprehension*. Austin, TX: Pro-Ed.

Wang, Z., Sabatini, J., O'Reilly, T., and Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *J. Educ. Psychol.* 111, 387–401. doi: 10.1037/edu0000302

Warren, T., White, S. J., and Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition* 111, 132–137. doi: 10.1016/j.cognition.2008.12.011

Wiederholt, J. L., and Bryant, B. R. (2001). *Gray oral reading test (GORT)*, 4th Edn. Austin, TX: Pro-Ed.

Wolf, M., O'Rourke, G. A., Gidney, C., Lovett, M., Cirino, P., and Morris, R. K. (2002). The second deficit: An investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. *Read. Writ.* 15, 43–72. doi: 10.1023/A:1013816320290

Woodcock, R. W., McGrew, K. S., and Mather, N. (2001). *Woodcock-johnson III tests of achievement*. Itasca, IL: Riverside Publishing.

Wurm, L. H., and Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *J. Mem. Lang.* 72, 37–48. doi: 10.1016/j.jml.2013.12.003

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fpsyg.2022.1028824

OPEN ACCESS

EDITED BY
Xiaowei Zhao,
Emmanuel College, United States

REVIEWED BY
Michael Wolmetz,
Johns Hopkins University,
United States
Christoph Aurnhammer,
Saarland University, Germany
Nicolas Dirix,
Ghent University, Belgium

*CORRESPONDENCE
Nora Hollenstein
nora.hollenstein@hum.ku.dk

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 26 August 2022
ACCEPTED 20 December 2022
PUBLISHED 12 January 2023

CITATION
Hollenstein N, Tröndle M, Plomecka M, Kiegeland S, Özyurt Y, Jäger LA and Langer N (2023) The ZuCo benchmark on cross-subject reading task classification with EEG and eye-tracking data. *Front. Psychol.* 13:1028824. doi: 10.3389/fpsyg.2022.1028824

The ZuCo benchmark on cross-subject reading task classification with EEG and eye-tracking data

Nora Hollenstein[1]*, Marius Tröndle[2], Martyna Plomecka[2], Samuel Kiegeland[3], Yilmazcan Özyurt[3], Lena A. Jäger[4,5] and Nicolas Langer[2]

[1]Center for Language Technology, University of Copenhagen, Copenhagen, Denmark, [2]Department of Psychology, University of Zurich, Zurich, Switzerland, [3]Department of Computer Science, ETH Zurich, Zurich, Switzerland, [4]Department of Computational Linguistics, University of Zurich, Zurich, Switzerland, [5]Department of Computer Science, University of Potsdam, Potsdam, Germany

We present a new machine learning benchmark for reading task classification with the goal of advancing EEG and eye-tracking research at the intersection between computational language processing and cognitive neuroscience. The benchmark task consists of a cross-subject classification to distinguish between two reading paradigms: normal reading and task-specific reading. The data for the benchmark is based on the Zurich Cognitive Language Processing Corpus (ZuCo 2.0), which provides simultaneous eye-tracking and EEG signals from natural reading of English sentences. The training dataset is publicly available, and we present a newly recorded hidden testset. We provide multiple solid baseline methods for this task and discuss future improvements. We release our code and provide an easy-to-use interface to evaluate new approaches with an accompanying public leaderboard: www.zuco-benchmark.com.

KEYWORDS

reading task classification, eye-tracking, EEG, machine learning, reading research, cross-subject evaluation

1. Introduction

Reading plays a fundamental role in the acquisition of information (e.g., encyclopedias) and communication (e.g., emails). As we read, our eyes gaze through the written sentences in a sequence of fixations and high-velocity saccades to extract visual information which are forwarded to the brain to obtain meaning. Thus, assessing where a person looks during reading while recording brain activity non-invasively with electroencephalography (EEG) provides powerful behavioral and physiological measures for cognitive neuroscience to further the understanding of human language processing. Most previous experimental reading research has used hand-picked reading materials in highly controlled experimental settings (Brennan, 2016; Nastase et al., 2020). The neural correlates of reading have traditionally been studied with serial word-by-word presentation with a fixed presentation time, which eliminates important aspects of the natural reading process and precludes direct comparisons between neural activity and

oculomotor behavior (Dimigen et al., 2011; Kliegl et al., 2012). The electrical neural correlates of normal reading of naturally occurring real-world sentences have been investigated less frequently due to a number of methodological challenges related to identifying the exact timing and type of visual stimuli presented during reading.

Because of recent methodological progress in stimulus presentation and data preprocessing (Dimigen et al., 2011; Ehinger and Dimigen, 2019), an excellent temporal resolution, and low costs, co-registered EEG, and eye-tracking have become important tools for studying the temporal dynamics of naturalistic reading (Frey et al., 2018; Hollenstein et al., 2018). Fixation-related potentials (FRPs), the evoked electrical responses time-locked to the onset of fixations, have become important tools for researchers to study various topics including free-viewing visual perception (e.g., Rämä and Baccino, 2010), brain-computer interfaces (e.g., Finke et al., 2016), and natural reading (e.g., Degno et al., 2019). In naturalistic reading paradigms, FRPs allow the study of the neural dynamics of how new information from a currently fixated word affects the ongoing language comprehension process.

In this work, we leverage these novel methodological advances to offer a machine learning (ML) benchmark challenge, formulated as a cross-subject classification task, to identify two reading tasks as accurately as possible. Specifically, the challenge is to discriminate between normal reading (with the only task of reading comprehension) and task-specific reading (TSR; with the purpose of finding specific information in the text) from eye-tracking and EEG data. Decoding mental states and detecting specific cognitive processes occurring in the brain during different reading tasks (i.e., *reading task classification*) are important challenges in cognitive neuroscience as well as in natural language processing (NLP). Applications of reading task classification include measuring attention and engagement (Miller, 2015; Abdelrahman et al., 2019), detecting proper reading vs. skimming (Biedert et al., 2012), as well as applications related to intent recognition within brain computer interfaces (Schalk et al., 2008). Other studies have demonstrated that recognizing reading patterns for estimating reading effort can improve the diagnosis of reading impairments such as dyslexia (Rello and Ballesteros, 2015; Raatikainen et al., 2021) and attention deficit disorder (Tor et al., 2021). Furthermore, it has been shown that using EEG and eye-tracking signals facilitates the prediction workload (Lobo et al., 2016) and investigation of language learning (Notaro and Diamond, 2018).

The accurate distinction of the cognitive processes occurring in different reading tasks is also important for ML and NLP. Identifying specific reading patterns can improve models of human reading and provide insights into human language understanding and how we perform linguistic tasks. This knowledge can then be applied to ML algorithms for NLP (e.g., information extraction applications). Computational models of language understanding can be adapted based on the insights from different reading and language processing tasks. Therefore, the identification of reading intents can be beneficial for computational methods of language understanding, but also for applications such as digital assistant tools, e.g., supporting translation processes, understanding how learners approach tasks in adaptive e-learning, and inferring document relevance.

A crucial potential of human physiological data in the context of NLP is that it can be leveraged to understand and to improve the manual labeling process required for generating training samples for supervised ML. For instance, Tokunaga et al. (2017) analyze eye-tracking data during the annotation of text to find effective gaze features for a specific NLP task and Tomanek et al. (2010) build cost models for active learning scenarios based on insights from eye-tracking data.

Reading task classification can help to improve the labeling processes by detecting tiredness from brain activity data and eye-tracking data, and subsequently to suggest breaks or task switching, or by using cognitive data directly to (pre-)annotate samples used for training ML models. If we can find and extract the relevant aspects of text understanding and annotation directly from the source, i.e., eye-tracking and brain activity signals during reading, we can potentially replace this expensive manual labeling work with ML models trained on physiological activity data recorded from humans while reading. Therefore, successful reading task classification could support the reduction of manual labor, improving label quality in ML systems as well as the job quality of annotators.

Essential for using neurophysiological signals to advance NLP is the availability of a large dataset providing concurrent measures of eye-tracking and EEG data, as well as ground truth labels for ML tasks. For the present benchmark, this is possible by leveraging a naturalistic dataset of reading English sentences, the Zurich Cognitive Language Processing Corpus (Hollenstein et al., 2018, 2020). The ZuCo dataset is publicly available and has recently been used in a variety of applications including leveraging EEG and eye-tracking data to improve NLP tasks (Barrett et al., 2018; Mathias et al., 2020; McGuire and Tomuro, 2021), evaluating the cognitive plausibility of computational language models (Hollenstein et al., 2019b; Hollenstein and Beinborn, 2021), investigating the neural dynamics of reading (Pfeiffer et al., 2020), developing models of human reading (Bautista and Naval, 2020; Bestgen, 2021).

Recently, ZuCo has also been leveraged for an ML competition on eye-tracking prediction (Hollenstein et al., 2021a).This competition revolves around a different task with a focus on computational language models in the field of natural language processing. The goal was to predict word-level eye-tracking features from normal reading such as mean fixation duration and fixation probability in a regression task. This shows that the ZuCo dataset has been used successfully for a wide range of ML tasks.

Moreover, the results of previous single-subject models for reading task classification (Hollenstein et al., 2021c;

Mathur et al., 2021) emphasize the potential of this task, but also highlight the performance gap between research-oriented single-subject models and more realistic cross-subject scenarios. The proposed benchmark therefore addresses this gap by focusing on the latter to improve the inter-subject generalization capabilities of these machine learning models. The recording of a new hidden testset with additional participants enables us to test this task in a suitable manner. Furthermore, by applying state-of-the-art EEG recording and preprocessing techniques, we ensure that this benchmark relies on a strong foundation, so that the resources and efforts of the research community can be spent wisely.

To conclude, the contributions of our work can be summarized as follows: First, we formulate a benchmark task for applying ML techniques to an important problem in cognitive science, namely, the classification of cognitive tasks. Second, we provide the data[1] and code[2] to reproduce our experiments. We provide a public benchmark and leaderboard on a new held-out test data. All information can be found here: www.zuco-benchmark.com. Finally, we propose and discuss models using various feature sets as baseline models for this benchmark task. We present detailed analyses of the results for both eye-tracking and EEG features and discuss the model performances.

2. Methods

The basis for this ML benchmark task is the Zurich Cognitive Language Processing Corpus 2.0 (ZuCo 2.0). ZuCo 2.0 was originally published in Hollenstein et al. (2020). In short, this corpus contains gaze and brain activity data of 18 participants reading 739 English sentences, 349 in a normal reading paradigm, and 390 in a task-specific paradigm, in which the participants actively search for a semantic relation type in the given sentence as a linguistic annotation task. This new dataset provides experiments designed to analyze the differences in cognitive processing between normal reading and task-specific reading.

In previous work, we recorded a first dataset (i.e., ZuCo 1.0) of simultaneous eye-tracking and EEG during natural reading (Hollenstein et al., 2018). ZuCo 1.0[3] consists of three reading tasks, two of which contain very similar reading material and experiments as presented in the current work. However, for ZuCo 1.0 the normal reading and task-specific reading paradigms were recorded in different sessions on different days. Therefore, the recorded data from ZuCo 1.0 is not appropriate as a means of comparison between normal reading and task-specific reading, since the differences in the brain activity data might result mostly from the different sessions due to the sensitivity of EEG. Therefore, while the data is available in the same format, it is not recommended to be used for this benchmark task. In the following section, we describe the compilation of the ZuCo 2.0 dataset.

TABLE 1 Descriptive statistics of reading materials (SD, standard deviation), including Flesch readibility scores.

	NR	TSR
Sentences	349	390
Sent. length	Mean (SD), range	Mean (SD), range
	19.6 (8.8), 5–53	21.3 (9.5), 5–53
Total words	6,828	8,310
Word types	2,412	2,437
Word length	Mean (SD), range	Mean (SD), range
	4.9 (2.7), 1–29	4.9 (2.7), 1–21
Flesch score	55.38	50.76

2.1. Reading materials

During the recording session, the participants read a total of 739 sentences that were selected from the Wikipedia corpus provided by (Culotta et al., 2006). This corpus was chosen because it provides annotations of semantic relations. Relation detection is a high-level semantic language understanding task requiring complex cognitive processing. ZuCo 2.0 includes seven of the originally defined relation types: *political_affiliation*, *education*, *founder*, *wife/husband*, *job_title*, *nationality*, and *employer*. The sentences were chosen with similar sentence lengths and Flesch reading ease scores (Flesch, 1948) between the two reading tasks. The Flesch score indicates how difficult an English text passage is to understand based on its structural characteristics, i.e., number of words and number of syllables. A higher Flesch score means the text is easier to read. The dataset statistics are shown in Table 1.

Of the 739 sentences, the participants read 349 sentences in a normal reading paradigm and 390 sentences in a task-specific reading paradigm, in which they had to determine whether a certain relation type occurred in the sentence or not. Table 2 shows the distribution of the different relation types in the sentences of the task-specific annotation paradigm. Purposefully, there are 63 duplicates between the normal reading and the task-specific sentences (8% of all sentences). The intention of these duplicate sentences is to provide a set of sentences read twice by all participants with a different task in mind. Hence, this enables the comparison of eye-tracking and

1 Benchmark data available here: https://osf.io/d7frw/.

2 Code for baseline methods available here: https://github.com/norahollenstein/zuco-benchmark.

3 Data available here: https://osf.io/q3zws/.

brain activity data when reading normally and when annotating specific relations. During both tasks, the participants were able to read in their own speed, using a control pad to move to the next sentence and to answer the control questions, which allowed for natural reading. Since all subject read at their own personal pace, the reading speed varies between subjects. Figure 1 shows the average sentence length, reading speed, and omission rate for each task.The sentence length (i.e., the number of words per sentence) was controlled in the selection of reading materials, so that it would not differ significantly between the two tasks (NR mean = 19.6, SD = 8.8; TSR mean = 21.3, SD = 9.5; $p = 0.02$ in a two-sided t-test).

2.1.1. Normal reading

In the first task, participants were instructed to read the sentences naturally, without any specific task other than comprehension. An example sentence is "He served in the United Stated Army in World War II, then got a law degree from Tulane University." The control condition for this task consisted of single-choice questions about the content of the previous sentence. Twelve percent of randomly selected sentences were followed by a comprehension question with three answer options on a new screen, for example, "Which university did he get his degree from? (1) Austin University, (2) Tulane University, (3) Louisiana State University."

TABLE 2 Distribution of relation types in the task-specific reading.

Relation type	Sentences
Political affiliation	45 (9)
Education	72 (10)
Wife	54 (12)
Job title	65 (11)
Employer	54 (10)
Nationality	60 (8)
Founder	40 (8)
Total	**390 (68)**

The right column contains the number of sentences, and the number control sentences without a relation in brackets.

2.1.2. Task-specific reading

In the second task, the participants were instructed to search for a specific semantic relation in each sentence they read. Instead of comprehension questions, the participants had to decide for each sentence whether it contains the relation or not, i.e., they were actively annotating each sentence. An example sentence containing the relation *founder* is "After this initial success, Ford left Edison Illuminating and, with other investors, formed the Detroit Automobile Company." Seventeen percent of the sentences did not include the particular relation type and were used as control conditions. All sentences within one recording block involved the same relation type. Each block was preceded by a short practice round, which described the relation type and was followed by three sample sentences, so that the participants would be familiar with the respective relation type.

2.2. Linguistic assessment

As a linguistic assessment, the vocabulary and language proficiency of the participants was tested with the LexTALE test (Lexical Test for Advanced Learners of English, Lemhöfer and Broersma, 2012). This is an unspeeded lexical decision task designed for intermediate to highly proficient language users. The average LexTALE score over all participants was 88.54%. Moreover, we also report the scores the participants achieved with their answers to the reading comprehension control questions and their relation annotations. The detailed scores for all participants are also presented in Table 3.

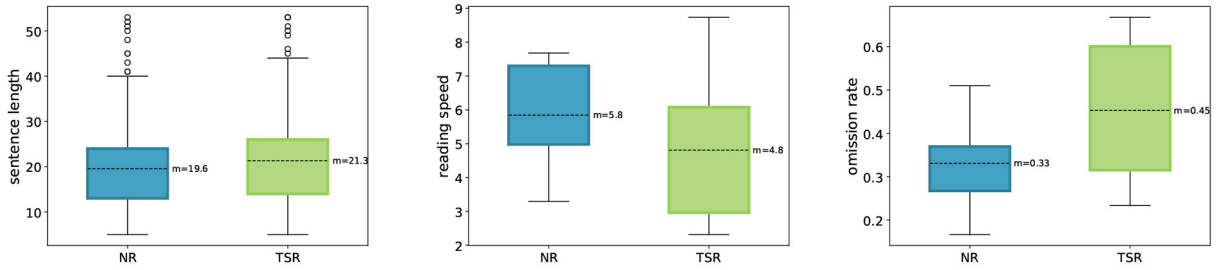


FIGURE 1
Sentence length (words per sentence), reading speed (seconds per sentence) and omission rate (percentage of words not fixated) comparison between normal reading (NR) and task-specific reading (TSR) of the sentence in ZuCo 2.0.

TABLE 3 Subject demographics for ZuCo 2.0, LexTALE scores, scores of the comprehension questions, and individual reading speed (i.e., seconds per sentence) for each task.

ID	Age	Gender	LexTALE	Comp. scores		Reading speed	
				NR	TSR	NR	TSR
YAC	32	female	76.25%	82.61%	83.85%	5.27	4.96
YAG	47	female	93.75%	91.30%	56.92%	7.64	8.73
YAK	31	female	100.00%	74.07%	96.41%	3.83	5.89
YDG	51	male	100.00%	91.30%	96.67%	4.97	3.93
YDR	25	male	85.00%	78.26%	96.92%	4.32	2.32
YFR	27	male	85.00%	89.13%	94.36%	6.48	4.79
YFS	39	male	90.00%	91.30%	96.15%	3.96	2.85
YHS	31	male	90.00%	78.26%	97.69%	3.30	2.40
YIS	52	male	97.50%	89.13%	98.46%	5.82	2.58
YLS	34	female	93.75%	91.30%	92.31%	5.57	5.85
YMD	31	female	100.00%	86.96%	95.64%	7.50	6.24
YRK	29	female	85.00%	97.83%	96.15%	7.35	7.70
YRP	23	female	82.50%	78.26%	90.00%	7.14	8.37
YSD	34	male	95.00%	93.48%	94.36%	5.01	2.87
YSL	32	female	71.25%	84.78%	83.85%	6.73	6.14
YTL*	36	male	81.25%	80.43%	94.10%	7.48	3.23
Mean	**34**	**44% m.**	**88.54%**	**86.36%**	**91.94%**	**5.84**	**4.81**

The * next to the subject ID marks a bilingual subject.

2.3. Participants

The subjects from ZuCo 2.0 are provided as training data for the current benchmark. For the ZuCo 2.0, we recorded data from 19 participants and discarded the data of one of them due to technical difficulties with the eye-tracking calibration. Another two subjects were discarded during data cleaning and preprocessing. Thus, we share the data of these 16 participants. All participants are healthy adults (between 23 and 52 years old; 10 females). Details on subject demographics can be found in Table 3. Their native language is English, originating from Australia, Canada, UK, USA or South Africa. Two participants are left-handed and three participants wear glasses for reading. All participants gave written consent for their participation and the re-use of the data prior to the start of the experiments. The study was conducted under approval by the Ethics Commission of the University of Zurich.

2.3.1. ZuCo 2.0 held-out testset

To provide a true hidden dataset for the current benchmark, we recorded data from 10 additional participants (i.e., a held-out testset). They underwent the identical procedure as in the ZuCo 2.0 dataset. All participants are healthy adults [mean age = 31.8 (SD = 5.11), four females]. All participants are right-handed. Their native language is English, originating from UK, Canada or USA. For an overview on subjects demographics, comprehension scores and reading speed please refer to Table 4. All participants gave written consent for their participation and the re-use of the data prior to the start of the experiments.

2.4. Procedure

Data acquisition took place in a sound-attenuated and dark experiment room. Participants were seated at a distance of 68 cm from a 24-inch monitor (ASUS ROG, Swift PG248Q, display dimensions 531 × 299 mm, resolution 800 × 600 pixels resulting in a display: 400 × 298.9 mm, a vertical refresh rate of 100 Hz). All sentences were presented at the same position on the screen and could span multiple lines. The sentences were presented in black on a light gray background with font size 20-point Arial, resulting in a letter height of 0.8 mm. The experiment was programmed in MATLAB 2016b (MathWorks, Inc. 2000), using PsychToolbox (Brainard, 1997). A stable head position was ensured *via* a chin rest. Participants were instructed to stay as still as possible during the recordings to avoid motor EEG artifacts. Participants completed the tasks sitting alone in

TABLE 4 Subject demographics for the new held-out test dataset, LexTALE scores, scores of the comprehension questions, and individual reading speed (i.e.,seconds per sentence) for each task.

ID	Age	Gender	LexTALE	Comp. scores		Reading speed	
				NR	TSR	NR	TSR
XAH	25	female	95.25%	91.30%	93.58%	5.58	3.94
XBB	37	male	95.75%	82.60%	93.84%	6.88	5.67
XBD	32	male	89.00%	91.30%	96.15%	7.31	4.48
XDT	25	male	97.50%	86.95%	93.85%	8.24	8.54
XLS	28	male	85.00%	89.13%	94.87%	7.52	5.68
XPB	29	male	97.50%	86.95%	91.02%	7.87	6.53
XSE	31	female	90.00%	89.13%	96.15%	7.23	3.75
XSS	42	female	97.50%	89.13%	96.67%	7.49	6.21
XTR	34	female	93.75%	89.13%	96.15%	9.18	5.91
XWS	35	male	100.00%	89.13%	95.64%	6.65	4.29
Mean	**31.8**	**60% m.**	**94.13%**	**88.48%**	**94.79%**	**7.40**	**5.50**

the room, while two research assistants were monitoring their progress in the adjoining room. All recording scripts including detailed participant instructions are available alongside the data. During both tasks, the participants were able to read in their own speed, using a control pad to move to the next sentence and to answer the control questions, which allowed for natural reading. All 739 sentences were recorded in a single session for each participant. The duration of the recording sessions was between 100 and 180 min, depending on the time required to set up and calibrate the devices, and the personal reading speed of the participants. Participants were also offered snacks and water during the breaks and were encouraged to rest. We recorded 14 blocks of ∼50 sentences, alternating between tasks: 50 sentences of normal reading, followed by 50 sentences of task-specific reading. The order of blocks and sentences within blocks was identical for all subjects. Each sentence block was preceded by a practice round of three sentences and followed by a short break to ensure a clear separation between the reading tasks. For the held-out test dataset, all blocks were merged and the order of the sentences was shuffled before sharing the data on OSF. This is done to prohibit the possibility that challenge participants would simply train a model to identify an experimental block rather than the type of reading for each sentence.

2.5. Data acquisition

2.5.1. Eye-tracking acquisition

Eye movements and pupil size were recorded with an infrared video-based eye tracker (EyeLink 1000 Plus, SR Research) at a sampling rate of 500 Hz and an instrumental spatial resolution of 0.01°. The eye tracker was calibrated with a nine-point grid at the beginning of the session and re-validated before each block of sentences. Participants were instructed to keep their gaze on a given point until it disappeared. If the average error of all points (calibration vs. validation) was below 1° of visual angle, the positions were accepted. Otherwise, the calibration was redone until this criterion was reached.

2.5.2. EEG acquisition

We recorded the high-density EEG data at a sampling rate of 500 Hz with a bandpass of 0.1–100 Hz, using a 128-channel EEG Geodesic Hydrocel system (Electrical Geodesics). The *Cz* electrode served as a recording reference. The impedance of each electrode was checked before recording and was kept below 40 $k\Omega$. Additionally, electrode impedance levels were checked after every third block of 50 sentences (approximately every 30 min) and reduced if necessary.

2.6. Data preprocessing and feature extraction

2.6.1. Eye-tracking preprocessing and feature extraction

2.6.1.1. Eye-tracking preprocessing

The eye tracker computed eye position data and identified events such as saccades, fixations, and blinks. Saccade onsets were detected using the eye-tracking software default settings: acceleration larger than 8,000°/s2, a velocity above 30°/s, and a deflection above 0.1°. The eye-tracking data consists of (x, y) gaze location entries for each individual time point (Figures 3A, B). Coordinates were given in pixels with respect to

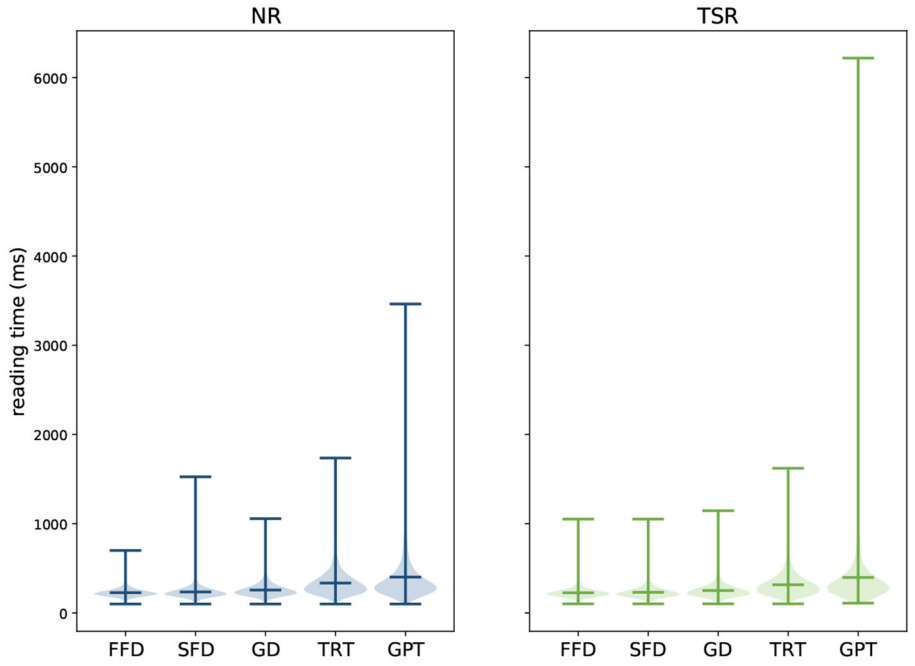


FIGURE 2
Violin plots showing means, distributions, and ranges of the reading time measures per word for each task and each eye-tracking feature (*x*-axis) in milliseconds.

the monitor coordinates [the upper left corner of the screen was (0, 0) and down/right was positive]. We provide this raw data as well as various engineered eye-tracking features.

2.6.1.2. Eye-tracking feature extraction

For this feature extraction, only fixations within the boundaries of each displayed word were extracted. A Gaussian mixture model was trained on the (*y*-axis) gaze data for each sentence to improve the allocation of eye fixations to the corresponding text lines. The number of text lines determined the number of Gaussians to be fitted within the model. Subsequently, each gaze data point was clustered to the matching Gaussian and the data were realigned. As a result, each gaze data point is clearly assigned to a specific text line. Data points distinctly not associated with reading (minimum distance of 50 pixels to the text) were excluded. Additionally, fixations shorter than 100 ms were excluded from the analyses, because these are unlikely to reflect fixations relevant for reading (Sereno and Rayner, 2003). On the basis of previous eye-tracking corpora, namely the GECO corpus (Cop et al., 2017) and ZuCo 1.0 (Hollenstein et al., 2018), we extracted the following features: (i) *gaze duration* (GD), the sum of all fixations on the current word in the first-pass reading before the eye moves out of the word; (ii) *total reading time* (TRT), the sum of all fixation durations on the current word, including regressions; (iii) *first fixation duration* (FFD), the duration of the first fixation on the prevailing word; (iv) *single fixation duration* (SFD), the duration of the first and only fixation on the current word; and (v) *go-past time* (GPT), the sum of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word. See Figure 2 for a visualization of the feature ranges of each reading task. For each of these eye-tracking features, we additionally computed the pupil size. Furthermore, we extracted the number of fixations and mean pupil size for each word and sentence. Additionally, on the sentence level, we extracted the mean and maximum saccade velocity, saccade amplitude and saccade duration. On the word level, saccade velocity, amplitude, and duration were extracted for in-going, outgoing, as well as saccades within a word. Finally, on the sentence level, omission rate is calculated, representing the proportion of words which were not fixated within each sentence.

2.6.2. EEG preprocessing and feature extraction

2.6.2.1. EEG preprocessing

Before the EEG preprocessing, data from all 14 blocks (seven NR and seven TSR) were first merged to avoid high predictive power based on the differences resulting from the preprocessing itself. To avoid loss of data by the subsequent automated preprocessing pipeline, the files of each recording blocked were screened to exclude highly artifactual data. Therefore, each block was temporarily filtered using a 2 Hz high-pass filter. Subsequently, outlying data points were removed

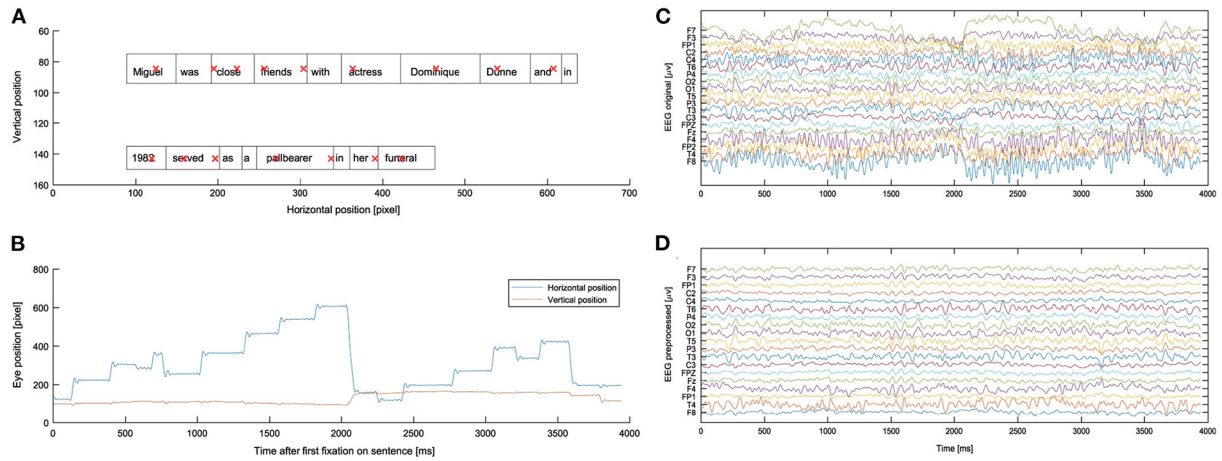


FIGURE 3
Visualization of eye-tracking and EEG data for a single sentence. **(A)** Prototypical sentence fixation data. Red crosses indicate fixations; boxes around the words indicate the wordbounds. **(B)** Fixation data plotted over time. **(C)** Raw EEG data during a single sentence. **(D)** Same data as in **(C)** after preprocessing.

if they exceeded a threshold of three standard deviations above or below the mean of the data. Only if the standard deviation of this temporarily pre-cleaned data was below a cut-off of 100 μV, the original corresponding block was used in the merging process. Applying this criterion, 4.02% of all blocks were excluded. The EEG preprocessing was conducted with the open-source MATLAB toolbox preprocessing pipeline Automagic (Pedroni et al., 2019), which combines state-of-the-art EEG preprocessing tools into a standardized and automated pipeline. The EEG preprocessing consisted of the following steps: First, bad channels were detected by the algorithms implemented in the EEGlab plugin `clean_rawdata`.[4] A channel was defined as a bad electrode when recorded data from that electrode was correlated at <0.85 to an estimate based on other channels. Furthermore, a channel was defined as bad if it had more line noise relative to its signal than all other channels (four standard deviations). Finally, if a channel had a longer flat-line than 5 s, it was considered bad. These bad channels were automatically removed and later interpolated using a spherical spline interpolation (EEGLAB function `eeg_interp.m`). The interpolation was performed as a final step before the automatic quality assessment of the EEG files. Next, data were filtered using a 2 Hz high-pass filter and line noise artifacts were removed by applying Zapline (de Cheveigné, 2020), removing seven power line components. Subsequently, independent component analysis (ICA) was performed. Components reflecting artifactual activity were classified by the pre-trained classifier ICLabel (Pion-Tonachini et al., 2019). Components that were classified as any class of artifacts (line noise, channel noise, muscle activity, eye activity, and cardiac artifacts) with a probability higher than 0.8 were removed from the data. Subsequently, residual bad channels were excluded if their standard deviation exceeded a threshold of 25 μV. Very high transient artifacts (>100 μV) were excluded from calculating the standard deviation of each channel. However, if this resulted in a significant loss of channel data (>50%), the channel was removed from the data. After this, the pipeline automatically assessed the quality of the resulting EEG files based on four criteria: First, a data file was marked as bad-quality EEG and not included in the analysis if the proportion of high-amplitude data points in the signals (>30 μV) was larger than 0.20. Second, more than 20% of time points showed a variance larger than $15\mu V$ across channels. Third, 30% of the channels showed high variance (>15 μV). Fourth, the ratio of bad channels was higher than 0.3. After Automagic preprocessing, 13 electrodes in the outermost circumference (chin and neck) were excluded from further processing as they capture little brain activity and mainly record muscular activity. The discarded electrode labels were E1, E8, E14, E17, E21, E25, E32, E48, E49, E56, E63, E68, E73, E81, E88, E94, E99, E107, E113, E119, E125, E126, E127, and E128. Additionally, 10 EOG electrodes were separated from the data and not used for further analysis, yielding a total number of 105 EEG electrodes. Subsequently, the data was converted to a common average reference.

2.6.2.2. EEG and eye-tracking synchronization

In a next step, the EEG and eye-tracking data were synchronized using the "EYE-EEG" toolbox (Dimigen et al., 2011) to enable EEG analyses time-locked to the onsets of fixations and saccades, and subsequently segment the EEG data based on the eye-tracking measures. The synchronization

4 http://sccn.ucsd.edu/wiki/Plugin_list_process

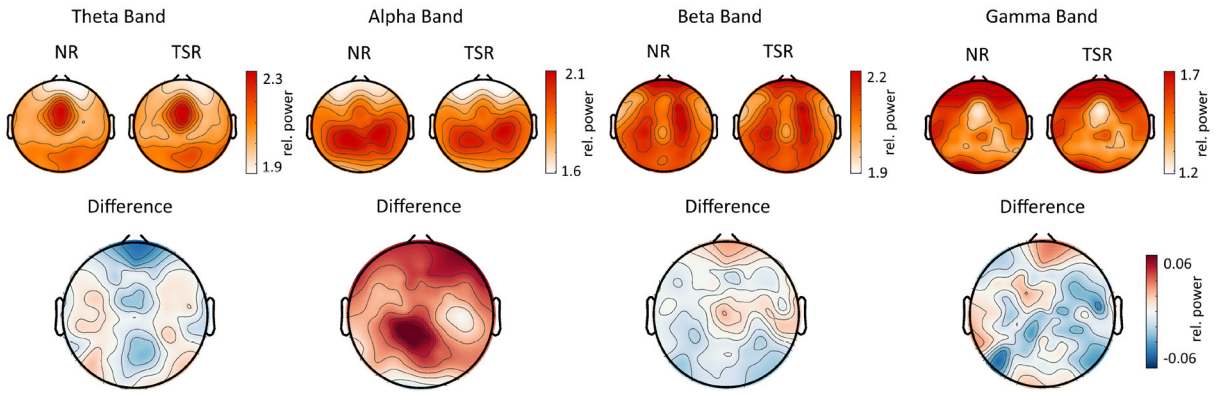


FIGURE 4
Topographical plots showing the mean EEG activity across all subjects from ZuCo 2.0. Averaged sentence level features are plotted in each reading condition as well as the difference between the tasks (NR minus TSR; scalp viewed from above, nose at the top). Only for the purpose of this visualization, relative power values are plotted (i.e., power in each frequency band divided by the average power between 1 and 50Hz), showing the expected typical power distribution across the scalp.

algorithm first identified the "shared" events. Next, a linear function was fitted to the shared event latencies to refine the start- and end-event latency estimation in the eye tracker recording. Finally, the synchronization quality was ensured by comparing the trigger latencies recorded in the EEG and eye-tracker data. All synchronization errors did not exceed 2 ms (i.e., one data point). Remaining eye artifacts in data were removed with Unfold toolbox (Ehinger and Dimigen, 2019) according to a method described in Pfeiffer et al. (2020). The effect of this preprocessing can be seen from Figures 3C, D.

2.6.2.3. EEG feature extraction

To compute oscillatory power measures, we band-pass filtered the continuous EEG signals across an entire reading task for four different frequency bands, resulting in a time-series for each frequency band. The distinct frequency bands were determined as follows: *theta_1* (4–6 Hz), *theta_2* (6.5–8 Hz), *alpha_1* (8.5–10 Hz), *alpha_2* (10.5–13 Hz), *beta_1* (13.5–18 Hz), *beta_2* (18.5–30 Hz), *gamma_1* (30.5–40 Hz), and *gamma_2* (40.5–49.5 Hz). Afterwards, we applied a Hilbert transformation to each of these time-series resulting in a complex time series. The Hilbert phase and amplitude estimation method yields results equivalent to sliding window Fourier transformation and wavelet approaches (Bruns, 2004). We chose specifically the Hilbert transformation to maintain temporal information for the amplitude of the frequency bands to enable the power computation of the different frequencies for time segments defined through fixations in the eye-tracking data. Finally, for each sentence as well as for each word within each sentence, and for each frequency band, the EEG features consist of a vector of 105 dimensions (one value for each EEG channel). On the level of individual words, these frequency band power features were calculated based on fixations of GD, TRT, FFD, SFD, and GPT (see above). For each EEG feature, all channels were subject to an artifact rejection criterion of 90 μV to exclude trials with transient noise. To descriptively compare the EEG activity and the extracted frequency band power between the NR and TSR sentences, the average of each condition as well as the differences (NR minus TSR) for the different sentence-level EEG features are plotted in Figure 4.

2.7. Data access

The raw and preprocessed EEG and eye-tracking data, as well as the features extracted from the preprocessed EEG and eye-tracking are provided for this benchmark. For the training data, the information about the task (normal reading or task-specific reading) is also available. Please note that for the held-out test dataset, we can only provide the preprocessed data and the extracted features. As the raw data were collected in different blocks of normal reading and task-specific reading, the participants could otherwise infer the outcome from the block separation. All the data can be accessed *via* OSF: https://osf.io/d7frw/.

3. Benchmark task

3.1. Task definition

We propose an ML benchmark for reading task identification. As described in Section 2, the ZuCo corpus provides data from two reading paradigms, normal reading (NR) and task-specific annotation reading (TSR). Consequently, we frame the problem as binary classification task with labels $Y \in \{\text{NR}, \text{TSR}\}$. The training data consists of sentences labeled

depending on which reading task they belonged to during the experiment. Each sentence is represented by a feature set X. The input features should be eye-tracking or EEG features, or a combination thereof.
The goal of the benchmark task is to build a binary classifier h to predict the label Y for each sentence given only the features X:

$$h : X \rightarrow \{\text{NR}, \text{TSR}\}. \quad (1)$$

Due to the naturalistic experiment design and the co-registration of EEG and eye movement signals, feature extraction is possible on various levels. There are no restrictions to the type and dimension of the input features or the model.

3.2. Performance metrics

The classifier's performance is evaluated by the classification accuracy, defined as the number of correct predictions divided by the total number of predictions. Since previous results have shown high performance on models trained and tested within-subject but low performance on cross-subject models (Hollenstein et al., 2021c), this benchmark aims to address this gap by focusing on the latter to improve the inter-subject generalization capabilities of the models. We propose a cross-subject evaluation, where each subject in the held-out testset is evaluated by a model trained on all subjects in the training split (i.e., the original ZuCo 2.0 dataset). Therefore, the main benchmark metric is defined as the mean classification accuracy across all subjects in the testset. As a second metric, we choose the F1-measure. In our classification setup, we do not distinguish between a positive and a negative class, i.e., there is no clear majority or minority class. For that reason, we choose to evaluate our classifier using the macro-averaged F1-scores. The benchmark task is evaluated on models from the following three categories: models trained on EEG features, models trained on eye-tracking features, and models trained on a combination of EEG and eye-tracking features.

3.3. Benchmark setup

We host the ZuCo benchmark on Eval-AI (Yadav et al., 2019) – an open source AI challenge platform for evaluating and comparing machine learning and artificial intelligence algorithms. The link to the reading task classification challenge and more information on how to participate is available here: https://github.com/norahollenstein/zuco-benchmark. This solution will help other researchers to participate in our machine learning challenge and enable us to automate the evaluation of the future submissions.

3.3.1. Evaluation strategy

Researchers that want to participant in the benchmark task can submit predictions from their models for the hidden testset. We specified the challenge configuration, evaluation code, and information about the data splits. Predictions for the testset labels can be submitted in the JSON file.

3.3.2. Leaderboard

The public leaderboard will include the scores on the chosen evaluation metrics as well as references to upcoming publications. Upon submission, the predictions will be handed over to challenge-specific workers that compare the predictions against corresponding ground-truth labels using the custom evaluation script provided by our team.

4. Baseline methods

4.1. Textual baselines

We set three minimal baselines for this benchmark task: (i) a random baseline, (ii) a word embedding baseline, and (iii) a text difficulty baseline. We will use the first one as the basis for model comparison, while the latter two serve merely as control conditions to validate the dataset and exclude linguistic properties as a possible confound in the reading task classification benchmark.

4.1.1. Random baseline

We compute a random baseline to assess the chance level of predicting the correct class. We randomly sample the labels according to the distribution of the training data. That means the label NR is chosen with a probability of $p_{NR} = \frac{390}{739} \approx 0.53$ and TSR is chosen with $p_{TSR} = 1 - p_{NR} \approx 0.47$.

4.1.2. Word embedding baseline

Even though the experimental design of ZuCo ensured the similarilty of the sentences in terms of sentence lengths and text complexity, we aim to ensure the sentences in the data are not easily separable merely by their linguistic characteristics. Therefore, we compare our models to a textual baseline as a sanity check. For this purpose, we use pre-trained textual representations, namely, the state-of-the-art contextualized BERT word embeddings (Devlin et al., 2019). We concatenate the embeddings of all words in a sentence and feed them into the LSTM model.

4.1.3. Text difficulty baseline

We also provide a baseline based on text readability. Although the sentences for both reading tasks were chosen to

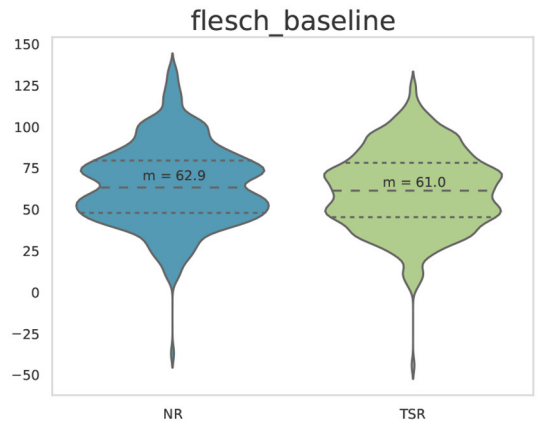


FIGURE 5
Flesch reading ease (FRE) scores for the NR and TSR sentences used in the ZuCo 2.0. dataset.

be of similar length and from the same text genre, we want to ensure that both tasks are not separable merely by the difficulty of the sentences. Therefore, we implement a text difficulty baseline, which classifies the sentences into NR and TSR based on their Flesch reading ease score (FRE; Flesch, 1948). This score indicates how difficult an English text passage is to understand based on the average number of words in a sentence and the average number of syllables in a word:

$$FRE = x - y\left(\frac{words}{sentences}\right) - z\left(\frac{syllables}{words}\right) \quad (2)$$

where x, y and z are language-specific weighting factors (for English $x = 206.835$, $y = 1.015$, $z = 84.6$). We compute FRE scores for each of the English sentences in the ZuCo data. Figure 5 shows the distribution of the FRE across the sentences of ZuCo 2.0.

4.2. EEG and eye-tracking models

We also present a set of initial models using EEG and eye-tracking features as a starting point for future models.[5] For each sentence in the dataset, the model input is composed of a vector of eye-tracking and/or EEG features corresponding to a single sentence in the dataset. Each sample in the training set is labeled with the reading task it was recorded in, normal reading (NR) or task-specific reading (TSR). We investigate the potential of using sentence-level eye-tracking and EEG features for the reading task classification. Hollenstein et al. (2021c) compared sentence-level and word-level features for this task previously and showed that sentence-level features perform better. However, challenge participants are also invited to use word-level and other features (see discussion in Section 6 for suggestions). The advantages of sentence-level features consist of the possibility of using simpler machine learning models and reduced training times (Hollenstein et al., 2021c). Sentence-level features are defined as metrics aggregated over all words in a given sentence.

4.2.1. Eye-tracking features

We include two types of sentence-level eye-tracking features. The features are summarized in Table 5. First, the fixation-based features - omission rate, number of fixations and reading speed - are aggregated metrics normalized by sentence length, i.e., the number of words in a sentence. Analogous to the word-level models, we also include saccade-based features. These include the mean and maximum duration, velocity and amplitude across all saccades that occurred within the reading time of a give sentence. We test these features individually and combined to investigate the performance increase achieved by adding more features.

4.2.2. EEG features

The sentence-level EEG features take into account the EEG activity over the whole sentence duration (even when no words were fixated). We aggregate over the preprocessed EEG signals of the full reading duration of a sentence. Each subfrequency band (e.g., *alpha_1* and *alpha_2*) were averaged to get one power measure for each frequency band, i.e., *theta* (4–8 Hz), *alpha* (8.5–13 Hz), *beta* (13.5–30 Hz), and *gamma* (30.5–49.5 Hz). The sentence-level EEG features are described in Table 6. We experiment with both aggregate metrics, i.e., the mean across all electrodes, and individual electrode features.

Examples of these features across all subjects, split by class (normal reading vs. task-specific reading) are shown in Figure 6 for ZuCo 2.0.

4.2.3. Principal component analysis

We use principal component analysis (PCA) to reduce the dimensionality of the EEG features. In an initial attempt, we fitted PCA on all training subjects and applied it to both the training and test split. This, however, led to no significant improvements in classification accuracy. Thus, we fit PCA to each subject individually. To prevent overfitting to the test subjects, we only consider subjects in the training data to determine the number of components. We fit PCA for each subject separately and calculate the number of components that explain 95% of the variance. We then choose the number of components of PCA as the median

5 The code is available here: https://github.com/norahollenstein/zuco-benchmark.

TABLE 5 Sentence-level eye-tracking features.

Name	Definition	Values
Fixation features		
omission_rate	Percentage of words in a sentence that is *not* fixated	1
fixation_number	Number of fixations in the sentence divided by the number of words	1
reading_speed	Sum of the duration of all fixations in the sentence divided by the number of words	1
Saccade features		
mean_sacc_dur	Sum of the duration of all saccades in the sentence divided by the number of words	1
max_sacc_dur	Maximum saccade duration per sentence	1
mean_sacc_velocity	Sum of the velocity of all saccades in the sentence divided by the number of saccades	1
max_sacc_velocity	Maximum saccade velocity per sentence	1
mean_sacc_amplitude	Sum of the amplitude of all saccades in the sentence divided by the number of saccades	1
max_sacc_amplitude	Maximum saccade amplitude per sentence	1
Combined features		
Combined ET features	Concatenation of all eye-tracking features	9

We use the combination of all features for our models.

TABLE 6 Sentence-level EEG features.

Name	Definition	Values
Mean features		
theta_mean	Mean theta band features averaged over all electrodes	1
alpha_mean	Mean alpha band features averaged over all electrodes	1
beta_mean	Mean beta band features averaged over all electrodes	1
gamma_mean	Mean gamma band features averaged over all electrodes	1
eeg_means	Mean frequency band features averaged over all electrodes, resulting in 1 feature value for each of the 8 frequency bands	8
Electrode features		
electrode_features_theta	Mean theta1 and theta2 values of all 105 electrodes	105
electrode_features_alpha	Mean alpha_1 and alpha_1 values of all 105 electrodes	105
electrode_features_beta	Mean beta_1 and beta_1 values of all 105 electrodes	105
electrode_features_gamma	Mean gamma_1 and gamma_1 values of all 105 electrodes	105
electrode_features_all	Concatenation of the four features above	420
Combined features		
ET & EEG mean features	Concatenation of sent_gaze_sacc and eeg_means	17

over all subjects in the training data, which makes it robust against outlier subjects. The result is a reduced dimensionality from 105 to 41 of both training and test data. Figure 7 shows that the amount of variance explained by the first components varies significantly between subjects. The first component, for instance, accounts for ∼24% of the variance for subject YTL, whereas it accounts for 49% of the variance for subject YAC.

To analyze how much the individual electrodes influence the principal components, we again fit PCA for each subject of the training data, such that the resulting components explain 95% of the variance. Assuming we have n original features and m principal components c, where each component is a linear combination of the original features, i.e., $c^j = \sum_i^n \beta_i^j x_i, j \in 1 \ldots m$. We then extract the amount of variance explained (w^j) by each component c^j and its weights β_i^j. We sum up all β_i^j

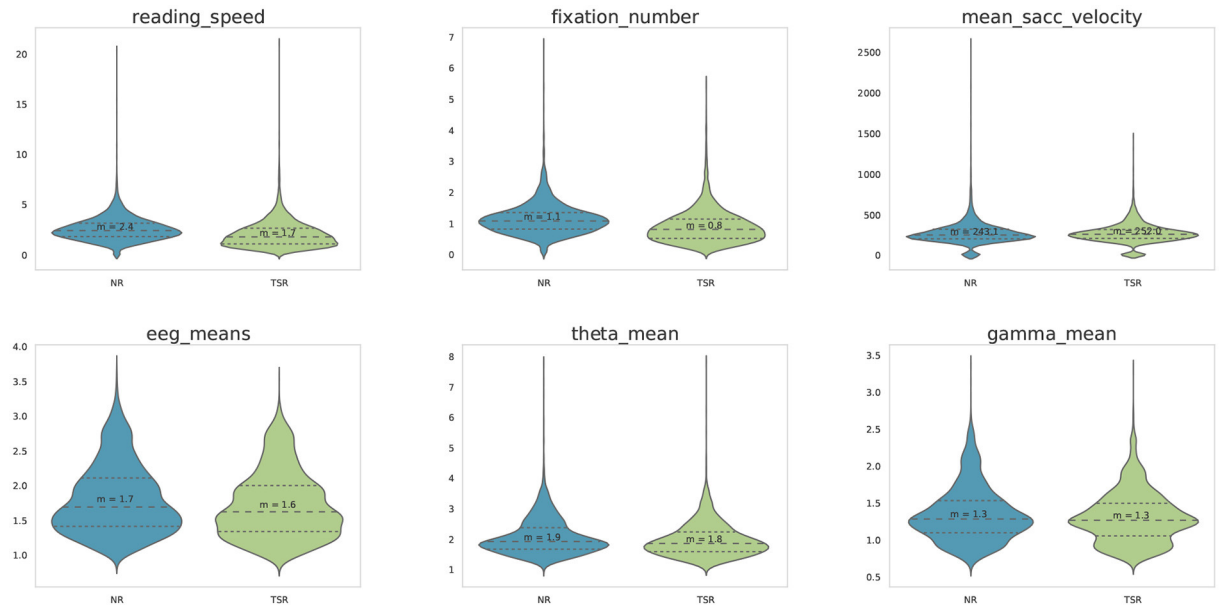


FIGURE 6
Examples of feature distributions across all subjects for the NR and TSR sentences included in the ZuCo 2.0. dataset.

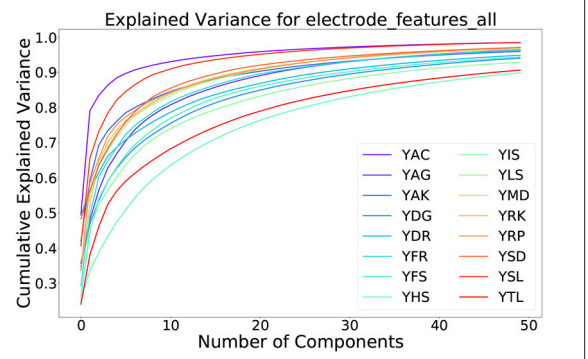


FIGURE 7
Variance explained with increasing number of PCA components for the training subjects in ZuCo 2.0.

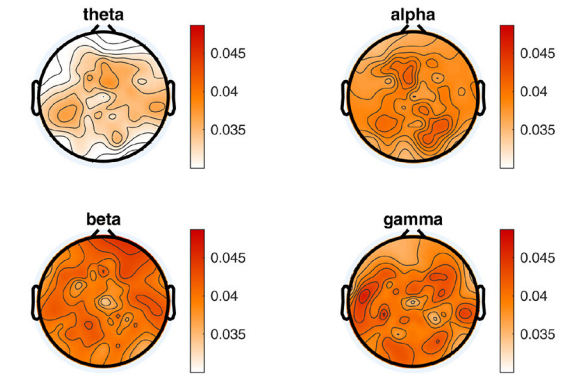


FIGURE 8
Topographical distribution of electrode importance for the principle components, divided into the 4 different frequency bands. Electrode importance is calculated by determining the influence of each electrode on the principle components and weighting them by amount of explained variance.

weighted by w^j, such that the resulting $\beta_i = \sum_j^m w^j \beta_i^j$ represents the relevance of feature x_i.

Following this procedure, we split the results into frequency bands and present the corresponding topography plots averaged over all training subjects in ZuCo 2.0 in Figure 8.

4.2.4. Model

The input to the sentence-level model is a single vector representing each sentence. We scale the feature values to a range between $\{0, 1\}$. We train a support vector machine for classification with a linear kernel. We use the `scikit-learn` SVC implementation.[6] For the cross-subject evaluation, the models are trained on all samples from all subjects in ZuCo 2.0 and tested on the samples from new subjects in the held-out testset.

6 https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

TABLE 7 The mean accuracy and F1-score over all subjects for each feature-set in the benchmark task.

Feature set	Accuracy	F1
Random	0.50	0.50
FRE baseline	0.53	0.35
BERT baseline	0.65	0.64
Eye-tracking features	0.69	0.67
Eye-tracking and EEG mean features	0.68	0.66
Concatenated EEG electrode features	0.55	0.46
Concatenated EEG electrode features (with PCA)	0.58	0.56

5. Results

5.1. Results of textual baselines

As described in the previous section, we set three minimal baselines for this benchmark task: (i) a random baseline, i.e., chance level for binary classification, (ii) a word embedding baseline, namely BERT word embeddings, and (iii) a text difficulty baseline, based on the Flesch reading ease score (FRE). The random baseline for binary classification is at 0.50 accuracy. The word embedding baseline yield a classification accuracy of 0.65 for ZuCo 2.0. The text difficulty baseline is also above random performance with a classification accuracy of 0.53 for ZuCo 2.0. Table 7 shows the accuracy and F1-score for all baselines.

5.2. Results of EEG and eye-tracking models

As described in Section 3, we consider three different feature sets, EEG, eye-tracking, and the combination of all features. For each feature set and each subject, we report the accuracy and the F1-score. For each subject in the hidden testset, we compute the results *via* bootstrapping, sampling 500 times with replacement, and using a sample size equal to the original data. For all results, we report the comparison to the random and textual baselines as well as the 95% confidence intervals for each subject. Table 7 shows a summary of the results. The corresponding tables with the detailed numbers for all subjects and feature sets are shown in Appendix 1.

First, the results for the eye-tracking features are shown in Figure 9. These results clearly show all subjects outperforming the random baseline and FRE control model except for one subject each for accuracy and F1-score. All subjects except one perform better than the random baseline, and three subjects perform significantly better than the BERT word embedding control model. The mean accuracy across all subjects in the testset is 0.69, and the mean F1-score is 0.67. Furthermore, the results for the combined eye-tracking and EEG mean feature set in Figure 10 do not yield an increase in performance compared to using only the eye-tracking features (mean accuracy: 0.68; F1-score: 0.66). Interestingly, the best and worst performing subjects vary between different feature combinations, and between accuracy and F1-score.

Next, we show the results using the concatenated EEG electrode features[7] in Figure 11. With this feature set, the mean accuracy across all subjects in the testset is 0.55, and the mean F1-score is 0.46. The accuracy scores are notably higher than for the F1-score. Finally, when using the same features but applying the PCA preprocessing, the models yield the results presented in Figure 12. The scores for the accuracy are similar but have a slightly higher mean of 0.58 (compared to 0.55 without PCA). However, the F1-scores with PCA are significantly higher with a mean of 0.56 (compared to 0.46 without PCA). While with these EEG electrode features the models outperform the random and text difficulty baseline for some test subjects, they do not achieve to outperform the strong embedding baseline. Additionally, we experimented with combining the BERT embeddings with the EEG and eye-tracking feature sets in the SVM models. However, the combination of linguistic and physiological features did not yield any improvements.

6. Discussion

The present benchmark challenge has the main goal of advancing reading task classification through eye-tracking and EEG data. The challenge participants are invited to develop ML models to identify whether subjects are reading a sentence with the goal of reading comprehension (i.e., normal reading) or whether the subjects are reading a sentence to search for a specific semantic relation in the sentence (i.e., task-specific reading). The objective is to investigate which eye movement and brain activity features are most suited to solve this problem. Understanding the physiological aspects of the reading process (i.e., the cognitive load and reading intent) can advance our understanding of human language processing and general attentional processes. On the other hand, natural language processing and machine learning would benefit, as classifiers that outperform current textual baselines could improve the quality and process of collecting annotated data (e.g., through gaze-aided unsupervised labeling).

Several previous studies have used ML models to accurately perform a reading task classification. Cole et al. (2011) used eye-tracking data to discriminate between a scanning task and a reading comprehension task. Furthermore, Biedert et al. (2012) developed a real-time classifier able to distinguish reading from

7 These figures show the results for absolute EEG power. The results for relative EEG power are depicted in the Appendix 1 in Figure 13.

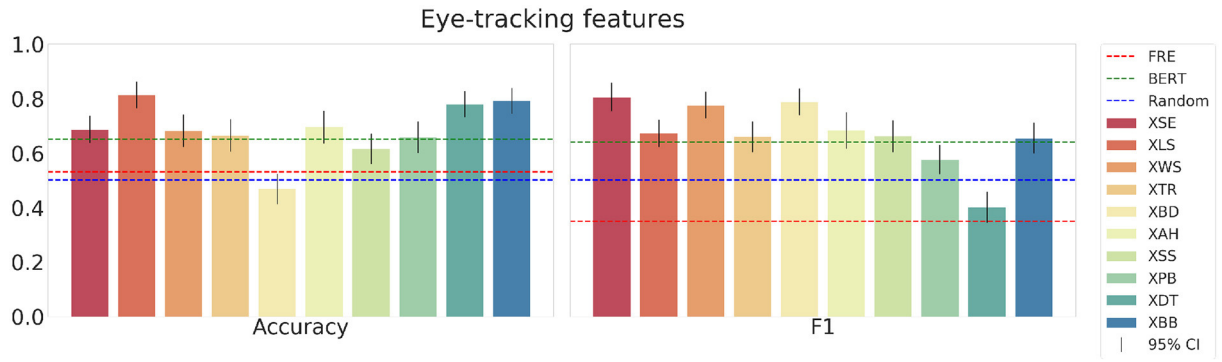


FIGURE 9
The mean accuracy **(left)**, mean F1-score **(right)** with corresponding 95% confidence intervals and textual baselines are plotted for each subject in the held-out test dataset using the concatenated eye-tracking features.

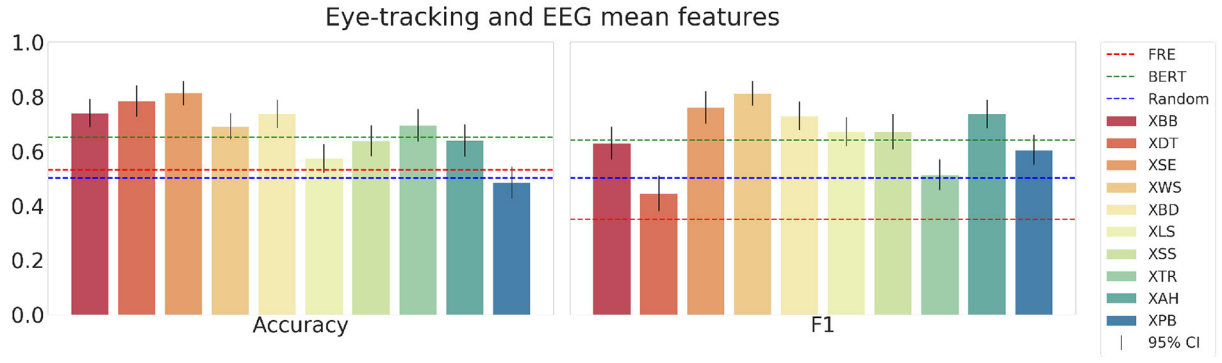


FIGURE 10
The mean accuracy **(left)**, mean F1-score **(right)** with corresponding 95% confidence intervals and textual baselines are plotted for each subject in the held-out test dataset using the eye-tracking and EEG mean features.

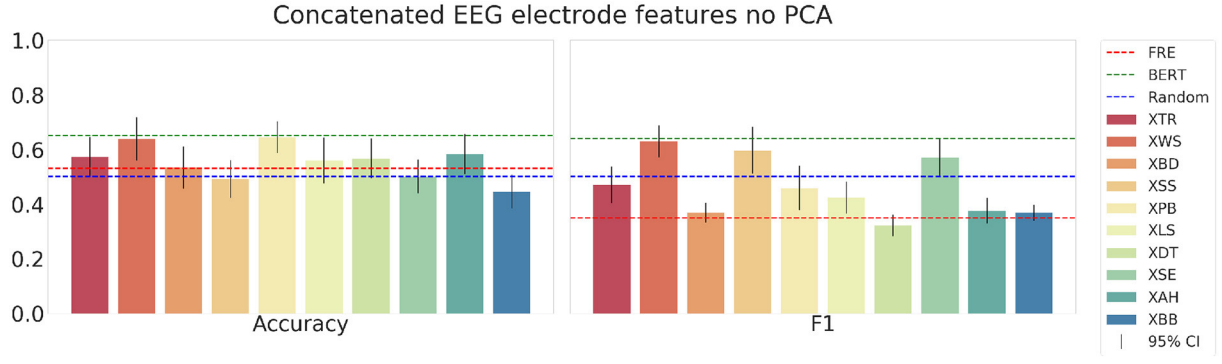


FIGURE 11
The mean accuracy **(left)**, mean F1-score **(right)** with corresponding 95% confidence intervals and textual baselines are plotted for each subject in the held-out test dataset using the concatenated EEG electrode features without PCA pre-processing.

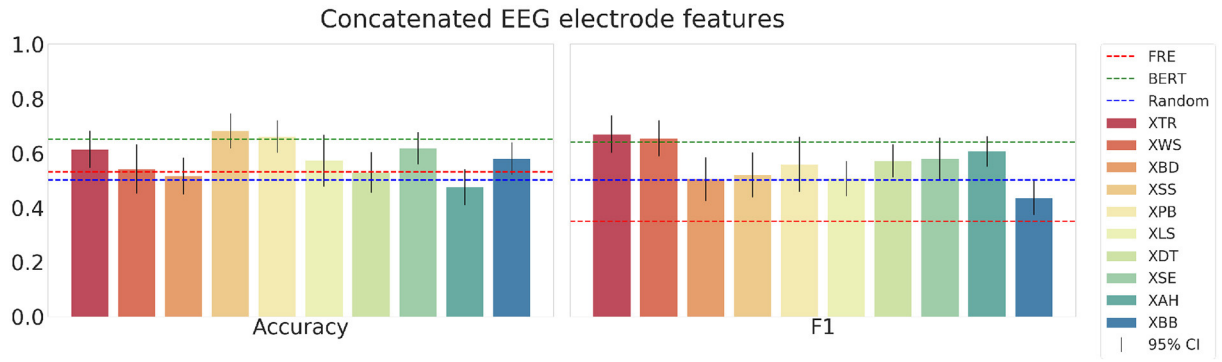


FIGURE 12
The mean accuracy **(left)** mean, F1-score **(right)** with corresponding 95% confidence intervals and textual baselines are plotted for each subject in the held-out test dataset using the concatenated EEG electrode features after pre-processing with PCA.

skimming patterns. In a related study, Kelton et al. (2019) investigated the influence of different content and tasks on the performance to determine whether subjects are reading or skimming a news article. Other neuroimaging methods such as fMRI have been combined with eye-tracking to examine the neural basis of sentence comprehension (e.g., Bonhage et al., 2015) or the discrimination between normal and non-word text (Choi et al., 2014). In another fMRI study, which simultaneously recorded eye-tracking data, Ceh et al. (2021) observed that internally and externally directed cognition are characterized by distinct brain activity. In addition, several research groups provide publicly available fMRI data to study naturalistic reading comprehension (Dehghani et al., 2017; Lopopolo et al., 2018; Pereira et al., 2018; Shain et al., 2020; Nastase et al., 2021). While functional MRI has a better spatial resolution compared to EEG, is a very costly method with restricted real-life usability. Whereas eye-tracking and EEG systems are of lower cost and can be used in more naturalistic situations. Several other publicly datasets recorded eye-tracking (e.g., Cop et al., 2017; Luke and Christianson, 2018; Jäger et al., 2021) or EEG from continuous speech stimuli (e.g., Broderick et al., 2018; Brennan and Hale, 2019). These datasets provide the possibility to improve and evaluate machine learning systems for NLP. However, to the best of our knowledge, the ZuCo dataset is the largest publicly available dataset that features simultaneous eye movement and EEG data recorded in a naturalistic reading setup. One recent addition is the CoCoNUt dataset by Frank and Aumeistere (2022), which contains 200 Dutch sentences with combined EEG and eye-tracking recordings. However, the selection of sentences is not completely natural, as it is guided by sentence length and word frequency. Thus, ZuCo is specifically tailored to leverage EEG and eye-tracking data to improve natural language processing tasks in a naturalistic setting. The field of machine learning contains a range of tasks on different modalities such as language (text), computer vision (video, images), and speech recognition (audio). Recently, Akbari et al. (2021) have shown superior performance of ML models with multimodal representations on downstream tasks such as image classification. Therefore, from an NLP perspective, another extension to this benchmark could be to investigate whether leveraging multimodal embeddings is beneficial for reading task classification.

In a recent study, the ZuCo data has been used already for reading task identification (Mathur et al., 2021) using a complex convolutional network, which is evaluated on a fixed cross-subject scenario on the sentences from ZuCo 2.0. However, the relatively poor performance of their model evaluated in a fixed cross-subject scenario, still leaves room for improvement and opens research questions regarding the selection of features. Hollenstein et al. (2021c) have recently presented extensive work on reading task classification, corroborating the advantages of the ZuCo dataset for this ML task. The authors found that, while high accuracy can be achieved on within-subject models, the performance drops for cross-subject evaluations. There is clearly room for improvement in the performance of the results presented in this work. However, these are still very promising results considering the complex nature of human physiological data.

A current bottleneck in machine learning is the lack of generalization capabilities of these models, meaning that the models perform poorly on data from other domains that are not included in their training data. For instance, ML models perform less accurately across languages, across image or text domains, or across subjects. The latter is of great importance in neuroscientific research which aims at a principled understanding of human brain activity as a response to complex stimuli (Nastase et al., 2019), as well as for practical applications such as brain-computer interfaces (Chiang et al., 2019). Specifically, when trained on physiological data, the rules identified by ML models for a given task ideally hold for the entire population. Considering the ever-increasing complexity of ML models due to their large number of parameters,

they are prone to overfit to their training set (which does not characterize the entire population), leading to spurious correlations. Therefore, to validate the gained insights on the physiological data, ML models need to be evaluated on held-out subjects as a proxy to the model's generalization capability. These results inspired the proposed benchmark based on the ZuCo dataset. The benchmark task and baseline models follow the rules suggested by Scheinost et al. (2019) to take into account subject-specific differences in predictive modeling.

In the current paper, we provide evidence that both eye-tracking and brain activity data can improve reading task classification compared to purely text-based baselines. The best-performing model is based on sentence-level eye-tracking features. Combining eye-tracking and EEG mean features yields promising results, but not better than only eye-tracking features. One explanation for this is that the combination of eye-tracking and EEG features decreases the signal-to-noise ratio even more than for only one type of cognitive processing signal. Another explanation is that the eye-tracking and EEG signals contain redundant information. This is always a risk when using co-registered data of EEG and eye-tracking signals within the same task. Specifically, eye movement artifacts could be contained in the EEG data. However, in this work, we use state-of-the-art methods to remove eye movement artifacts in the EEG data (through ICA and Unfold). In short, there are possible gains in performance to be achieved by more sophisticated combinations of eye movement and brain activity features.

There are various ways to leverage eye-tracking and EEG data. Currently, we extracted high-level eye-tracking features based on fixations (e.g, number of fixations and omission rate) and on saccades (e.g., mean velocity and maximum amplitude). The ZuCo dataset provides additional reading-related features such as mean fixation duration, total reading time or go-past time, but also pupil size information or even the raw data could be used in future approaches. Using raw data has shown great promise to model eye-tracking data (e.g., Jäger et al., 2020), and one of the main advantages of the ZuCo dataset is that it allows feature extraction on different levels. Moreover, our EEG features include mean features aggregated over all electrodes as well as electrode-based frequency measures, which have been shown to improve NLP tasks in the past (Hollenstein et al., 2019a, 2021b; Sun et al., 2020; Wang and Ji, 2021). Nonetheless, we want to highlight that preprocessed EEG data permits the examination of additional measures, such as source-level based features (e.g., source-level power estimates) and functional connectivity measures at the level of the underlying neuronal generators. Other EEG analysis methods allow the extract measures of spatio-temporal dynamics of brain activity (e.g., microstates) (Michel and Koenig, 2018) and event-related potentials such as N400 components (Frank et al., 2013; Brouwer et al., 2017). Interestingly, Hollenstein et al. (2021c) found that gamma band features worked best in a within-subject setting. However, we found that concatenating all EEG electrode features is more beneficial in a cross-subject setting. Finally, the cross-subject performance can be further increased by using a dimensionality reduction (PCA) on the concatenated EEG features. Future methods could focus on new approaches for EEG feature selection and aggregation.

The simultaneous recording of EEG and eye-tracking allows us to investigate specific feature sets on different levels of analysis, e.g., sentence level, word level, fixation level. Nevertheless, one should note that the ZuCo dataset includes reading individual sentences rather than full document, which influences the reading behavior. Reading studies with longer text spans should be considered in future work. Additionally, the naturalistic setup of the experiments used in this work are crucial for this benchmark task and for neuroscience in general (Nastase et al., 2020). Not only does it increase the ecological validity of the recordings by allowing natural reading without controlling the individual reading speed, but it also supports the extraction of signals on various linguistic levels (Hasson and Egidi, 2015; Brennan, 2016; Alday, 2019; Kandylaki and Bornkessel-Schlesewsky, 2019; Hamilton and Huth, 2020). Frey et al. (2018) investigated how two different reading tasks modulate both eye movements and brain activity. In line with our findings, their results show that eye movement patterns were top-down modulated by different task demands. Moreover, their brain activity analysis suggests that the decision-making process during task-specific reading elicits a greater load in working memory than the one generated in a normal reading task. In summary, eye-tracking and EEG data offer an immensely diverse amount of potential measures, which might contain unique valuable information. Thus, we aim to inspire benchmark challenge participants to explore and extract alternative features from the available preprocessed data.

7. Conclusion

We presented a new ML benchmark using eye-tracking and EEG data to classify reading tasks. The goal of the benchmark challenge is to distinguish between normal reading and task-specific reading in a cross-subject evaluation scenario. We provide multiple initial models for this task and show that ML models trained on eye-tracking and EEG features can outperform strong textual baselines.

The standardized Zurich Cognitive Language Processing Corpus (ZuCo) dataset facilitates the creation of such a machine learning benchmark. We use the ZuCo 2.0 dataset as training data. To make our benchmark task more robust, we have additionally recorded further eye-tracking and EEG data from natural reading from additional subjects in a hidden testset. ZuCo's rich structure and high-density coverage of simultaneous EEG and eye-tracking signals can also help to advance other areas that study the combination of gaze position and brain activity to identify variations in attention, reading patterns and

reading intents, as well as participants' compliance with the task demands and cross-subject variability.

Our dataset and benchmark setup allows us to easily add additional machine learning tasks to the leaderboard in the future. For instance, we can add additional NLP tasks since the ZuCo datasets provide ground truth labels for sentiment analysis or relation detection from text. Additionally, adding tasks such as eye movement and ERP prediction would be beneficial for various research communities. For example, the prediction of eye movement patterns has gained interest also in the NLP community (Hollenstein et al., 2021a). The main goal of this work is to create a platform for discussion and future research on a common benchmark task for reading task classification based on eye movement and brain activity data. We hope that this benchmark allows other researchers to make progress in this interdisciplinary research field.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary material.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics commission of the University of Zurich. The patients/participants provided their written informed consent to participate in this study.

Author contributions

NH: lead author, data collection, writing and editing, and machine learning. MT and MP: neuroscience experts, writing and editing, preprocessing and data analysis, and data collection. SK and YÖ: machine learning experts, writing and editing, and preprocessing and data analysis. LJ and NL: PIs and writing and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by the Swiss National Science Foundation under grant 100014_175875 (NL) and by the German Federal Ministry of Education and Research under grant 01|S20043 (LJ).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1028824/full#supplementary-material

References

Abdelrahman, Y., Khan, A. A., Newn, J., Velloso, E., Safwat, S. A., Bailey, J., et al. (2019). "Classifying attention types with thermal imaging and eye tracking," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3 (New York, NY: Association for Computing Machinery), 1–27. doi: 10.1145/3351227

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., et al. (2021). VATT: transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv*. [preprint]. doi: 10.48550/arXiv.2104.11178

Alday, P. M. (2019). M/EEG analysis of naturalistic stories: a review from speech to language processing. *Lang. Cogn. Neurosci.* 34, 457–473. doi: 10.1080/23273798.2018.1546882

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., and Søgaard, A. (2018). "Sequence classification with human attention," in *Proceedings of the 22nd Conference on Computational Natural Language Learning* (Brussels, Belgium: Association for Computational Linguistics), 302–312. doi: 10.18653/v1/K18-1030

Bautista, L. G., and Naval, P. (2020). "Towards learning to read like humans," in *International Conference on Computational Collective Intelligence* (New York, NY: Springer), 779–791. doi: 10.1007/978-3-030-63007-2_61

Bestgen, Y. (2021). "LAST at CMCL 2021 shared task: predicting gaze data during reading with a gradient boosting decision tree approach," in *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics* (Association for Computational Linguistics), 90–96. doi: 10.18653/v1/2021.cmcl-1.10

Biedert, R., Hees, J., Dengel, A., and Buscher, G. (2012). "A robust realtime reading-skimming classifier," in *Proceedings of the Symposium on Eye Tracking Research and Applications* (New York, NY: Association for Computing Machinery), 123–130. doi: 10.1145/2168556.2168575

Bonhage, C. E., Mueller, J. L., Friederici, A. D., and Fiebach, C. J. (2015). Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension. *Cortex* 68, 33–47. doi: 10.1016/j.cortex.2015.04.011

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897X00357

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Lang. Linguist. Compass* 10, 299–313. doi: 10.1111/lnc3.12198

Brennan, J. R., and Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS ONE* 14, e0207741. doi: 10.1371/journal.pone.0207741

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* 28, 803–809. doi: 10.1016/j.cub.2018.01.080

Brouwer, H., Crocker, M. W., Venhuizen, N. J., and Hoeks, J. C. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cogn. Sci.* 41, 1318–1352. doi: 10.1111/cogs.12461

Bruns, A. (2004). Fourier-, hilbert-and wavelet-based signal analysis: are they really different approaches? *J. Neurosci. Methods* 137, 321–332. doi: 10.1016/j.jneumeth.2004.03.002

Ceh, S. M., Annerer-Walcher, S., Koschutnig, K., Körner, C., Fink, A., Benedek, M., et al. (2021). Neurophysiological indicators of internal attention: an fMRI-eye-tracking coregistration study. *Cortex* 143, 29–46. doi: 10.1016/j.cortex.2021.07.005

Chiang, K.-J., Wei, C.-S., Nakanishi, M., and Jung, T.-P. (2019). "Cross-subject transfer learning improves the practicality of real-world applications of brain-computer interfaces," in *9th International IEEE/EMBS Conference on Neural Engineering* (San Francisco, CA), 424–427. doi: 10.1109/NER.2019.8716958

Choi, W., Desai, R. H., and Henderson, J. M. (2014). The neural substrates of natural reading: a comparison of normal and nonword text using eyetracking and fmri. *Front. Hum. Neurosci.* 8, 1024. doi: 10.3389/fnhum.2014.01024

Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J., Zhang, X., et al. (2011). Task and user effects on reading patterns in information search. *Interact. Comput.* 23, 346–362. doi: 10.1016/j.intcom.2011.04.007

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting GECO: an eyetracking corpus of monolingual and bilingual sentence reading. *Behav. Res. Methods* 49, 602–615. doi: 10.3758/s13428-016-0734-0

Culotta, A., McCallum, A., and Betz, J. (2006). "Integrating probabilistic extraction models and data mining to discover relations and patterns in text," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (New York, NY), 296–303. doi: 10.3115/1220835.1220873

[Dataset] Hollenstein, N., Tröndle, M., Plomecka, M., Kiegeland, S., Özyurt, Y., Jäger, L. A., et al. (2021c). Reading task classification using EEG and eye-tracking data. *arXiv [Preprint]. arXiv: 2112.06310*. Available online at: https://arxiv.org/pdf/2112.06310.pdf

[Dataset] Jäger, L., Kern, T., and Haller, P. (2021). Potsdam Textbook Corpus (PoTeC): Eye Tracking Data from Experts and Non-experts Reading Scientific Texts. Available on OSF. doi: 10.17605/OSF.IO/DN5HP

de Cheveignè, A. (2020). Zapline: a simple and effective method to remove power line artifacts. *Neuroimage* 207, 116356. doi: 10.1016/j.neuroimage.2019.116356

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., Liversedge, S. P., et al. (2019). Parafoveal previews and lexical frequency in natural reading: evidence from eye movements and fixation-related potentials. *J. Exp. Psychol. Gen.* 148, 453. doi: 10.1037/xge0000494

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., et al. (2017). Decoding the neural representation of story meanings across languages. *Hum. Brain Mapp.* 38, 6096–6106. doi: 10.1002/hbm.23814

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Minneapolis, MN: Association for Computational Linguistics), 4171–4186.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., and Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *J. Exp. Psychol. Gen.* 140, 552. doi: 10.1037/a0023885

Ehinger, B. V., and Dimigen, O. (2019). Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ* 7, e7838. doi: 10.7717/peerj.7838

Finke, A., Essig, K., Marchioro, G., and Ritter, H. (2016). Toward FRP-based brain-machine interfaces–single–trial classification of fixation-related potentials. *PLoS ONE* 11, e0146848. doi: 10.1371/journal.pone.0146848

Flesch, R. (1948). A new readability yardstick. *J. Appl. Psychol.* 32, 221. doi: 10.1037/h0057532

Frank, S. L., and Aumeistere, A. (2022). An eye-tracking-with-EEG coregistration corpus of narrative sentences. *PsyArXiv*. doi: 10.31234/osf.io/j5fgd

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2013). "Word surprisal predicts n400 amplitude during reading," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (Sofia, Bulgaria: Association for Computational Linguistics), 878–883.

Frey, A., Lemaire, B., Vercueil, L., and Guèrin-Duguè, A. (2018). An eye fixation-related potential study in two reading tasks: reading to memorize and reading to make a decision. *Brain Topogr.* 31, 640–660. doi: 10.1007/s10548-018-0629-8

Hamilton, L. S., and Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. *Lang. Cogn. Neurosci.* 35, 573–582. doi: 10.1080/23273798.2018.1499946

Hasson, U., and Egidi, G. (2015). "What are naturalistic comprehension paradigms teaching us about language?" in *Cognitive Neuroscience of Natural Language Use*, ed R. M. Willems (Cambridge: Cambridge University Press), 228–255. doi: 10.1017/CBO9781107323667.011

Hollenstein, N., Barrett, M., Troendle, M., Bigiolli, F., Langer, N., Zhang, C., et al. (2019a). Advancing NLP with cognitive language processing signals. *arXiv [Preprint]*. doi: 10.48550/arXiv.1904.02682

Hollenstein, N., and Beinborn, L. (2021). "Relative importance in sentence processing," in Proceedings of the *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Association for Computational Linguistics), 141–150. doi: 10.18653/v1/2021.acl-short.19

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., Santus, E., et al. (2021a). "CMCL 2021 shared task on eye-tracking prediction," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 72–78. doi: 10.18653/v1/2021.cmcl-1.7

Hollenstein, N., de la Torre, A., Langer, N., and Zhang, C. (2019b). "CogniVal: a framework for cognitive word embedding evaluation," in *Proceedings of the 23nd Conference on Computational Natural Language Learning* (Hong Kong: Association for Computational Linguistics), 538–549. doi: 10.18653/v1/K19-1050

Hollenstein, N., Renggli, C., Glaus, B., Barrett, M., Troendle, M., Langer, N., et al. (2021b). Decoding EEG brain activity for multi-modal natural language processing. *Front. Hum. Neurosci.* 15, 378. doi: 10.3389/fnhum.2021.659410

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., Langer, N., et al. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Sci. Data* 5, 180291. doi: 10.1038/sdata.2018.291

Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. (2020). "ZuCo 2.0: a dataset of physiological recordings during natural reading and annotation," in *Proceedings of The 12th Language Resources and Evaluation Conference* (Marseille), 138–146.

Jäger, L. A., Makowski, S., Prasse, P., Liehr, S., Seidler, M., Scheffer, T., et al. (2020). "Deep eyedentification: biometric identification using micro-movements of the eye," *in Machine Learning and Knowledge Discovery in Databases. Proceedings of the 2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (Würzburg, Germany), 299–314. doi: 10.1007/978-3-030-46147-8_18

Kandylaki, K. D., and Bornkessel-Schlesewsky, I. (2019). From story comprehension to the neurobiology of language. *Lang. Cogn. Neurosci.* 34, 405–410. doi: 10.1080/23273798.2019.1584679

Kelton, C., Wei, Z., Ahn, S., Balasubramanian, A., Das, S. R., Samaras, D., et al. (2019). "Reading detection in real-time," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research and Applications* (Denver, CO: ACM Press), 1–5. doi: 10.1145/3314111.3319916

Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A. M., and Sommer, W. (2012). Eye movements and brain electric potentials during reading. *Psychol. Res.* 76, 145–158. doi: 10.1007/s00426-011-0376-x

Lemhöfer, K., and Broersma, M. (2012). Introducing LexTALE: a quick and valid lexical test for advanced learners of English. *Behav. Res. Methods* 44, 325–343. doi: 10.3758/s13428-011-0146-0

Lobo, J. L., Ser, J. D., De Simone, F., Presta, R., Collina, S., and Moravek, Z. (2016). "Cognitive workload classification using eye-tracking and EEG data," in *Proceedings of the International Conference on Human-Computer Interaction in Aerospace* (New York, NY), 1–8. doi: 10.1145/2950112.2964585

Lopopolo, A., Frank, S. L., Van den Bosch, A., Nijhof, A., and Willems, R. M. (2018). "The Narrative Brain Dataset (NBD), an fMRI dataset for the study of natural language processing in the brain," in *LREC 2018 Workshop on Linguistic and Neuro-Cognitive Resources (LiNCR)* (Paris: LREC), 8–11.

Luke, S. G., and Christianson, K. (2018). The provo corpus: a large eye-tracking corpus with predictability norms. *Behav. Res. Methods* 50, 826–833. doi: 10.3758/s13428-017-0908-4

Mathias, S., Kanojia, D., Mishra, A., and Bhattacharya, P. (2020). "A survey on using gaze behaviour for natural language processing," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (Yohokama), 4907–4913. doi: 10.24963/ijcai.2020/683

Mathur, P., Mittal, T., and Manocha, D. (2021). "Dynamic graph modeling of simultaneous EEG and eye-tracking data for reading task identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Toronto, ON), 1250–1254. doi: 10.1109/ICASSP39728.2021.9414343

MathWorks, Inc. (2000). *MATLAB: The Language of Technical Computing. External interfaces*. MathWorks, Incorporated.

McGuire, E., and Tomuro, N. (2021). "Relation classification with cognitive attention supervision," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 222–232. doi: 10.18653/v1/2021.cmcl-1.26

Michel, C. M., and Koenig, T. (2018). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage* 180, 577–593. doi: 10.1016/j.neuroimage.2017.11.062

Miller, B. W. (2015). Using reading times and eye-movements to measure cognitive engagement. *Educ. Psychol.* 50, 31–42. doi: 10.1080/00461520.2015.1004068

Nastase, S. A., Gazzola, V., Hasson, U., and Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Soc. Cogn. Affect. Neurosci.* 14, 667–685. doi: 10.1093/scan/nsz037

Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage* 222, 117254. doi: 10.1016/j.neuroimage.2020.117254

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., et al. (2021). Narratives: fMRI data for evaluating models of naturalistic language comprehension. *bioRxiv*. doi: 10.1101/2020.12.23.424091

Notaro, G. M., and Diamond, S. G. (2018). Simultaneous EEG, eye-tracking, behavioral, and screen-capture data during online German language learning. *Data Brief* 21, 1937–1943. doi: 10.1016/j.dib.2018.11.044

Pedroni, A., Bahreini, A., and Langer, N. (2019). Automagic: standardized preprocessing of big EEG data. *Neuroimage* 200, 460–473. doi: 10.1016/j.neuroimage.2019.06.046

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., et al. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* 9, 963. doi: 10.1038/s41467-018-03068-4

Pfeiffer, C., Hollenstein, N., Zhang, C., and Langer, N. (2020). Neural dynamics of sentiment processing during naturalistic sentence reading. *Neuroimage* 218, 116934. doi: 10.1016/j.neuroimage.2020.116934

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). Iclabel: an automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* 198, 181–197. doi: 10.1016/j.neuroimage.2019.05.026

Raatikainen, P., Hautala, J., Loberg, O., Kärkkäinen, T., Leppänen, P., Nieminen, P., et al. (2021). Detection of developmental dyslexia with machine learning using eye movement data. *Array* 12, 100087. doi: 10.1016/j.array.2021.100087

Rämä, P., and Baccino, T. (2010). Eye fixation-related potentials (EFRPs) during object identification. *Vis. Neurosci.* 27, 187–192. doi: 10.1017/S0952523810000283

Rello, L., and Ballesteros, M. (2015). "Detecting readers with dyslexia using machine learning with eye tracking measures," in *Proceedings of the 12th International Web for All Conference* (New York, NY: Association for Computing Machinery), 1–8. doi: 10.1145/2745555.2746644

Schalk, G., Brunner, P., Gerhardt, L. A., Bischof, H., and Wolpaw, J. R. (2008). Brain-computer interfaces (BCIS): detection instead of classification. *J. Neurosci. Methods* 167, 51–62. doi: 10.1016/j.jneumeth.2007.08.010

Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. M. R., Salehi, M., et al. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage* 193, 35–45. doi: 10.1016/j.neuroimage.2019.02.057

Sereno, S. C., and Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends Cogn. Sci.* 7, 489–493. doi: 10.1016/j.tics.2003.09.010

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* 138, 107307. doi: 10.1016/j.neuropsychologia.2019.107307

Sun, P., Anumanchipalli, G. K., and Chang, E. F. (2020). Brain2Char: a deep architecture for decoding text from brain recordings. *J. Neural Eng.* 17, 066015. doi: 10.1088/1741-2552/abc742

Tokunaga, T., Nishikawa, H., and Iwakura, T. (2017). An eye-tracking study of named entity annotation. *Proceedings of the International Conference Recent Advances in Natural Language Processing* (Varna, Bulgaria: INCOMA Ltd.), 758–764. doi: 10.26615/978-954-452-049-6_097

Tomanek, K., Hahn, U., Lohmann, S., and Ziegler, J. (2010). "A cognitive cost model of annotations based on eye-tracking data," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Uppsala, Sweden), 1158–1167.

Tor, H. T., Ooi, C. P., Lim-Ashworth, N. S. J., Wei, J. K. E., Jahmunah, V., Oh, S. L., et al. (2021). Automated detection of conduct disorder and attention deficit hyperactivity disorder using decomposition and nonlinear techniques with EEG signals. *Comput. Methods Programs Biomed.* 200, 105941. doi: 10.1016/j.cmpb.2021.105941

Wang, Z., and Ji, H. (2021). Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. *arXiv*. [preprint]. doi: 10.48550/arXiv.2112.02690

Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., et al. (2019). EvalAI: towards better evaluation systems for AI agents. *arXiv*. [preprint]. doi: 10.48550/arXiv.1902.03570

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 02 February 2023
DOI 10.3389/fpsyg.2023.1112365

OPEN ACCESS

EDITED BY
Nora Hollenstein,
University of Copenhagen, Denmark

REVIEWED BY
Joseph Marvin Imperial,
University of Bath, United Kingdom
Yohei Oseki,
The University of Tokyo, Japan

*CORRESPONDENCE
Lavinia Salicchi
✉ lavinia.salicchi@connect.polyu.hk

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 30 November 2022
ACCEPTED 13 January 2023
PUBLISHED 02 February 2023

CITATION
Salicchi L, Chersoni E and Lenci A (2023) A study on surprisal and semantic relatedness for eye-tracking data prediction.
Front. Psychol. 14:1112365.
doi: 10.3389/fpsyg.2023.1112365

A study on surprisal and semantic relatedness for eye-tracking data prediction

Lavinia Salicchi[1]*, Emmanuele Chersoni[1] and Alessandro Lenci[2]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China, [2]Computational Linguistics Laboratory (CoLing Lab), University of Pisa, Pisa, Italy

Previous research in computational linguistics dedicated a lot of effort to using language modeling and/or distributional semantic models to predict metrics extracted from eye-tracking data. However, it is not clear whether the two components have a distinct contribution, with recent studies claiming that surprisal scores estimated with large-scale, deep learning-based language models subsume the semantic relatedness component. In our study, we propose a regression experiment for estimating different eye-tracking metrics on two English corpora, contrasting the quality of the predictions with and without the surprisal and the relatedness components. Different types of relatedness scores derived from both static and contextual models have also been tested. Our results suggest that both components play a role in the prediction, with semantic relatedness surprisingly contributing also to the prediction of function words. Moreover, they show that when the metric is computed with the contextual embeddings of the BERT model, it is able to explain a higher amount of variance.

KEYWORDS

cognitive modeling, surprisal, semantic relatedness, cosine similarity, language models, distributional semantics, eye-tracking

1. Introduction

Eye-tracking data recorded during reading provide important evidence about the factors influencing language comprehension (Rayner et al., 1989; Rayner, 1998). In the investigation of potential predictors of human reading patterns, cognitive studies have focused their attention on two specific factors, among the others: (i) the semantic coherence of a word with the rest of the sentence (Ehrlich and Rayner, 1981; Pynte et al., 2008; Mitchell et al., 2010), which is typically assessed *via semantic relatedness* metrics (usually the *cosine*) computed with *distributional word embeddings*, and (ii) the predictability of the word from its previous context, as measured by *surprisal* (Hale, 2001; Levy, 2008). Initially, the two factors were considered separately, and the general idea was that words having low semantic coherence and low in-context predictability (i.e., high surprisal) induce longer reading times. This hypothesis was instead questioned by Frank (2017), who argued that previous findings had to be attributed to a confound between semantic relatedness and word predictability and that the effect of the former disappeared once surprisal was factored out.

Our work aims at providing further evidence about the complex interplay between semantic relatedness and surprisal as predictors of eye-tracking data. For example, it is unclear whether the fact that no independent effect of relatedness has been found depends on the specific word embedding model being used for measuring it. In fact, there is a large variety of Distributional Semantic Models (DSMs) that are trained with different objectives, and they have been shown to perform differently depending on the task (Lenci et al., 2022). Moreover, the recent introduction of contextual embedding models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) has also radically changed the way semantic relatedness can be assessed. In particular,

contextual embeddings now make it possible to compare the semantic representations of *words in specific contexts* (*token-level representations*), and not just type-level representations that tend to conflate multiple senses of the same word.

The goals of this paper can thus be summarized as follows:

1. Investigating whether distributional measures of semantic relatedness between a word and its previous contexts are indeed made redundant by surprisal, or have instead an autonomous explanatory role to model eye-tracking data;
2. Looking into different types of word embeddings, to check whether "classical" static models and contextual ones interact differently or not with surprisal.

To explore these issues, we implemented four different linear models to predict three eye-tracking features on two eye-tracking corpora: i) a baseline with word-level features, ii) a model with baseline features and the surprisal between target word and context, iii) a model with baseline features and the relatedness between the vector representing the target word and the vector representing the context, and iv) a model with all the above-mentioned regression features. While surprisal has been consistently computed using a state-of-the-art neural language model GPT2-xl (Radford et al., 2019), the vectors employed in the cosine similarity calculation were obtained using either SGNS (Mikolov et al., 2013) or BERT (Devlin et al., 2019), to compare static and contextual word embedding models.

Our results show that the models including both relatedness and surprisal perform better than the other three, suggesting that, despite the overlap between the two, they contribute differently in explaining the variance in the data. Furthermore, when comparing the models using only relatedness, we noticed that BERT vectors outperform SGNS ones, confirming the added value of contextual embeddings when modeling the relatedness of words in contexts. Finally, we investigated how our models predict eye-tracking feature values for different parts of speech, and we found that while surprisal helps on content words, semantic relatedness contributes to improving the predictions on both function and content words.

2. Computational models of human reading times: Surprisal and semantic relatedness

Since the cognitive processes of meaning construction involve the integration of individual word meanings into the syntactic and semantic context, the literature in natural language processing and cognitive science got interested in how such contextual effects on word fixations could be modeled. A first class of computational models has relied on distributional semantics to assess the relatedness of a word with its wider semantic context (Section 2.1); another class of models has explored the connection between the logarithmic probabilities of words in context and their processing difficulty (Section 2.2).

2.1. Computational measures for semantic coherence

A fruitful line of research has been investigating the usage of cosine similarity between word embeddings for predicting reading times. The employment of word vectors for modeling reading times originated from classical DSMs (Lenci and Sahlgren, 2023). Pynte et al. (2008) and Mitchell et al. (2010) used the semantic distance between a target word and the context as a predictor, measured as 1 min the traditional cosine similarity metric (Turney and Pantel, 2010; Lenci, 2018). The context was in turn modeled as the sum of the distributional vectors representing the words before the target. These studies found strong correlations between semantic distance and reading times: The more semantically related the words, the shorter the fixation durations.

Originally, vector spaces were obtained from the extraction and counting (hence the name of *count models*) of the co-occurrences between the target words and the relevant linguistic contexts. Raw co-occurrences were usually weighted *via* different types of statistical association measures [e.g., Mutual Information, log-likelihood; see Evert (2005) for an overview] and then the vector space was optionally transformed with some algebraic operation for dimensionality reduction, such as Singular Value Decomposition (Landauer and Dumais, 1997; Bullinaria and Levy, 2012). The contexts could consist either in the words occurring within a window surrounding the target (Lund and Burgess, 1996; Sahlgren, 2008), or in the words linked to the target by syntactic (Padó and Lapata, 2007; Baroni and Lenci, 2010) or semantic relations (Sayeed et al., 2015).

Later, with the increasing success of deep learning techniques in Natural Language Processing, the so-called *predict models* established themselves as a new standard (Mikolov et al., 2013; Bojanowski et al., 2017). In such models, the learning of word vectors is based on neural network training and framed as a self-supervised language modeling task. One of the most popular predict DSMs is Word2Vec (Mikolov et al., 2013), which includes two main architectures: CBOW, trained for predicting a target word given the context surrounding it, and Skip-Gram, whose learning objective is to predict the surrounding context given a target word. The most common implementation of Skip-Gram makes use of negative sampling (SGNS), whose objective is to discriminate between word sequences that are actually occurring in the data (positive samples) and "corrupted" samples, which are obtained by randomly replacing a word in a true sequence from the corpus (negative samples).

One of the main limitations of "traditional" word embeddings, both count and predict ones, is that they provide *static* representations of the semantics of a word. They assign a single embedding to each word type, thereby conflating the possible senses of a lexeme and hampering the possibility to address the pervasive phenomena of polysemy and homography. For example, *bank* as a financial agency will have the same vector representation of *bank* as the bank of the river. This way, lexical semantic representations are built at the *type* level only, and the embedding will be a sort of distributional summary of all the instances of a word, no matter how different their senses might be (and probably, the most frequent senses would obscure the minority ones).

The most recent generation of DSMs is said to be *contextual* because they produce a distinct vector for each word instance in context, that is a *token* level representation(Peters et al., 2018; Devlin

et al., 2019; Liu et al., 2019). Contextual DSMs generally rely on a multi-encoder network and the word vectors are learned as a function of the internal states, so that a word appearing in different sentence contexts determines different activation states and, as a consequence, is represented by a different vector.

Most contextual DSMs are based on *Transformers* (Vaswani et al., 2017), which use a self-attention mechanism (Bahdanau et al., 2014) for getting the most salient elements in a sentence context and assign them higher weights. BERT (Devlin et al., 2019) is probably the most popular model for generating contextual word representations. BERT is trained on a masked language modeling objective function: random words in the input sentences are replaced by a `'[MASK]'` token and the model attempts to predict the masked word based on the surrounding context. Simultaneously, BERT is optimized on a next sentence prediction task, as the model receives sentence pairs in input and has to predict whether the second sentence is subsequent to the first one in the training data. It should be noticed that BERT is defined as *deeply bidirectional* as, in fact, it takes into account the left-hand and the right-hand context of a word to predict the word filling the masked token. The contextual embeddings produced by BERT have been shown to improve the state-of-the-art performance in several Natural Language Processing tasks (Devlin et al., 2019) and it has been reported that its multilingual versions (i.e., Multilingual BERT, XLM) are able to predict human fixations in multiple languages (Hollenstein et al., 2021, 2022a,b). Significantly, it was shown that it is possible to extract semantic representations at the type level from BERT just by averaging token vectors of randomly-sampled sentences, and those can achieve a performance close to traditional word embeddings on word similarity tasks (Bommasani et al., 2020; Chronis and Erk, 2020; Lenci et al., 2022) and on word association modeling (Rodriguez and Merlo, 2020).

2.2. Computational measures for word predictability

A significant part of the psycholinguistic and computational studies modeled naturalistic reading data by means of language model probabilities, being inspired by *surprisal theory* (Hale, 2001, 2016), with the idea that the predictability of a word is the main factor determining the reading times. More specifically, the processing difficulty of a word is considered to be proportional to its *surprisal*, that is, the negative logarithm of the probability of the word given the context. Several studies based on language models adopted surprisal theory as a reference framework for the prediction of eye-tracking data (Demberg and Keller, 2008; Frank and Bod, 2011; Fossum and Levy, 2012; Monsalve et al., 2012; Smith and Levy, 2013). The predictions were typically evaluated on the Dundee Corpus (Kennedy et al., 2003), as one of the earliest corpora with gold standard annotations of eye-tracking measures.

Later research has focused on the quality of the language model to estimate conditional probabilities, finding that models with lower perplexity are a better fit to human reading times (Goodkind and Bicknell, 2018). Following studies confirmed the model perplexity as a significant determinant, making use of more and more advanced neural architectures, such as LSTM (van Schijndel and Linzen, 2018), GRU (Aurnhammer and Frank, 2019), Transformers (Merkx and Frank, 2021), GPT-2 (Wilcox et al., 2020).

Is contextual predictability, that is surprisal, all we need to model human reading behavior? Some recent results suggest that this may not be the case. Goodkind and Bicknell (2021), for example, investigated the role played on local word statistics, such as word bigram and trigram probability, in sentence processing, and consequently their impact on reading times, finding that they affect processing independently of surprisal. Moreover, Hofmann et al. (2021) compared different models for computing surprisal as predictors of eye-tracking fixations and found that they explain different and independent proportions of variance in the viewing parameters. For example, classical n-gram-based language models are better at predicting metrics related to short-range access, while RNN models better predict the early preprocessing of the next word.

The models of the GPT family are based on Transformer architectures (Radford et al., 2018, 2019; Brown et al., 2020). Differently from BERT, GPT is a uni-directional, autoregressive Transformer language model, which means that the training objective is to predict the next word, given all of the previous words. GPT-2, in particular, has been commonly used in eye-tracking studies, as the surprisal scores computed by this language model have been proved to be strong predictors of reading times and eye fixations in English (Hao et al., 2020; Wilcox et al., 2020; Merkx and Frank, 2021) and in other languages (e.g., Dutch, German, Hindi, Chinese, Russian) (Salicchi et al., 2022).

The research work on semantic relatedness and surprisal led Frank (2017) to ask whether these two factors have actually independent effects in the modeling of reading times. The question was motivated by the fact that not all the studies on reading times found effects associated with semantic relatedness (e.g., Traxler et al., 2000; Gordon et al., 2006), although vector space metrics clearly proved to be useful for modeling other types of experimental data on naturalistic reading, such as the N400 amplitude in EEG recordings (Frank and Willems, 2017). Frank suggested that, since DSMs like Word2Vec (Mikolov et al., 2013) are based on word co-occurrence and are optimized for predicting words in context, previous results were due to a confound between semantic relatedness and word predictability. Indeed, when surprisal was factored out, the author showed that the semantic distance effects disappeared. Moreover, the different results obtained in modeling the N400 component in the EEG data were attributed to differences in the stimuli presentation method: while in eye-tracking participants read the text naturally, in many EEG studies words are presented one at a time with unnaturally long durations. Following the findings of Wlotko and Federmeier (2015) and Frank (2017) pointed out that, the more natural the presentation rates of the words in the experimental setting in EEG, the smaller the semantic relatedness effects on N400 data tend to be, with no effects at all for behavioral metrics on naturalistic reading. Is distributional semantic relatedness really made redundant by surprisal, or were the results by Frank (2017) also conditioned by the specific type of embeddings used in the experiments? The analyzes in Sections 3, 4 aim at clarifying this issue.

3. Materials and methods

3.1. Definition of eye-tracking metrics in psycholinguistic studies

Several metrics have been defined to describe eye movement features (Rayner, 1998). In this work, we focus on first fixation

duration, number of fixations and total reading time. The first fixation duration (FFD), that is the time spent fixing a word for the first time, is typically associated with lexical information processing, like lexical access (Inhoff, 1984), which is heavily affected by word frequency (Balota and Chumbley, 1984). Fast word recognition is obtained when a word can be recognized with a single glance. In this sense, a short FFD reflects a quick and successful lexical access (Hofmann et al., 2021).

However, several words may not be accessed immediately. Words may receive multiple fixations before the eyes move to the next word, and this is reflected by the number of fixations (NF), depending on the integration of the word within the sentence semantics or syntax (Frazier and Rayner, 1982). An alternative metric for this "delayed" lexical access is known as *gaze duration*, which computes directly the sum of the duration of individual fixations before moving to the next word (Inhoff and Radach, 1998; Rayner, 1998).

Finally, the total reading time (TRT), as the sum of all fixation durations on the word, including regressions, is affected by both lexical and sentence-level processing. The TRT is likely to indicate the time required for the full semantic integration of the word in the sentence context (Radach and Kennedy, 2013).

What are the factors affecting word fixations during reading? There is a general consensus that word position, word length, and the number of syllables within the word affect language processing and, consequently, reading behavior and fixations (Just and Carpenter, 1980). It has also been observed that low-frequency words tend to have longer gaze durations and, additionally, they lead to longer gaze on the immediately following words, a phenomenon typically referred to as *spillover effect* (Rayner and Duffy, 1986; Rayner et al., 1989; Remington et al., 2018). A common explanation is that rare and longer words have a higher cognitive load, as they require more time for the semantic integration in the sentence context (Pollatsek et al., 2008), and therefore they may influence the processing of the following words.

3.2. Eye-tracking corpora

Traditional corpora annotated with eye-tracking data consist of short isolated sentences (or even single words) with particular structures or lexemes, in order to investigate specific syntactic and semantic phenomena. In the present work, we use GECO (Cop et al., 2017) and Provo (Luke and Christianson, 2018), two eye-tracking corpora containing long, complete, and coherent texts.

GECO is a bilingual corpus in English and Dutch composed of the entire Agatha Christie's novel *The Mysterious Affair at Styles*. The corpus is freely downloadable with a related dataset containing eye-tracking data of 33 subjects (19 of them bilingual, 14 English monolingual) reading the full novel text, presented paragraph-by-paragraph on a screen[1]. In total, GECO is composed of 54,364 tokens.

Provo contains 55 short English texts about various topics, with 2.5 sentences and 50 words on average, for a total of 2,689 tokens, and a vocabulary of 1,197 words. These texts were read by 84 native English speakers and their eye-tracking measures were collected and made publicly available online[2].

GECO and Provo are particularly interesting for our goals because they are corpora of naturalistic reading since data have been recorded from subjects reading real texts, instead of short stimuli created *in vitro*. For every word in the corpora, we extracted the mean total reading time, mean first fixation duration, and mean number of fixations. Mean values were obtained by averaging over the subjects. The choice of modeling mean eye-tracking measures is justified by the high inter-subject consistency of the recorded data.

TABLE 1 Summary of the linear models implemented for the experiments.

Model name	Features
BL	Word frequency Word length Word position within the sentence Previous word frequency Previous word length Whether or not the previous word was fixated
BL-cos	Baseline features (same as BL) Cosine similarity (BERT vectors)
	Baseline features (same as BL) Cosine similarity (SGNS vectors)
BL-sur	Baseline features (same as BL) Surprisal (GPT2-xl)
BL-sur-cos	Baseline features (same as BL) Surprisal (GPT2-xl) Cosine similarity (SGNS vectors)
	Baseline features (same as BL) Surprisal (GPT2-xl) Cosine similarity (BERT vectors)

3.3. Method

We implemented and compared four main types of linear models (see Table 1):

1. A baseline model with word-related statistics that are known to influence sentence and word processing (i.e., word frequency, word length, word position within the sentence, previous word frequency, previous word length, and whether or not the previous word was fixated);
2. Two models combining baseline features and cosine similarity, one using Skip-Gram vectors (SGNS), one using BERT vectors;
3. One model with baseline features + surprisal computed using GPT2-xl;
4. Two models with baseline features + surprisal computed using GPT2-xl + cosine similarity, one using SGNS vectors, one using BERT vectors.

Recent works have cast doubts on the application of cosine in similarity task while employing contextual vector models. In fact, in contextual embeddings a small number of dimensions (e.g., 3-5) tend to dominate the similarity metric, accounting for most of the data variance (Timkey and van Schijndel, 2021). Moreover, it has been shown that the removal of the outlier dimensions leads to drastic

1 https://expsy.ugent.be/downloads/geco/

2 https://osf.io/sjefs/

performance drops both in language modeling and in downstream tasks (Kovaleva et al., 2021).

To address this issue, for similarity tasks it has been suggested to correct the comparisons by discounting the "rogue" dimensions or to adopt metrics based on the rank of the dimensions themselves, rather than on their absolute values (Timkey and van Schijndel, 2021). In order to take into account the potential effect of rogue dimensions on computing cosine similarity with BERT, we followed the latter suggestion and we also implemented two further models, in which we use Spearman correlation instead of cosine similarity.

Rank-based metrics have been reported to outperform vector cosine in semantic relatedness tasks (Santus et al., 2016a,b, 2018; Zhelezniak et al., 2019), and it has been shown that Spearman itself is more correlated with human judgments than cosine (Timkey and van Schijndel, 2021). For each of the resulting eight models, the values to be predicted were first fixation duration (FFD), number of fixations (NF) and total reading time (TRT). We predicted those metrics on both GECO and Provo corpus. We also experimented with models with and without interactions between the features. The models were implemented using the generalized linear models available in R, which have also been used for the statistical analysis.

After we fitted the data of the eye-tracking features with each model, we compared them using the corrected Akaike Information Criterion (AICc) in order to determine the extent to which the goodness of fit improves with the addition of semantic relatedness and surprisal as predictors. Additionally, we also analyzed i) the correlations between linear model errors (as Mean Absolute Error, MAE) and word features, and ii) which parts of speech are easier or harder for each model to predict.

3.4. Regression features

3.4.1. Baseline features

The baseline model includes the following word features: i) the target word and previous word length, computed as the number of letters within the word to be predicted; ii) the target word and previous word frequency, whose values are extracted from Wikipedia;[3] iii) the target word position, as the index of the word within the current sentence; iv) a Boolean value corresponding to 1 if the word preceding the target word was fixated, 0 otherwise. The baseline features are the same used by Frank (2017).

3.4.2. Metrics of semantic relatedness

To compute the semantic relatedness between the context and the target word, we extracted vectors for each word, represented the sentence context with a vector, and finally computed, alternatively, the *cosine similarity* or the *Spearman correlation* between the context and the target vectors (the latter metric was used only with the BERT vectors only).

With **SGNS** embeddings, we extracted the pre-trained vectors for each word, and we computed the context vector using an additive model: We summed the vectors of all the words preceding the target and took this as the context representation. For example, given the sentence *The dog chases the cat*, if the target word is *chases*, the context vector will be $\overrightarrow{The} + \overrightarrow{dog}$, while if the target word is *cat*, the context vector will be $\overrightarrow{The} + \overrightarrow{dog} + \overrightarrow{chases} + \overrightarrow{the}$.

On the other hand, given the bidirectional nature of the **BERT** language model, the input to extract the embeddings from this model required a special preprocessing, since we wanted to avoid the model to "see the future," by having the target word vector including information also from the right-hand context. Therefore, we fed BERT with sub-sentences. For instance, given the sentence *The dog chases the cat*, we generated the following sub-sentences:

S[0] = [*The*]
S[1] = [*The dog*]
S[2] = [*The dog chases*]
S[3] = [*The dog chases the*]
S[4] = [*The dog chases the cat*]

For each target word, we extracted its vector, when the lexeme occurs at the end of a sub-sentence (e.g., *The* will be extracted in S[0], *dog* in S[1], *chases* in S[2], and so on).

Regarding the context, we used the vector of the special token [CLS], which is created by BERT as a global representation of the input sentence, taking into account how salient each word is for the sentence's meaning. Again, to avoid a representation of the target word itself within the [CLS] vector, we computed the cosine similarity and the Spearman correlation between the target word embedding, and the [CLS] vector of the previous sub-sentence. For example, if *cat* is the target word, we computed the cosine similarity between $\overrightarrow{cat}$ from S[4] and $\overrightarrow{CLS_{S[3]}}$. In order to find the optimal layer for the computation of the similarity scores, we extracted vectors from all the 24 layers of BERT Large and computed the Spearman correlations with each one of the target features.

The results can be seen in Figure 1. Consistently with the findings of Salicchi et al. (2021), the layers with the highest absolute correlation values are the ones immediately before the last one. We chose layer 22 as the one with the highest inverse correlation to our data.

3.4.3. Surprisal

To model the influence of word predictability on eye-tracking measures, we included in the regression models the surprisal of the target words given their previous context. For each target word we computed the surprisal as the negative logarithm of its probability given all the words preceding the target:

$$surprisal(w_n) = -\log P(w_n|w_0, w_1, ..., w_{n-1}) \quad (1)$$

The probability P is computed by GPT2-xl, the largest publicly available version of GPT-2. Similarly to the original model, GPT2-xl was also trained on the WebText corpus (40 GB of text data), but it has a larger architecture (48 layers, for a total of 1542M parameters) and was shown to have the

3 The Wikipedia frequencies were extracted from https://github.com/IlyaSemenov/wikipedia-word-frequency

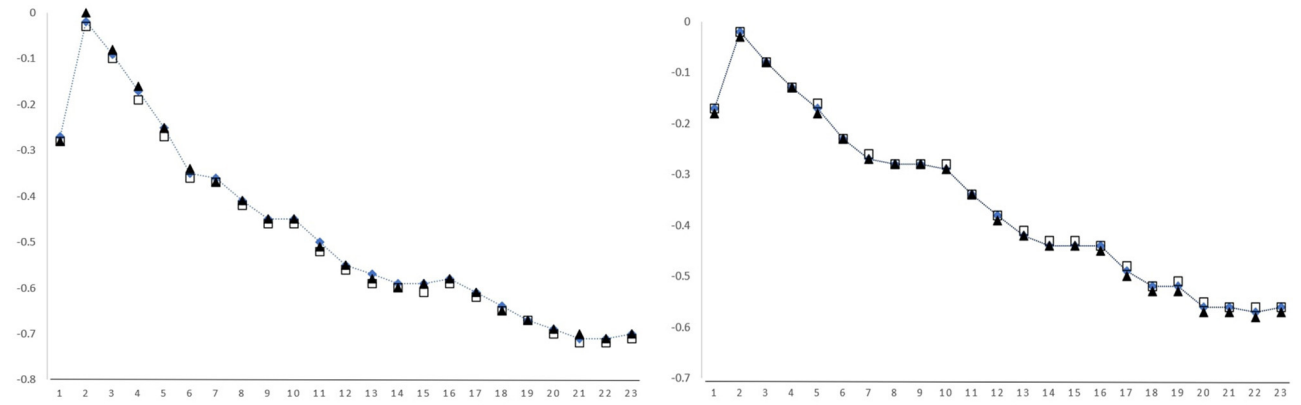


FIGURE 1
Spearman correlations between TRT (dot line), FFD (square), and NF (triangle) and the cosine similarity using vectors produced by different layers of BERT Large, on GECO **(left)** and Provo **(right)**. Layer 24, whose values are systematically higher than the average, is intentionally left out for plot reading purposes.

TABLE 2 Average AICc, and AICc for TRT, FFD, and NF on GECO with SGNS vectors.

	Avg		TRT		FFD		NF	
Model	AICc	Delta	AICc	Delta	AICc	Delta	AICc	Delta
BL-sur-cos	60,286	0	88,611	0	80,296	0	11,951	0
BL-sur	60,492	206	88,835	224	80,576	280	12,065	115
BL-cos	60,982	696	89,409	798	80,903	607	12,634	683
BL	61,466	1,180	89,948	1,337	81,483	1,186	12,969	1,018

TABLE 3 Average AICc, and AICc for TRT, FFD, and NF on GECO with BERT vectors.

	Avg		TRT		FFD		NF	
Model	AICc	Delta	AICc	Delta	AICc	Delta	AICc	Delta
BL-sur-cos	59,566	0	87,758	0	79,232	0	11,709	0
BL-cos	60,151	585	88,413	654	79,697	465	12,346	637
BL-sur-Spearman	60,467	901	88,803	1,045	80,538	1,307	12,060	350
BL-sur	60,492	926	88,835	1,077	80,576	1,345	12,065	356
BL-Spearman	61,430	1,864	89,902	2,145	81,432	2,200	12,957	1,247
BL	61,466	1,900	89,948	2,190	81,483	2,251	12,969	1,259

lowest perplexity on the evaluation corpora of Radford et al. (2019).

4. Results and discussion

4.1. General analysis

4.1.1. Cosine similarity vs. Spearman correlation

We first checked whether Spearman correlation was a better similarity metric than cosine with BERT contextual embeddings. Therefore, we compared BL-cos and BL-Spearman, namely models with baseline features and the similarity metric only, and we compared BL-sur-cos and BL-sur-Spearman, which are the models using baseline features, surprisal, and the similarity metric. The AICc values reported in Tables 2–5 clearly show that cosine similarity is a better predictor of eye-tracking features than Spearman correlation: on GECO, the difference between BL-cos and BL-Spearman is 1,279, and between BL-sur-cos and BL-sur-Spearman is 901; on Provo the differences are 333 and 318, respectively. Given these results, we henceforth focus our analyzes only on cosine similarity and its relationship with surprisal. Our findings suggest that, within the linear models we propose, BERT embeddings anisotropy does not affect the eye movements modeling, and therefore, cosine similarity is a suitable feature to be used for this eye tracking feature prediction task.

4.1.2. Linear models comparison

For each implemented model, we used AICc values to determine which one was the best fit for the data. On both corpora, we notice that the best predictor of eye-tracking features is BL-sur-cos,

TABLE 4 Average AICc, and AICc for TRT, FFD, and NF on Provo with SGNS vectors.

	Avg		TRT		FFD		NF	
Model	AICc	Delta	AICc	Delta	AICc	Delta	AICc	Delta
BL-sur-cos	279	0	1,309	0	288	0	−762	0
BL-sur	391	112	1,436	127	441	153	−704	58
BL-cos	437	158	1,468	159	406	118	−594	168
BL	619	340	1,683	374	643	354	−470	292

TABLE 5 Average AICc, and AICc for TRT, FFD, and NF on Provo with BERT vectors.

	Avg		TRT		FFD		NF	
Model	AICc	Delta	AICc	Delta	AICc	Delta	AICc	Delta
BL-sur-cos	67	0	1,081	0	−88	0	−791	0
BL-cos	196	129	1,216	135	−0.26	87	−627	165
BL-sur-Spearman	385	318	1,429	348	434	521	−707	85
BL-sur	391	324	1,436	355	441	529	−704	88
BL-Spearman	529	462	1,674	593	633	721	−474	315
BL	619	552	1,683	602	643	730	−470	321

including the interactions between baseline features, but with no interactions between cosine and surprisal. The fact that the regression model using both surprisal and cosine consistently performs better than the ones using only one of the two is strong evidence that they are both explanatory factors of reading times. Furthermore, while comparing BL-cos-sur with SGNS embeddings, and BL-cos-sur with BERT embeddings, it is possible to notice how the usage of the latter set of vectors improves the model (AICc values on GECO: 60,286 with SGNS-59,566 with BERT; AICc values on Provo: 279 with SGNS-67 with BERT).

Looking at the *p*-values of the regression features of our BL-sur-cos model, we observe that both cosine similarity and surprisal are statistically highly significant at $p < 0.001$ (for a complete analysis of regression features significance scores see Appendix 1). Although the combination of both cosine similarity and surprisal is the best performing model on both corpora, it is useful to focus also on the performances of BL-cos, and BL-sur while employing different vector models for BL-cos, to get further insights on the different contributions of surprisal and cosine similarity. We performed nested model comparisons with the R *anova* function using BL-sur-cos and three partial models: one excluding the cosine similarity (BL-sur), and the other two excluding surprisal (BL-cos with BERT vectors and BL-cos with SGNS vectors), in order to check whether the two features make independent contributions. We obtained strongly significant *p*-values ($p < 0.001$) on both corpora, regardless of vector type and for all the eye-tracking features, indicating that both semantic relatedness and surprisal provide an independent and significant contribution.

Focusing now on BL-cos and BL-sur, the performance on **GECO** is reported in Tables 2, 3. BL-cos with BERT vectors: Delta cosine similarity is 585, Delta surprisal is 926 (surprisal: +341) (Table 3); BL-cos with SGNS vectors: Delta surprisal is 206, Delta cosine similarity is 696 (surprisal: −490) (Table 2); On **Provo** instead BL-cos with BERT vectors: Delta cosine similarity is 129, Delta surprisal is 324 (surprisal: +195) (Table 5); BL-cos with SGNS vectors: Delta surprisal is 112, Delta cosine is 158 (surprisal: −46) (Table 4). This first analysis shows that BL-cos and BL-sur have *quantitatively* similar behavior, suggesting that cosine and surprisal help to predict eye-tracking values to the same extent. A difference in the salience of the two features is instead highlighted by the Part-of-Speech analysis (see the related subsection below).

It is also clear that models using SGNS vectors have poorer performances than the ones relying on BERT. Not only, as already mentioned, the usage of BERT embeddings improves the performances of the BL-cos-sur model, but while comparing the BL-cos models and the BL-sur model, the first shows better performances than the latter only when BERT vectors are involved. This difference in the capability of BL-cos models in predicting eye-tracking features suggests that the findings in Frank (2017) might be influenced by the specific type of embedding model used for the experiments (SGNS).

Once confirmed that the model including both surprisal and cosine similarity is the one performing better, we performed further analysis focused on BL, BL-sur, and BL-cos only, in order to understand the individual contribution of the two computational metrics.

4.1.3. Error analysis

In order to have a more fine-grained view of the performance differences between models BL-cos and BL-sur, we also analyzed the correlation between the Mean Absolute Error (MAE) of the models and word-level features. We tested the following features: target and previous word length, target and previous word frequency, target word length, target word position, fixation of the previous word (a boolean feature), and the reading complexity of the sentence from the beginning to the target word, which we computed using the Dale-Chall readability formula (Dale and Chall, 1948).

After we averaged the correlations among all the eye-tracking features to be predicted (see Appendix 2) we noticed that almost

TABLE 6 Average MAE on Provo and GECO content and function words from models BL, BL-cos, and BL-sur for the three eye-tracking features and their mean.

Feature	Model	Word type			
		Content		Function	
		Provo	GECO	Provo	GECO
TRT	BL	0.228	0.337	0.290	0.457
	BL-cos	0.217	0.333	0.281	0.457
	BL-sur	0.215	0.330	0.275	0.454
FFD	BL	0.180	0.295	0.246	0.425
	BL-cos	0.159	0.281	0.216	0.422
	BL-sur	0.172	0.289	0.236	0.423
NF	BL	0.178	0.228	0.147	0.187
	BL-cos	0.177	0.228	0.132	0.184
	BL-sur	0.170	0.226	0.140	0.185
Avg	BL	0.195	0.287	0.228	0.356
	BL-cos	**0.185**	**0.281**	**0.210**	0.354
	BL-sur	0.186	0.282	0.217	0.354

The bold formatting indicates the lowest MAE averaged over the 3 eye tracking features.

all the values are negative, suggesting that: (i) longer and more frequent words are easier to be predicted; (ii) words at the beginning of the sentence are harder to predict for our models, plausibly because a wider and richer context benefits both cosine similarity and surprisal; (iii) sentences with higher readability make better predictions possible. Even so, the correlations between MAE and these features are generally low, ranging from 0.002 for previous word length to 0.1 for target word length. However, it is possible to use these values for a comparison between models BL-cos and BL-sur. We notice that surprisal seems to be more sensitive to target word frequency and previous word fixation if compared to cosine similarity, while the latter shows slightly higher correlations with target word length and position within the sentence.

4.1.4. POS analysis

Both GECO and Provo provide information regarding the part of speech (POS) of each word in the corpora. We used this information to check the performances of BL-cos and BL-sur on different POS. We first checked the average MAE of BL, BL-cos, and BL-sur for function words (pronouns, conjunctions, determiners, numeral, existential there's, prepositions, interjections) and content words (nouns, verbs, adverbs, adjectives) for each eye-tracking feature (Table 6). Then for a more detailed analysis, we ranked the words following the MAE values, and finally, we focused on the 10, 100, 500, and 1,000 words with the highest MAE.

We found that for all three models function words are harder to be predicted than content words, especially coordinating conjunctions and pronouns. Noticeably, previous research had already found that the semantics of function words is difficult to model even for Transformers (Kim et al., 2019), and that fine-tuned multilingual Transformer model struggle the most with the prediction of their fixation metrics (Hollenstein et al., 2022b). Regarding the performances of BL-cos and BL-sur, even if both cosine similarity and surprisal help in lowering the average MAE, if compared to the baseline, cosine similarity employment improves slightly more the performance of the model for both content words and function words.

4.2. Eye-tracking features analysis

While comparing the different models, it was clear that some performance differences were due to the eye-tracking feature the models had to predict. For example, the data showed in the Avg column of Tables 2–5 are mean values computed using the AICc scores of TRT, FFD, and NF, but if we focus on the performances of models BL-cos and BL-sur, depending on the target eye-tracking features, we notice some interesting and substantial differences: on TRT cosine similarity-only and surprisal-only models follow the general tendency we described in Section 4.1 (i.e., surprisal better than cosine similarity when BL-cos makes use of SGNS vectors to compute cosine), but with cosine similarity performing generally slightly better than surprisal; on FFD the model using baseline regression features and cosine similarity only performs consistently better, except when using SGNS on GECO (but not on Provo), while on NF model BL-sur outperforms BL-cos on both corpora, even when using BERT vectors in BL-cos.

In the analysis of the correlations between models MAE and word features, we found that for TRT and FFD, the highest correlation (especially on GECO) is the one between MAE and the word length. Since it is a negative correlation, we can conclude that shorter words induce higher MAE: The shorter the word, the harder for the model to predict the feature value. On the other hand, with NF, word length has the highest, but *positive*, correlation with the MAE, thus suggesting that for this eye-tracking feature shorter words are easier to be predicted. Finally, for all the eye-tracking features on both corpora, word frequency is negatively correlated. As expected, prediction is more difficult for the rarest words.

When we checked the contribution of BL-cos and BL-sur in comparison to the baseline for different parts of speech, we noticed that for FFD cosine similarity generally decreases the MAE, while for TRT surprisal gives a generally higher contribution, except for verbs and adjectives (Tables 7, 8). Regarding NF, cosine similarity lowers the MAE for function words, while surprisal has a major impact on content words. However, for the NF feature content words are less easily predicted.

We surmise that the different performances of BL-sur and BL-cos in predicting these three eye-tracking features might be explained by taking into account the reading process stage each feature is related to. On one hand, since FFD is typically associated with early stages of reading, such as lexical information process, it is not surprising that the model relying on semantic relatedness between the context and the target word performs better. On the other hand,

TABLE 7 Average MAE on Provo content words.

Model	TRT				FFD				NF			
	N	RB	V	J	N	RB	V	J	N	RB	V	J
BL	0.243	0.242	0.210	0.209	0.190	0.184	0.171	0.166	0.195	0.180	0.152	0.180
BL-cos	0.231	0.243	**0.198**	**0.199**	**0.178**	0.184	**0.157**	**0.154**	0.192	0.178	**0.149**	0.179
BL-sur	**0.228**	**0.221**	0.200	0.205	0.181	**0.176**	0.163	0.163	**0.183**	**0.171**	0.150	**0.178**

N, nouns; RB, adverbs; V, verbs; J, adjectives-for the three eye-tracking features. The bold formatting indicates the values with the lowest MAE of each POS within each eye-tracking feature.

TABLE 8 Average MAE on GECO content words.

Model	TRT				FFD				NF			
	N	RB	V	J	N	RB	V	J	N	RB	V	J
BL	0.335	0.365	0.334	0.309	0.289	0.322	0.294	0.273	0.238	0.226	0.217	0.242
BL-cos	0.328	0.367	0.332	0.301	0.280	0.323	0.292	0.264	0.237	0.226	0.217	0.241
BL-sur	**0.323**	**0.360**	0.332	**0.299**	0.280	**0.320**	0.292	**0.262**	**0.234**	**0.225**	**0.216**	**0.240**

N, nouns; RB, adverbs; V, verbs; J, adjectives-for the three eye-tracking features. The bold formatting indicates the values with the lowest MAE of each POS within each eye-tracking feature.

the performances of BL-cos and BL-sur on TRT and NF, features that reflect later stages of the reading process, including information-structural integration, may suggest that predictability is a key factor in handling syntagmatic relations and integrating semantic and syntactic information.

5. Conclusion

In this paper, we implemented four different kinds of regression models to predict three eye-tracking features of two corpora collecting eye movements data, with the aim of investigating the role and interplay between distributional measures of target-context semantic relatedness, and target surprisal, as computed with a state-of-the-art neural language model. The main research question was whether semantic relatedness is indeed made redundant by surprisal, as argued by Frank (2017), or instead plays an independent role in explaining eye-tracking data. The models include: (i) a baseline with word-level features, (ii) the same baseline with cosine similarity, (iii) the baseline with surprisal, iv) the baseline with both cosine similarity and surprisal.

Our results show that the complete model systematically outperforms the others for every eye-tracking feature and that both semantic relatedness and surprisal benefit the prediction of eye-tracking features, given the performance drop while factoring one of them out. Surprisal and distributional semantic relatedness clearly overlap, especially since the latter is nowadays commonly computed using word embeddings produced by DSMs trained with a prediction objective, like the one that surprisal formalizes. Yet, they capture different linguistic dimensions. Surprisal models the *syntagmatic* predictability of a word, given the preceding ones. On the other hand, both static and contextual DSMs use prediction as a distributional signal to form internal representations of lexical meaning that capture information more directly pertaining to the *paradigmatic* dimension, such as belonging to the same semantic classes and domains or sharing similar features. For instance, the words *pie* and *cake* are paradigmatically related because they share several salient attributes, such as being edible, sweet, etc. (Chersoni et al., 2021) showed that word embeddings encode a vast range of linguistically and cognitively relevant semantic features. Therefore, the results of our analyzes suggest that, despite their overlap, corpus-based semantic relatedness and surprisal capture different dimensions that play an autonomous role during reading. While surprisal reflects how predictable the target word is from the previous context, semantic relatedness models how coherent the meaning of the target is with respect to the context one (e.g., they belong to the same semantic field or describe a prototypical situation). Frank and Willems (2017) found that syntagmatic surprisal and paradigmatic semantic relatedness can have neurally distinguishable effects during language comprehension. Our analyzes show that their independent effect can be detected in eye-tracking data too.

We also analyzed whether the relatedness and surprisal have a differential effect depending on the target part-of-speech. Comparing the average MAE of our models, we noticed that surprisal mainly helps to improve the model's performances on content words, while the contribution of semantic relatedness includes function words as well. Finally, we investigated whether the interplay between surprisal and relatedness is affected by the type of word embeddings used to compute the latter, in particular considering the difference between static DSMs (SGNS) and contextual ones (BERT). The experiments show that when using BERT vectors, which are inherently able to account for context-dependent meaning shifts and carry out an implicit form of word-sense disambiguation, the model **BL-cos** performs better than **BL-sur**, while static vectors make the latter outrank the model using semantic relatedness only. Overall, our findings suggest that the kind of word embedding employed for computing vector distances has a significant impact, which may explain the differences from the findings by Frank (2017).

The present work admittedly has some limitations. For example, we employed and compared a restricted pool of language models and word embedding models, and a possible future direction could be testing other, more recent models (e.g., XLNet Yang et al. 2019, among others, RoBERTa Liu et al., 2019), or different static embedding models (e.g., GloVe Pennington et al., 2014, FastText Bojanowski et al., 2017). A particularly interesting issue, raised by some recent works, is the relationship between the size of a language model and its capacity to model human behavioral data (Oh and Schuler, 2022; Shain et al., 2022). In particular, Oh and Schuler

(2022) found that larger language models are worse at predicting human reading times: larger models tend to be less surprised by open-class words because they have been trained on many more word sequences than those available to humans. Moreover, phenomena of inverse scaling have also been reported for language modeling of negations (Jang et al., 2022) and quantifiers (Kalouli et al., 2022; Michaelov and Bergen, 2022). It might be worth testing whether this increasing lack of alignment with human performance as scale increases can be observed also at the level of similarity estimation with the embeddings, or it is an effect limited to language model predictions. With this purpose, it could be interesting to compare embedding models of different size with BERT, and see if there are differences in modeling open class vs. function words.

Another limitation is due to the fact that we used English materials only, and this leaves open the question whether our results would apply to other languages. An interesting research path to pursue is to compare models with cosine similarity and surprisal using multilingual data. In fact, we plan to extend our analyzes to the recently-published MECO corpus (Siegelman et al., 2022), which provides eye-tracking data on comparable texts for 13 different languages.

Finally, if the importance and independence of surprisal and semantic relatedness are clear, given the results shown in the present paper, a preliminary feature importance analysis using a random forest regression model (see Appendix 3) revealed how target and previous word lengths are the features with the higher impact, and most importantly, surprisal systematically seems to have a larger effect on the model compared to cosine similarity. These preliminary results suggest one further possible research direction: the employment and comparison of different models and a consequent feature importance analysis, in order to find even more generalizable insights regarding the role of semantic relatedness and predictability in the reading process.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AL and EC contributed to the conception and design of the study. LS was responsible for the coding part, the data analysis, and the creation of the first draft of the manuscript. AL, EC, and LS contributed equally to the final form of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This project was supported by the COnversational BRAins (CoBra) European Training Network (H-ZG9X). EC was supported by the Startup Fund (1-BD8S) by the Hong Kong Polytechnic University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1112365/full#supplementary-material

References

Aurnhammer, C., and Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*. 134, 107198 doi: 10.1016/j.neuropsychologia.2019.107198

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv*:1409.0473.

Balota, D. A., and Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *J. Exp. Psychol.* 10, 340. doi: 10.1037/0096-1523.10.3.340

Baroni, M., and Lenci, A. (2010). Distributional memory: a general framework for corpus-based semantics. *Comput. Linguist*. 36, 673–721. doi: 10.1162/coli_a_00016

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Computat. Linguist*. 5, 135–146. doi: 10.1162/tacl_a_00051

Bommasani, R., Davis, K., and Cardie, C. (2020). "Interpreting pretrained contextualized representations via reductions to static embeddings," in *Proceedings of ACL*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems, Vol. 33*, 1877–1901.

Bullinaria, J. A., and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behav. Res. Methods* 44, 890–907. doi: 10.3758/s13428-011-0183-8

Chersoni, E., Santus, E., Huang, C.-R., and Lenci, A. (2021). Decoding word embeddings with brain-based semantic features. *Comput. Linguist*. 47, 663–698. doi: 10.1162/coli_a_00412

Chronis, G., and Erk, K. (2020). "When is a bishop not like a rook? When it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships," in *Proceedings of CONLL*.

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

Dale, E., and Chall, J. S. (1948). A formula for predicting readability: instructions. *Educ. Res. Bull*. 27, 37–54.

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL* (Minneapolis, MN).

Ehrlich, S. E., and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav*. 20, 641–665. doi: 10.1016/S0022-5371(81)90220-6

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations* (Ph.D. thesis). University of Stuttgart.

Fossum, V., and Levy, R. (2012). "Sequential vs. hierarchical syntactic models of human incremental sentence processing," in *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics* (Montreal, QC).

Frank, S. L. (2017). "Word embedding distance does not predict word reading time," in *Proceedings of CogSci* (London).

Frank, S. L., and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychol. Sci*. 22, 829–834. doi: 10.1177/0956797611409589

Frank, S. L., and Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Lang. Cogn. Neurosci*. 32, 1192–1203. doi: 10.1080/23273798.2017.1323109

Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol*. 14, 178–210. doi: 10.1016/0010-0285(82)90008-1

Goodkind, A., and Bicknell, K. (2018). "Predictive power of word surprisal for reading times is a linear function of language model quality," in *Proceedings of the LSA Workshop on Cognitive Modeling and Computational Linguistics* (Salt Lake City, UT).

Goodkind, A., and Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv preprint* arXiv:2103.04469. doi: 10.48550/arXiv.2103.04469

Gordon, P. C., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: evidence from eye tracking during reading. *J. Exp. Psychol. Learn. Mem. Cogn*. 32, 1304. doi: 10.1037/0278-7393.32.6.1304

Hale, J. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of NAACL* (Pittsburgh, PA).

Hale, J. (2016). Information-theoretical complexity metrics. *Lang. Linguist. Compass*. 10, 397–412. doi: 10.1111/lnc3.12196

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., and Frank, R. (2020). "Probabilistic predictions of people perusing: evaluating metrics of language model performance for psycholinguistic modeling," in *Proceedings of the EMNLP Workshop on Cognitive Modeling and Computational Linguistics*.

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., and Kuchinke, L. (2021). Language models explain word reading times better than empirical predictability. *Fronti. Artif. Intell*. 4, 730570. doi: 10.3389/frai.2021.730570

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., and Santus, E. (2022a). "CMCL 2022 shared task on multilingual and crosslingual prediction of human reading behavior," in *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics* (Dublin).

Hollenstein, N., Gonzalez-Dios, I., Beinborn, L., and Jaeger, L. (2022b). "Patterns of text readability in human and predicted eye movements," in *Proceedings of the AACL Workshop on Cognitive Aspects of the Lexicon* (Taipei).

Hollenstein, N., Pirovano, F., Zhang, C., Jäger, L., and Beinborn, L. (2021). "Multilingual language models predict human reading behavior," in *Proceedings of NAACL*.

Inhoff, A. W. (1984). Two stages of word processing during eye fixations in the reading of prose. *J. Verbal Learn. Verbal Behav*. 23, 612–624. doi: 10.1016/S0022-5371(84)90382-7

Inhoff, A. W., and Radach, R. (1998). "Definition and computation of oculomotor measures in the study of cognitive processes," in *Eye Guidance in Reading and Scene Perception*, 29–53.

Jang, J., Ye, S., and Seo, M. (2022). Can large language models truly understand prompts? a case study with negated prompts. *arXiv preprint* arXiv:2209.12711. doi: 10.48550/arXiv.2209.12711

Just, M. A., and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychol. Rev*. 87, 329–354. doi: 10.1037/0033-295X.87.4.329

Kalouli, A.-L., Sevastjanova, R., Beck, C., and Romero, M. (2022). "Negation, coordination, and quantifiers in contextualized language models," in *Proceedings of COLING* (Gyeongju).

Kennedy, A., Hill, R., and Pynte, J. (2003). "The dundee corpus," in *Proceedings of the European Conference on Eye Movement* (Dundee).

Kim, N., Patel, R., Poliak, A., Wang, A., Xia, P., McCoy, R. T., et al. (2019). "Probing what different NLP tasks teach machines about function word comprehension," in *Proceedings of *SEM* (Minneapolis, MN).

Kovaleva, O., Kulshreshtha, S., Rogers, A., and Rumshisky, A. (2021). "BERT busters: outlier dimensions that disrupt transformers," in *Findings of ACL*.

Landauer, T. K., and Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev*. 104, 211. doi: 10.1037/0033-295X.104.2.211

Lenci, A. (2018). Distributional models of word meaning. *Ann. Rev. Linguist*. 4, 151–171. doi: 10.1146/annurev-linguistics-030514-125254

Lenci, A., and Sahlgren, M. (2023). *Distributional Semantics*. Cambridge: Cambridge University Press.

Lenci, A., Sahlgren, M., Jeuniaux, P., Gyllensten, A. C., and Miliani, M. (2022). A comprehensive comparative evaluation and analysis of distributional semantic models. *Lang. Resour. Evaluat*. 56, 1269–1313. doi: 10.1007/s10579-021-09575-z

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692

Luke, S. G., and Christianson, K. (2018). The provo corpus: a large eye-tracking corpus with predictability norms. *Behav. Res. Methods* 50, 826–833. doi: 10.3758/s13428-017-0908-4

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instruments Comput*. 28, 203–208. doi: 10.3758/BF03204766

Merkx, D., and Frank, S. L. (2021). "Human sentence processing: recurrence or attention?" in *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Michaelov, J. A., and Bergen, B. K. (2022). 'Rarely' a problem? language models exhibit inverse scaling in their predictions following 'few'-type quantifiers. *arXiv preprint* arXiv:2212.08700. doi: 10.48550/arXiv.2212.08700

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781

Mitchell, J., Lapata, M., Demberg, V., and Keller, F. (2010). "Syntactic and semantic factors in processing difficulty: an integrated measure," in *Proceedings of ACL* (Uppsala).

Monsalve, I. F., Frank, S. L., and Vigliocco, G. (2012). "Lexical surprisal as a general predictor of reading time," in *Proceedings of EACL* (Avignon).

Oh, B.-D., and Schuler, W. (2022). "Entropy-and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal," in *Proceedings of EMNLP* (Abu Dhabi).

Padó, S., and Lapata, M. (2007). Dependency-based construction of semantic space models. *Comput. Linguist*. 33, 161–199. doi: 10.1162/coli.2007.33.2.161

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: global vectors for word representation," in *Proceedings of EMNLP* (Doha) .

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). "Deep contextualized word representations," in *Proceedings of NAACL* (New Orleans, LA).

Pollatsek, A., Juhasz, B. J., Reichle, E. D., Machacek, D., and Rayner, K. (2008). Immediate and delayed effects of word frequency and word length on eye movements in reading: a reversed delayed effect of word length. *J. Exp. Psychol*. 34, 726. doi: 10.1037/0096-1523.34.3.726

Pynte, J., New, B., and Kennedy, A. (2008). On-line contextual influences during reading normal text: a multiple-regression analysis. *Vision Res*. 48, 2172–2183. doi: 10.1016/j.visres.2008.02.004

Radach, R., and Kennedy, A. (2013). Eye movements in reading: some theoretical context. *Q. J. Exp. Psychol*. 66, 429–452. doi: 10.1080/17470218.2012.750676

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). *Improving Language Understanding by Generative Pre-training*. Open-AI Blog.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. In Open-AI Blog.

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

Rayner, K., and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Mem. Cogn*. 14, 191–201. doi: 10.3758/BF03197692

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., and Clifton Jr, C. (1989). Eye movements and on-line language comprehension processes. *Lang. Cogn. Process*. 4, SI21-SI49. doi: 10.1080/01690968908406362

Remington, R. W., Burt, J. S., and Becker, S. I. (2018). The curious case of spillover: does it tell us much about saccade timing in reading? *Attent. Percept. Psychophys*. 80, 1683–1690. doi: 10.3758/s13414-018-1544-5

Rodriguez, M. A., and Merlo, P. (2020). "Word associations and the distance properties of context-aware word embeddings," in *Proceedings of CONLL*.

Sahlgren, M. (2008). The distributional hypothesis. *Italian J. Comput. Linguist*. 20, 33–53.

Salicchi, L., Lenci, A., and Chersoni, E. (2021). "Looking for a role for word embeddings in eye-tracking features prediction: does semantic similarity help?" in *Proceedings of IWCS* (Dublin).

Salicchi, L., Xiang, R., and Hsu, Y.-Y. (2022). "HkAmsters at CMCL 2022 shared task: predicting eye-tracking data from a gradient boosting framework with linguistic features," in *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Santus, E., Chersoni, E., Lenci, A., Huang, C.-R., and Blache,P. (2016a). "Testing APsyn against Vector cosine on similarity estimation," in *Proceedings of PACLIC* (Seoul).

Santus, E., Chiu, T.-S., Lu, Q., Lenci, A., and Huang, C.-R. (2016b). "What a Nerd! beating students and vector cosine in the ESL and TOEFL datasets," in Proceedings *of LREC* (Portorož).

Santus, E., Wang, H., Chersoni, E., and Zhang, Y. (2018). "A rank-based similarity metric for word embeddings," in *Proceedings of ACL* (Melbourne, VIC).

Sayeed, A., Shkadzko, P., and Demberg, V. (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. *Italian J. Comput. Linguist.* 1, 31–46. doi: 10.4000/ijcol.298

Shain, C., Meister, C., Pimentel, T., Cotterell, R., and Levy, R. P. (2022). Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv*. doi: 10.31234/osf.io/4hyna

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., et al. (2022). Expanding horizons of cross-linguistic research on reading: the multilingual eye-movement corpus (meco). *Behav. Res. Methods* 2022, 1–21. doi: 10.3758/s13428-021-01772-6

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Timkey, W., and van Schijndel, M. (2021). "All bark and no bite: rogue dimensions in transformer language models obscure representational quality," in *Proceedings of EMNLP* (Punta Cana).

Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., and Morris, R. K. (2000). Priming in sentence processing: intralexical spreading activation, schemas, and situation models. *J. Psycholinguist. Res.* 29, 581–595. doi: 10.1023/A:10264162 25168

Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188. doi: 10.1613/jair.2934

van Schijndel, M., and Linzen, T. (2018). "A neural model of adaptation in reading," in *Proceedings of EMNLP* (Brussels).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2020). "On the predictive power of neural language models for human real-time comprehension behavior," in *Proceedings of CogSci*.

Wlotko, E. W., and Federmeier, K. D. (2015). Time for prediction? the effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex* 68, 20–32. doi: 10.1016/j.cortex.2015.03.014

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). "XlNet: generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems, Vol. 32* (Vancouver, BC).

Zhelezniak, V., Savkov, A., Shen, A., and Hammerla, N. (2019). "Correlation coefficients and semantic textual similarity," in *Proceedings of NAACL* (Minneapolis, MN).

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 06 February 2023
DOI 10.3389/fpsyg.2023.1076379

OPEN ACCESS

EDITED BY
Marijan Palmovic,
University of Zagreb, Croatia

REVIEWED BY
Lieve Macken,
Ghent University, Belgium
Laura Kamandulyte Merfeldiene,
Vytautas Magnus University, Lithuania

*CORRESPONDENCE
Ramunė Kasperė
✉ ramune.kaspere@ktu.lt

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 21 October 2022
ACCEPTED 12 January 2023
PUBLISHED 06 February 2023

CITATION
Kasperė R, Motiejūnienė J, Patašienė I, Patašius M and Horbačauskienė J (2023) Is machine translation a dim technology for its users? An eye tracking study. *Front. Psychol.* 14:1076379. doi: 10.3389/fpsyg.2023.1076379

Is machine translation a dim technology for its users? An eye tracking study

Ramunė Kasperė[1]*, Jurgita Motiejūnienė[1], Irena Patašienė[2], Martynas Patašius[2] and Jolita Horbačauskienė[1]

[1]Faculty of Social Sciences, Arts and Humanities, Kaunas University of Technology, Kaunas, Lithuania, [2]Faculty of Informatics, Kaunas University of Technology, Kaunas, Lithuania

State-of-the-art research shows that the impact of language technologies on public awareness and attitudes toward using machine translation has been changing. As machine translation acceptability is considered to be a multilayered concept, this paper employs criteria of usability, satisfaction and quality as components of acceptability measurement. The study seeks to determine whether there are any differences in the machine-translation acceptability between professional users, i.e., translators and language editors, and non-professional users, i.e., ordinary users of machine translation who use it for non-professional everyday purposes. The main research questions whether non-professional users process raw machine translation output in the same way as professional users and whether there is a difference in the processing of raw machine-translated output between users with different levels of machine-translated text acceptability are analyzed. The results of an eye tracking experiment, measuring fixation time, dwell time and glance count, indicate a difference between professional and non-professional users' cognitive processing and acceptability of machine translation output: translators and language editors spend more time overall reading the machine-translated texts, possibly because of their deeper critical awareness as well as professional attitude toward the text. In terms of acceptability overall, professional translators critically assess machine translation on all components of which confirms the findings of previous similar research. However, the study draws attention to non-professional users' lower awareness regarding machine translation quality. The study was conducted within a research project that received funding from the Research Council of Lithuania (LMTLT, agreement No S-MOD-21-2), seeking to explore and evaluate the impact on society of machine translation technological solutions.

KEYWORDS

machine translation, acceptability, usability, quality, satisfaction, end-users, professional translators

1. Introduction

Neural machine translation is more and more frequently used in the translation and localization market. Following the AI Index Report, artificial intelligence has allowed improving machine translation in certain language pairs almost to human quality (Perrault et al., 2019). According to some scores, "[t]he fastest improvement was for Chinese-to-English, followed by English-to-German and Russian-to-English" (Perrault et al., 2019). However, the performance varies between different language pairs and that depends on language pair popularity, which "defines how much investment goes into data acquisition" (Perrault et al., 2019).

For these reasons, researchers and research administrators have recently been paying attention to the effects that artificial intelligence and developed technologies bring about on the translation industry, translator's profession, career and daily tasks, as well as training and

skills needed, but also in terms of the perceptions within society. In this perspective, some important research papers have been published in the past few years where translation scholars have concluded that, for example, artificial intelligence-powered machine translation and other language related technologies have fundamentally changed public awareness and attitudes toward multilingual communication (Vieira et al., 2021). Such technologies are now increasingly being used to overcome language barriers not only in situations of personal use but also in high-risk environments, such as health care systems, courts, police and so on. The availability and impact of machine translation accessibility and impact on society, including the importance of full participation of various social groups in communication processes, are being analyzed and evaluated (Vieira et al., 2021). On the other hand, public awareness of the capabilities as well as the quality of machine translation is identified as insufficient (Kasperė and Motiejūnienė, 2021).

Although machine translation is breaking down language barriers, and its accuracy and efficiency are getting closer to human-level translation, human effort is needed to reduce the negative impact of machine translation in society (Hoi, 2020). The communication processes supported by machine translation can be of high quality if the process participants are aware of the quality shortcomings (Yasuoka and Bjorn, 2011). Studies have also found that machine translation can help reduce the exclusion of ethnic minorities in a wide variety of fields (Taylor et al., 2015).

There is a plethora of research on the quality of machine translation and use of post editing (see Ueffing, 2018; Ortega et al., 2019; Vardaro et al., 2019; Nurminen and Koponen, 2020; Rossi and Carré, 2022, to mention but a few). The benefits of machine translation post editing in different language pairs have been acknowledged in multiple studies employing a diversity of research designs (see Carl et al., 2011, 2015; Moorkens, 2018; Stasimioti and Sosoni, 2021). Studies have also addressed the issue of machine translation acceptability (see Castilho, 2016; Castilho and O'Brien, 2018; Rivera-Trigueros, 2021; Taivalkoski-Shilov et al., 2022). However, the attitudes and perceptions of translation students, novice translators, professional translators and posteditors have been mainly taken into the focus, possibly due to a somewhat easier access to respondents and more convenient research design (see Moorkens and O'Brien, 2015; Rossi and Chevrot, 2019; Ferreira et al., 2021). The acceptability of machine-translated content by non-professional users has not been extensively studied. The ordinary users' perspective is important because of the variety of purposes for which they take machine translation for granted and use it daily (Kasperė and Motiejūnienė, 2021).

The study[1] seeks to investigate the acceptability of raw machine translation texts in Lithuanian, a low-resource language. In this paper, we report the results of an eye tracking experiment with professional translators and non-professional users of machine translation with the focus on acceptability. The inter-group and intra-group comparisons of raw machine-translated text acceptability are made. The research questions are as follows: do non-professional users process raw machine translation output in the same way as professional users? Is there a difference in the processing of raw machine-translated output among non-professional users with different levels of acceptability of machine-translated text? Is there a difference between professional and non-professional users' comprehension of the raw machine-translated output?

2. Literature overview

Machine translation acceptability is a multilayered concept. Criteria of usability, satisfaction and quality have been indicated to be the components of acceptability. Castilho and O'Brien (2018) define acceptability as machine translation output quality in terms of correctness, cohesion and coherence from the reader's perspective. Even if the text contains errors, it does not mean that it is considered unacceptable. If the needs of the readers are satisfied, the text has served its mission (Castilho and O'Brien, 2016, 2018). In order to measure acceptability, Castilho (2016) defines the three criteria. Usability is related to efficiency and effectiveness of the text and may be measured by exerted cognitive effort; satisfaction, which is understood as a user's positive attitude toward the translated text, may be measured through web surveys, post-task satisfaction questionnaires or moderators' ratings; and quality is defined by fluency, adequacy, syntax and grammar, and style in translated content or as text easeability, readability, etc. (Castilho, 2016). For the purposes of this research, acceptability is understood as a notion combining satisfaction, usability and quality as assumed by the ordinary readers of the text who have no linguistic background or related, e.g., translator, training.

Research employing eye tracking methodology is common in Translation Studies (Carl et al., 2011; Castilho, 2016; Daems et al., 2017; Moorkens, 2018; Vardaro et al., 2019; Ferreira et al., 2021; Stasimioti and Sosoni, 2021). Among the existing body of scientific literature on the acceptability criteria of machine translation, of particular mention are those published papers that employ eye tracking experiments. Since acceptability is a vague notion representing quite subjective understanding and judgement, eye tracking studies present relevant insights into the readers' cognitive processing of the (machine-translated) text they are reading. The research reveals that the required cognitive load is generally to a greater or lesser extent higher in cases where machine translation is provided in comparison with human-translated or post-edited text.

Jakobsen and Jensen (2008) report the results of a translation process study, focusing on the differences between the reading of a text with the aim of understanding its meaning and reading the same text (or a very similar text) with the expectation of having to translate it next. The authors recorded eye movements of six translation students and six professional translators who were asked to perform four tasks at the speed at which they normally work, namely read a text for comprehension, read a text in preparation for translating it later on, read a text while performing its oral translation and read a text while typing a written translation. The researchers compared task duration, total number of fixations, total gaze time and average duration of individual fixations for each task and found out that the purpose of reading had a clear effect on eye movements and gaze duration. Overall, the increases in the number of fixations from the first to the last task of the experiment were statistically significant (Jakobsen and Jensen, 2008).

1 Approval to conduct this study was obtained from the Research Ethics Committee of Kaunas University of Technology (No. M6-2021-04 as of 2021-06-16).

In a study by Guerberof Arenas et al. (2021), researching the effect of different translation modalities on users through an eye tracking experiment, 79 end users' (Japanese, German, Spanish, English) experiences with published translated, machine-translated and published English versions were compared. The authors focused on the number of successful tasks performed by end users, the time necessary for performing successful tasks in different translation modalities, the satisfaction level of end users in relation to different translation modalities and the amount of cognitive effort necessary for carrying out tasks in different translation modalities (Guerberof Arenas et al., 2021). They measured usability, i.e., effectiveness (by asking participants to perform some tasks), efficiency (by measuring the time to complete the tasks and by measuring cognitive effort using an eye tracker) and satisfaction. The authors came to the conclusion that the effectiveness variable was not found to be significantly different when the subjects read the published translated version, a machine-translated version and the published English version of the text although efficiency and satisfaction were significantly different, especially for less experienced participants. The results of the eye tracking experiment revealed that end users' cognitive load was higher for machine-translated and human translated versions than for the English original. The findings also indicated that the language and the translation modality played a significant role in the usability, regardless of whether end users finished the given tasks and even if they were unaware that MT was used (Guerberof Arenas et al., 2021).

In a study by Hu et al. (2020), an eye tracking experiment involving 66 Chinese participants with low proficiency in English who also had to fill in questionnaires on comprehension testing and attitudes showed that the quality of raw machine-translated output was considered somewhat lower, but almost as good as that of a post-edited machine-translated output, although the research design involved non-professional post-editing of machine-translated text.

Some earlier user-centered studies where raw machine translation was analyzed *via* eye tracking, screen recording experiments and post-task questionnaires determined a lower usability of machine-translated instructions in comparison with post-edited output (Castilho et al., 2014; Doherty and O'Brien, 2014; Doherty, 2016).

In a study of non-professional users where acceptability of a machine-translated text from English into Lithuanian was tested, an eye tracking experiment revealed that the cognitive processing was greater, i.e., required a longer gaze time and fixation count, on machine translation errors in comparison with correct segments of text (Kasperavičienė et al., 2020). The machine-translated segments with errors required more attention and cognitive effort from the readers, but the results regarding overall acceptability of the raw machine-translated text obtained *via* a post-task survey did not correlate with the readers' gaze time spent on segments with errors.

Literary texts have also received some attention with regard to the differences between human and machine translations from English into Dutch as perceived by end users. Colman et al. (2021) employed eye tracking to analyze end users' reading process and determine the extent to which machine translation impacts the reading process. An increased number of eye fixations and increased gaze duration while reading machine translation segments was found in comparison with human translation (Colman et al., 2021).

Although scarce, there is some research, based on research designs employing methodologies other than eye tracking, determining how the acceptability of machine-translated texts in various languages is perceived by non-professionals or low proficiency future professionals. The broad public uses machine translation for many reasons and purposes and they may not fully understand or consider how machine translation really works and what quality it generates. In a study of 400 surveyed participants, acceptability of the text that had been machine translated from English to Lithuanian was found to be affected by such factors as age and education. The less educated and senior participants were more prone to consider machine translation reliable and satisfactory (Kasperė et al., 2021).

In a study by Rossetti et al. (2020), 61 participants were surveyed in order to get insight into the "impact of machine translation and postediting awareness" on comprehension and trust. The participants were asked to read and evaluate crisis messages in English and Italian using ratings and open-ended questions on comprehensibility and trust. The authors found insignificant differences in the end users' comprehension and trust between raw machine-translated and post-edited text (Rossetti et al., 2020). However, users with low proficiency of English were more positive toward raw machine-translated text in terms of its comprehension and trust (Rossetti et al., 2020).

In another study with translation agencies, professional translators and clients/users of professional translation, the level of user awareness of machine translation was studied through surveys (García, 2010). Acceptability and evaluation of machine translation from Chinese into English was at the focus. The researcher found out that <5% of professional translators considered the quality of machine translation very high. The translation agencies expressed a very similar view on machine translation to that of the translators. The clients/users of professional translations (about 30%) who were aware of and requested machine translation had an intermediate or positive assessment of the quality of machine translation (García, 2010).

As the amount of content to be translated is growing, there is a demand to cut the cost of translation orders, which leads to a growing need for research and testing how translators work with machine translation (Moorkens and O'Brien, 2015) and the newly-arising need to learn how the end users are aware of, perceive, use and accept machine-translated content.

3. Materials and methods

Machine translation quality overall can be assessed in various ways: by applying automatic quality estimation metrics, by carrying out an error analysis by professionals/experts, employing cognitive experimental methods with human experts or professionals or semi-experts or non-experts, determining acceptability of the output of non-experts/non-professionals/amateur users, *via* qualitative methods, etc. Recently, cognitive experimental methods for machine translation quality assessment have been increasingly employed, e.g., eye tracking, key logging, screen recording, post-performance (retrospective) interviews, think-aloud protocols, etc. In an eye tracking experiment, fixation count and time, gaze time, saccades, pupil dilation, and other variables can be measured, although researchers have determined that, for example, pupil dilation may not adequately reflect cognitive effort involved or provide valid and reliable data. To test the validity of the data, cognitive translation researchers have employed complementary methods, including other experimental methods, interviews or surveys. Translation

research studies employing eye tracking have mostly relied on post-performance or retrospective interviews/surveys, and the number of subjects involved in an eye tracking experiment for translation research varies between 2 and 84 (per language). The most common eye movement measures taken into account and described in translation research are fixation time and fixation count (Kasperė and Motiejūnienė, 2021).

3.1. Experiment

For the current study, we used an eye tracking experiment along with a questionnaire in order to ensure the validity of results obtained. Before the experiment, a larger-scale population survey was conducted to find out the purposes, typical circumstances of machine translation use and systems employed non-professional users (Kasperė et al., 2021). In the survey, the respondents were asked to indicate a machine translation tool that they used most often. The reported results of the survey revealed that the absolute majority of the respondents indicated that they used Google Translate as the tool for machine translation (Kasperė et al., 2021). We, therefore, also employed it for the machine translation of the text in the research design of this particular study. Google Translate has over 500 million users per month and over 140 billion words are translated per day (Schuster et al., 2016; Hu et al., 2020). The text chosen for a reading task in the experiment was a recipe of a dish. The motivation behind selecting the text of a recipe for this experiment lies in the findings of the above-mentioned study where the respondents indicated various reasons for using machine translation in their everyday activities, one of the most common being household purposes (Kasperė et al., 2021). The text of a recipe, originally in English, was machine translated using Google Translate to Lithuanian. The translated excerpt given to the subjects as a reading task contained 371 words and was arranged on three slides 13–15 lines each.

In the machine-translated excerpt, we selected areas of interest with errors and areas of interest without errors. According to scientific literature, the perceptual span in western languages is about 13–15 characters to the right of the center of vision, and 3–4 to the left (McConkie and Rayner, 1975; Rayner, 1998). Therefore, all our selected areas of interest (both with and without errors) included 18–20 characters. In the raw translated text prepared for the experiment, 12 distinct errors were selected as areas of interest. Another 12 areas of interest without errors were selected as control. To identify the errors, we used the Multidimensional Quality Metrics, which is a typology of errors developed for assessment of the quality of human translated, machine translated and post-edited texts. This system covers more than 100 error types and can be adapted to all languages (Lommel et al., 2014). Within this classification, the following main types of errors are as indicated: terminology; accuracy (for example, addition, mistranslation, omission, untranslated text, etc.); linguistic conventions (also called fluency in the previous versions of the taxonomy, related to errors in grammar, punctuation, spelling, unintelligible text, etc.), design and markup (errors related to visual presentation of a translated text, such as text formatting, layout); locale conventions (errors related to locale-specific content); style (errors related to inappropriate organizational or language style); and audience appropriateness (for example, errors related to culture-specific reference) (MQM Commitee, 2022). The 12 identified errors fell into two 2 different categories of errors, namely accuracy and linguistic conventions. Accuracy errors were those of mistranslation, untranslated text, omission, and addition. Errors that fell within the linguistic conventions category were those of an incorrect word form (ending) resulting in inappropriate agreement between the words in a phrase.

Eye tracking was performed using a commercial non-invasive eye tracking device SensoMotoric Instruments GmbH Scientific RED-B.6-1524-6150133939 and SMI BeGaze 3.7.42 software for data analysis. For each area of interest (AOI), several eye movement measures were taken into consideration: fixation time (total time of fixations that happened in the AOI), dwell time (total time of fixations and saccades that happened in the AOI), and glance count (the number of times when the gaze entered the AOI).

3.2. Research participants

In total, there were 30 subjects in the experiment: 11 professional translators, language editors and revisers and 19 non-professional users of machine translation, who were of different educational backgrounds, age, occupation. All subjects were native speakers of Lithuanian. Among the non-professional users, 13 had a university degree and 6 had secondary education. The subjects gave consent to participate in the experiment on a voluntary basis. They were informed that the text they were reading was a machine translation. The subjects were also told that they would have to answer questions about the text afterwards filling in a post-task questionnaire. There were 4 reading comprehension questions, all related to the errors in the text, including 2 true/false questions and 2 open questions. The post-task questionnaire also had 9 statements, 3 per each component of acceptability (i.e., quality, usability and satisfaction). The statements could be assessed by the subjects on a 5-point Likert scale, where 1-completely disagree, 2-somewhat disagree, 3-neither agree nor disagree, 4-somewhat agree, and 5-completely agree. In total, in this part of the questionnaire, the subjects of the experiment could accumulate a maximum of 45 points: 15 for quality, 15 for usability and 15 for satisfaction. The questions and the statements provided to the subjects in a post-task questionnaire were presented in their native, i.e., Lithuanian, language.

3.3. Data analysis

IBM SPSS Statistics 27 was used for descriptive and relationship analysis. Descriptive statistics were calculated for quantitative nominal and ordinal data. The relationships between data were investigated using column plots and box plots. Although the convenience sample was used, limiting the usefulness of hypothesis testing, several non-parametric tests (one-sample Kolmogorov-Smirnov test, independent-samples Mann-Whitney *U*-test, independent-samples Moses test of extreme reaction) with a significance level of 0.05 were used to explore what hypotheses would be more promising for further research.

4. Results

The findings of our study demonstrate that the average fixation time on the areas of interest with errors of both groups of the subjects was longer than on the areas of interest without errors, which confirms findings of other studies that errors attract more readers' attention and require more cognitive effort than correct text (see Figure 1). The average fixation time on the areas of interest with errors (in percentage from total time of the trial) was 12.6 vs. 11.7% for professional and non-professional users of machine translation, respectively. On the other hand, professionals also demonstrated a longer average fixation time on areas of interest without errors than non-professionals, i.e., 11.8 vs. 10.4%. The longer average fixation time on both types of areas of interest within the professionals' cohort might be interpreted that professional translators and language specialists who work with texts on a daily basis have different skills and a more pronounced critical look at any text. Such a hypothesis would still have to be tested on a broader scale experiment.

Independent-samples Mann-Whitney *U*-test would indicate that the hypothesis that the fixation time of professionals and non-professionals for AOIs with errors has the same distribution (more precisely, the hypothesis that the probability of fixation time being higher for a random professional than for random non-professional is 0.5) could not be rejected ($p = 0.792$). Still, the independent-samples Moses test of extreme reaction suggests that, while hypothesis about the distributions having the same range could not be rejected, the value of p is much closer to the level of significance ($p = 0.079$). Similar (although weaker) relationship holds for AOIs without errors ($p = 0.670$ and $p = 0.180$).

As Figure 2 shows, while the median of dwell time for AOIs with errors was very similar for professionals (17,851 ms) and non-professionals (18,722 ms), the spread of it was clearly different. That might be assumed to be rather surprising, for, intuitively, one might suppose that professionals are going to be more like each other than non-professionals.

Different eye movement measures, including the fixation time, were also compared in the groups of the subjects who scored high and low in the post-task survey for the questions demonstrating quality and usability of the text and the users' satisfaction with the text.

On all components of acceptability (see Figure 3 for quality, Figure 4 for usability, and Figure 5 for satisfaction), non-professional users scored higher than professionals. The average total quality scores were 6.2632 for non-professionals and 5.3636 for professionals (median 6 vs. 5, respectively). The average total usability scores were 6.6842, i.e., slightly better, for non-professionals compared with professionals, i.e., 6.0000 (median 7 vs. 6, respectively). In terms of the average total satisfaction scores, the non-professionals' scores were much more increased compared with professionals, i.e., 5.7895 vs. 3.5455 (median 6 vs. 3), respectively. This suggests that non-professional users were more positive toward the machine-translated text than professional users, perhaps because professional users are more aware of features of good translation and are able to notice when they are not present.

Independent-samples Mann-Whitney *U*-test would also indicate that the hypothesis that the total satisfaction score of professionals and non-professionals has the same distribution (more precisely, the hypothesis that the probability of this score being higher for a random professional than for a random non-professional is 0.5) can be rejected ($p < 0.001$). On the other hand, the same test does not suggest rejecting the hypotheses that the total usability score and the total quality score of professionals and non-professionals have the same distributions ($p = 0.427$ and $p = 0.381$).

One-sample Kolmogorov-Smirnov test suggests that the hypotheses of total quality score, total usability score and total satisfaction score having normal distribution could be rejected ($p = 0.020$, $p = 0.017$, $p < 0.001$), while the hypothesis that their sum has a normal distribution could not be rejected ($p = 0.200$).

The subjects from the group of non-professional users who thought that the quality was low (having scores lower than average; there were 16 such subjects out of 21) demonstrated a longer average fixation time both for AOIs with errors and without errors (12.2 vs. 10.7%, respectively) than those subjects who thought that the quality was high (10.4 vs. 9.4%, respectively) (see Figure 6).

The same pattern was observed for the usability and satisfaction components. The subjects who thought that the text was barely usable (having scores lower than average; there were 16 such subjects out of 21) showed a longer fixation time result that those who thought that the text was usable (12.0 vs. 11.0% and 10.6 vs. 9.6%, respectively) (see Figure 7).

The non-professional users who were less satisfied with the text (value lower than average; there were 17 such subjects out of 21) demonstrated a longer average fixation time result in comparison with those who were more satisfied with the text (12.0 vs. 10.6%, respectively) (see Figure 8).

All professional translators, language editors and revisers who read the raw machine-translated text provided to them in the experiment thought that the text quality was low, and they scored low on the questions of satisfaction in the post-task questionnaire on acceptability components. Only in terms of usability, the subjects of the professional translators' group were divided into those who thought that the text was usable to some extent (usability higher than average) and those who thought that the text was not usable. The results for average percentage of fixation time of the two groups of professional translators - low scorers and high scorers for usability statements—are shown in Figure 9. The subjects in the group of low scorers for the usability statements demonstrated a shorter average fixation time compared with those who scored higher, 12.4 vs. 14.3% for AOIs with errors and 11.7 vs. 12.7% for AOIs without errors, which also raises questions for further research, discussion and implications.

As Figure 10 shows, total satisfaction scores for non-professionals who looked at AOIs with errors for a shorter period of time than average varied greatly. The higher limit of those scores decreased for non-professionals who looked at such AOIs longer, while the lower limit tended to stay the same. On the other hand, the satisfaction scores for the professionals tended to stay the same, as for non-professionals who paid more attention to the AOIs with errors.

However, the independent-samples Mann-Whitney *U*-test would indicate that the hypothesis that the total satisfaction score of professionals and non-professionals has the same distribution (that the probability of this score being higher for a random professional than for a random non-professional is 0.5) cannot be rejected ($p = 0.157$).

Besides, the subjects' text comprehension was measured *via* a post-task reading comprehension questionnaire, consisting of 4 questions, i.e., 2 true/false questions and 2 open questions.

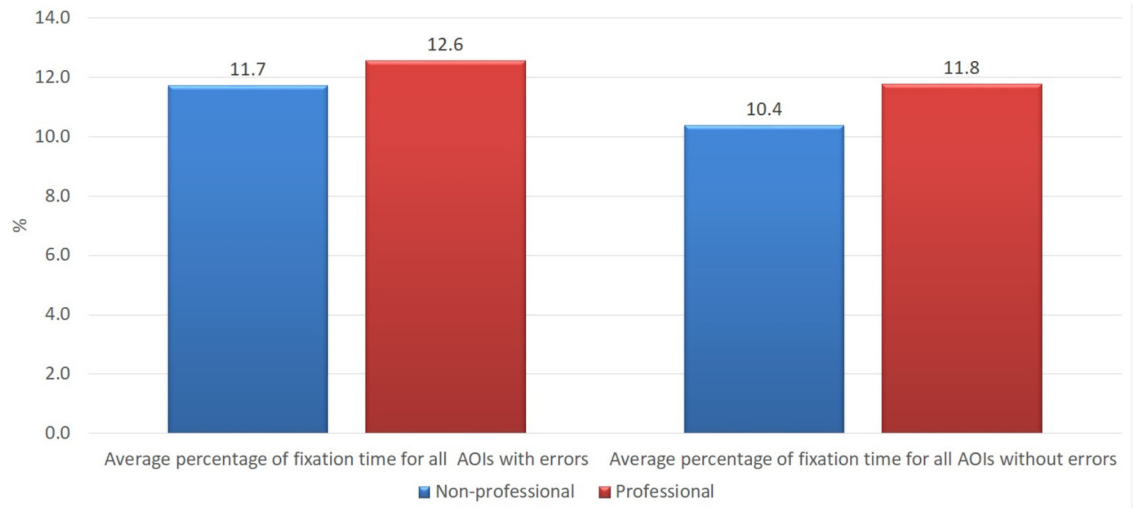


FIGURE 1
Average percentage of fixation time on all areas of interest with errors and without errors in the groups of professional and non-professional users of machine translation.

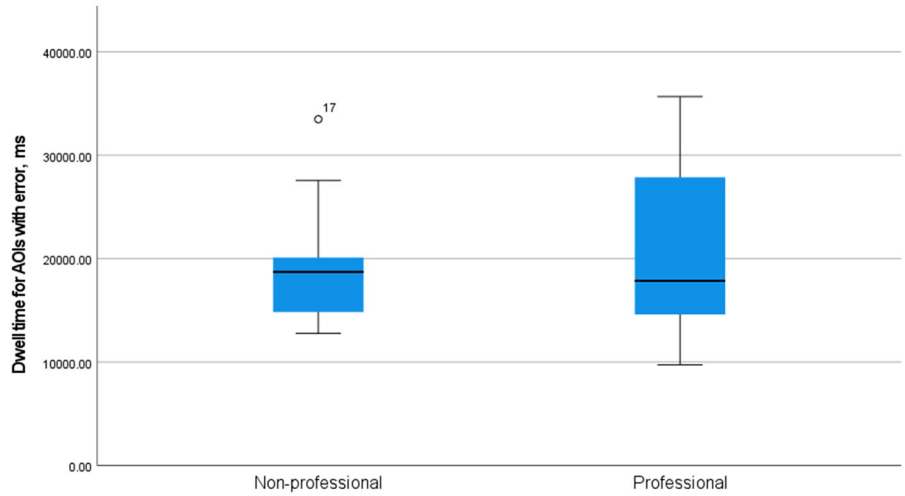


FIGURE 2
A simple boxplot of dwell time on AOIs with errors in the groups of professional and non-professional users of machine translation.

Figure 11 demonstrates how the subjects scored in both groups. The professionals scored better in text comprehension compared with non-professional users (median 2 vs. 3, respectively) (see Figure 11).

Figure 12 shows how fixation times for AOIs with errors correlate with the number of correctly answered questions. It may be seen that the pattern differs between professionals and non-professionals, with professionals having higher spread for more correct answers and non-professionals having higher spread for average number of correct answers. It is also interesting that the median dwell time was mostly the same for non-professionals giving different numbers of correct answers, while the median dwell times for professionals giving the highest and the lowest numbers of correct answers are lower than for professionals who gave a medium number of correct answers. Furthermore, both professionals and non-professionals who gave no correct answers (there were two such professionals and two such non-professionals) had low dwell times (with the maximum lower than the medians of every other group).

Figure 13 shows how glance counts for AOIs with errors correlate with the number of correctly answered questions. The differences between professionals and non-professionals may be observed, with professionals having higher spread and non-professionals having lower spread for the higher number of correct answers. Professionals tended to reach higher glance counts (for each number of correct answers, professionals tended to have a higher median glance count, with the exception of the group of no correct answers, which might have been an outlier). Furthermore, non-professionals who gave no correct answers had high glances counts (with median higher than the medians of every other group of non-professionals). As they also

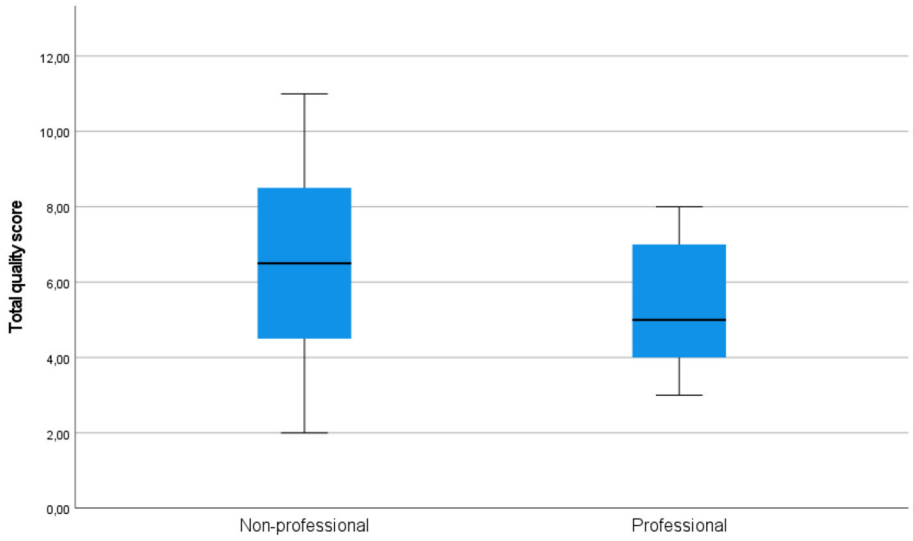


FIGURE 3
Total quality scores in the groups of professional and non-professional users of machine translation.

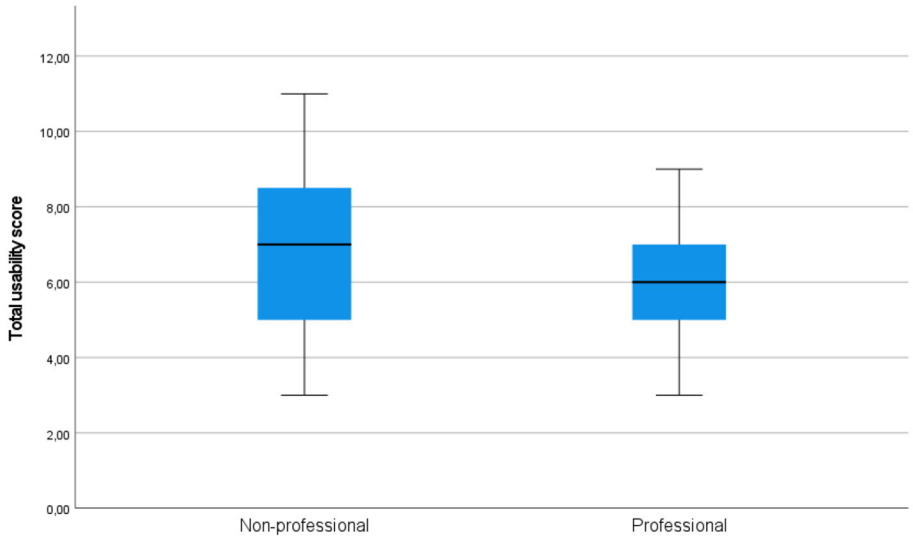


FIGURE 4
Total usability scores in the groups of professional and non-professional users of machine translation.

had low dwell times, this might indicate that the respondents who gave no correct answers were relatively inattentive.

Figure 14 shows how dwell times for AOIs with errors correlate with the number of correctly answered questions. The pattern again differs between professionals and non-professionals, with professionals having a higher spread for more correct answers and non-professionals having a higher spread for the average number of correct answers. Furthermore, both professionals and non-professionals who gave no correct answers had low fixation times (with the maximum lower than the medians of every other group), which may imply that less attention and effort while reading results in lower comprehension. Of course, such a finding needs to be tested and proven in a better targeted study, as in this particular case there might have been other factors like the text type, topic, tiredness, general absence of interest, etc. that influenced the results.

5. Discussion

Previous studies focusing solely on machine translation acceptability are few. Even fewer studies apply eye tracking to test machine translation acceptability. They mainly focus on the experiments with professional translators and/or translation students. To the best of our knowledge, there are no reported studies where machine translation acceptability by non-professional users was tested *via* an eye tracking experiment. No such research testing

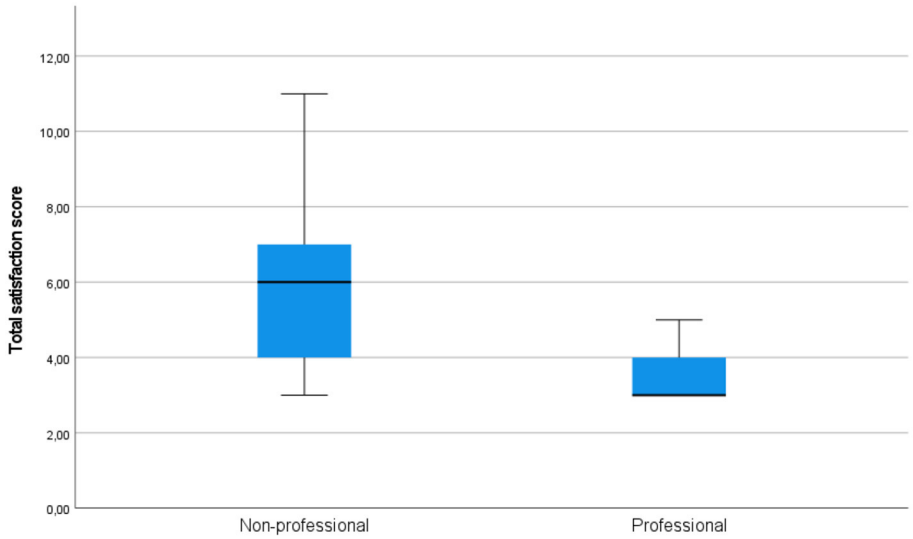


FIGURE 5
Total satisfaction scores in the groups of professional and non-professional users of machine translation.

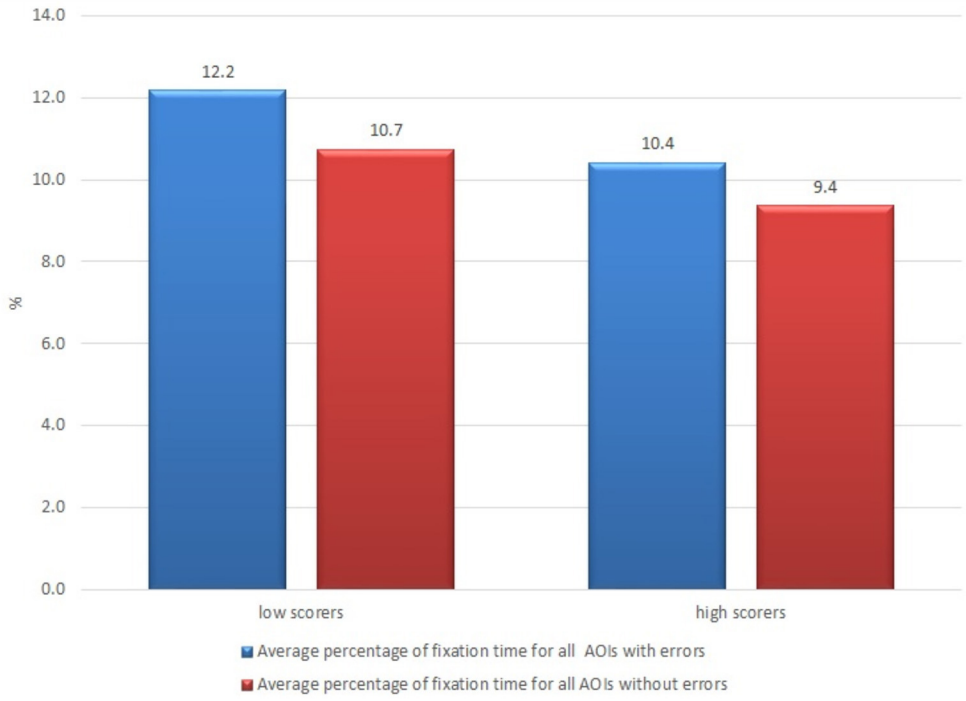


FIGURE 6
Average percentage of fixation time of non-professional users who rated quality of the raw machine-translated text higher and lower than the average.

acceptability of machine-translated text into Lithuanian has been conducted so far. Lithuanian, like many other smaller languages, is considered underresourced. It is also a morphologically rich synthetic language. Consequently, machine translation quality is less adequate than in other languages where investment into data acquisition and machine translation development is more substantial. Therefore, the views of Lithuanian language speakers, or smaller language speakers overall, toward machine translation might be diverse and involve many more risks or unexpected threats, if the output is used without critical awareness and judgment. For these reasons, comparisons between our results and previous research are only partial or indirect.

This study revolved around three research questions. The first question was related to comparison between professional and non-professional users' processing of raw machine translation output. The most obvious finding to emerge from this study is that there is a difference in the machine translation output cognitive processing and acceptability between professional and non-professional users. In comparison with non-professional users,

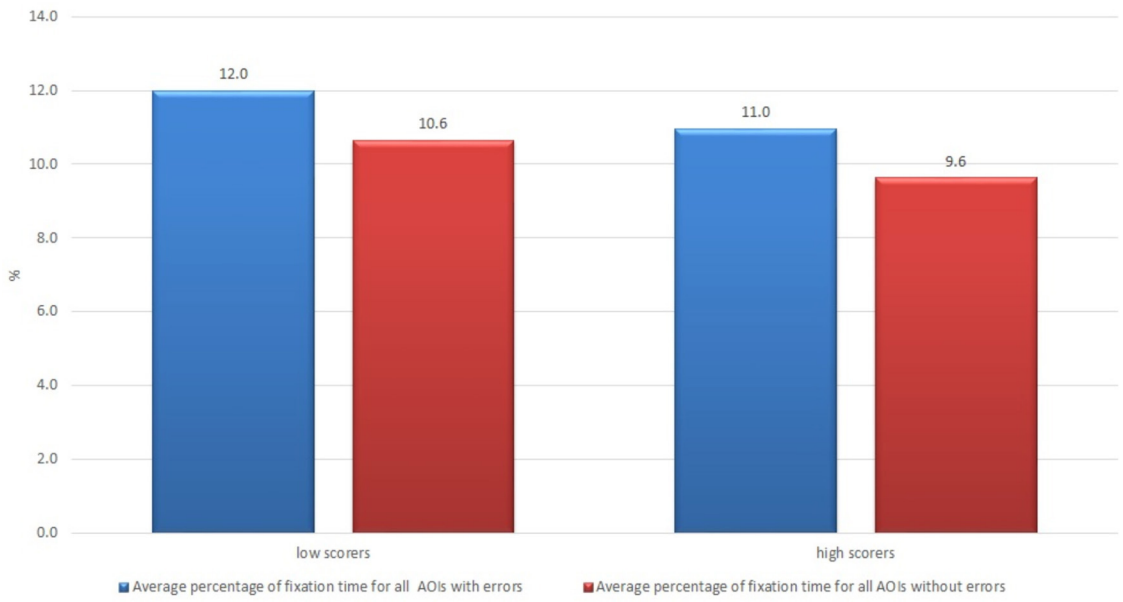


FIGURE 7
Average percentage of fixation time of non-professional users who rated usability of the raw machine-translated text high and low.

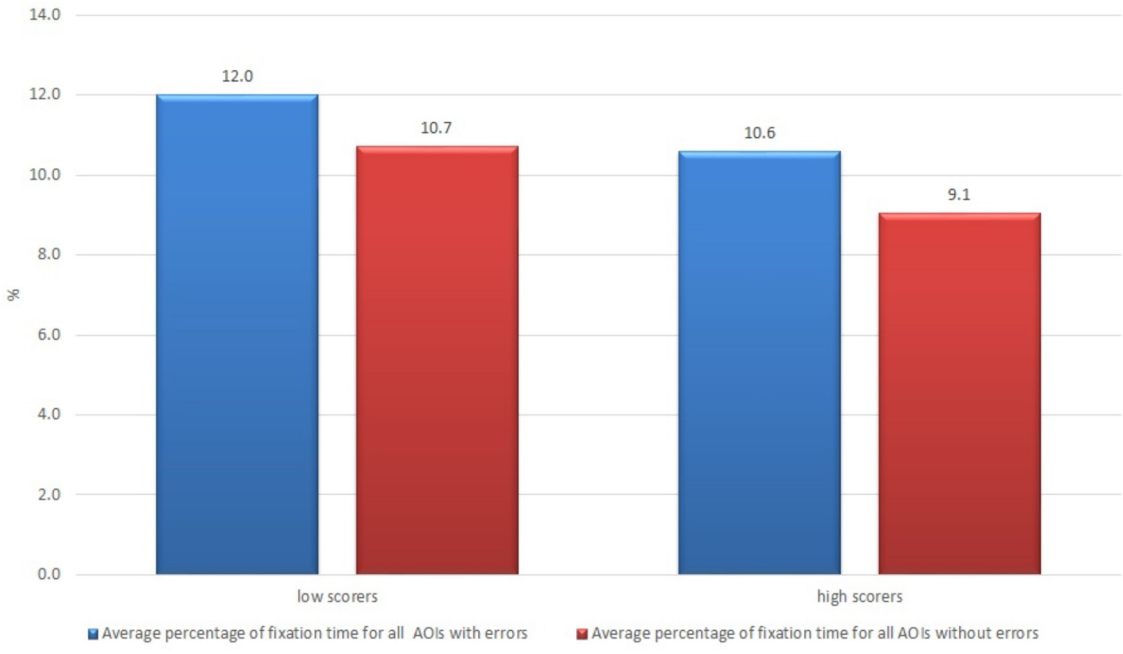


FIGURE 8
Average percentage of fixation time of non-professional users who were more and less satisfied with the raw machine-translated text.

professional users of machine translation, i.e., translators and language editors, spend more time overall reading the machine-translated texts, most probably because of their deeper critical awareness as well as proficient attitude toward the text. They also demonstrate a longer average fixation time and a greater average glance count on the machine translation errors. In terms of acceptability overall, professional users critically assess machine translation on all components of acceptability. This might possibly be explained by an assumption that professionals have less tolerance toward insufficient quality of machine translation, know how to prepare texts for publishable quality and see mistakes, inaccuracies and style issues in a text almost instantaneously. On the other hand, even if the text contains errors, it might still be usable.

The results obtained in this study seem to be to some extent consistent with the findings obtained in previous studies. García (2010) who investigated the level of user awareness of machine

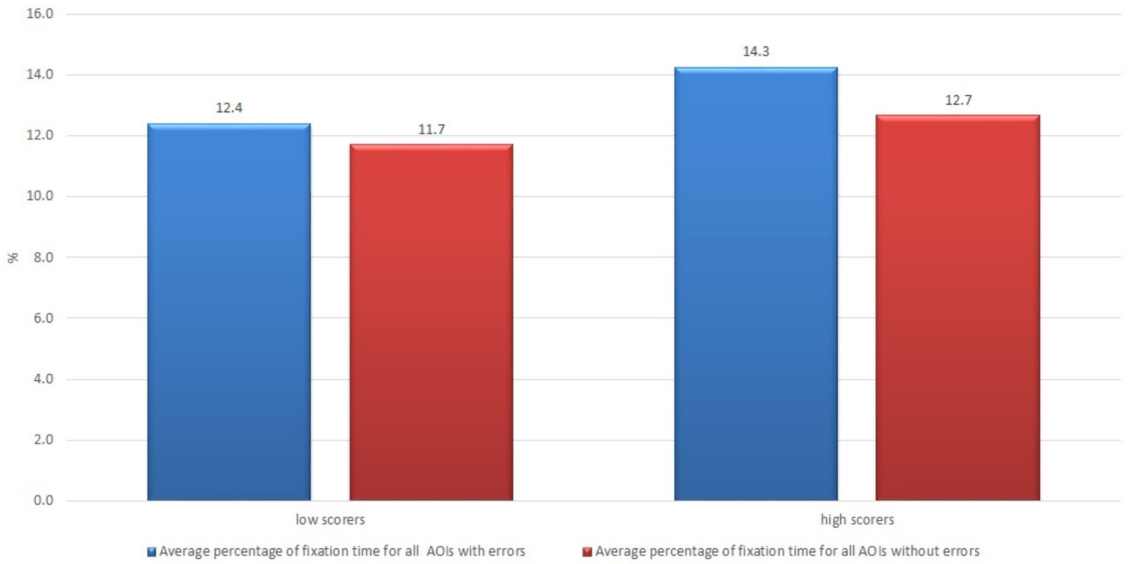


FIGURE 9
Average percentage of fixation time of professional translators who scored low and high regarding the usability with the raw machine-translated text in the post-task questionnaire.

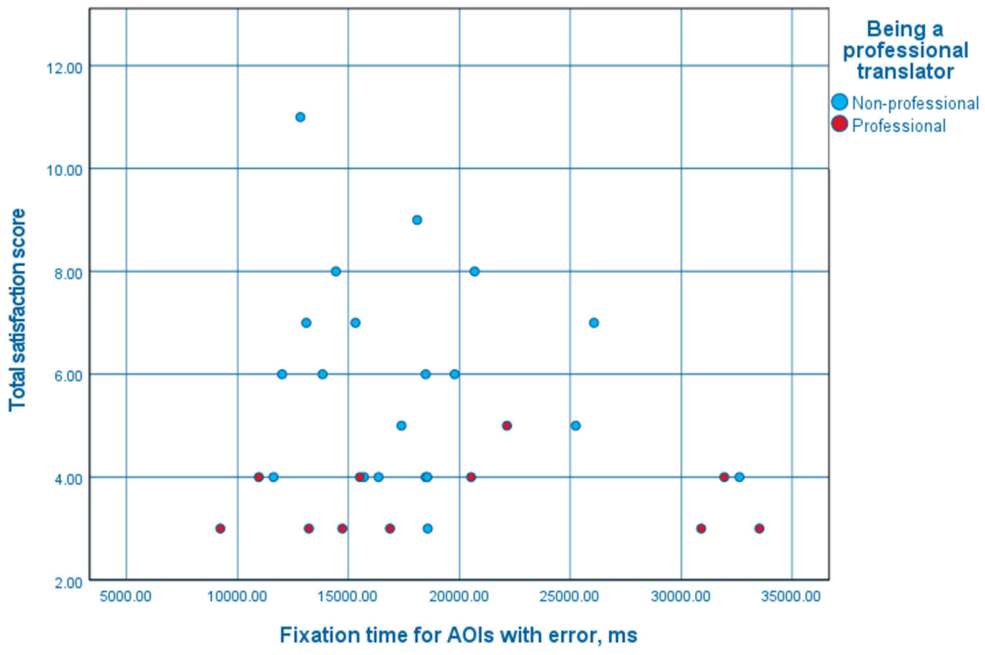


FIGURE 10
A scatter plot of fixation times for AOIs with errors and total satisfaction scores.

translation among professional translators and clients or users of translation found out that only a small proportion of professionals considered the quality of machine translation very high, which is not surprising since at the time machine translation had lower quality than the neural machine translation now. However, in the same study, the clients/users of translations demonstrated more positive assessment of the quality of machine translation compared to that of professional users (García, 2010). Our findings are also in line with the implications revealed by Vieira (2020) who concluded that there is a clear divide between the perceptions of professionals and non-professionals toward machine translation and its capabilities. In his study, Vieira acknowledged that the public coverage of machine translation veers more toward positive attitudes rather than negative. In our study, non-professional users—end-users with no linguistic background—also had more positive attitudes toward machine translation quality, usability and satisfaction compared with the professional translators' attitudes toward the text. However, in principle, our results may also be indirectly considered to be in

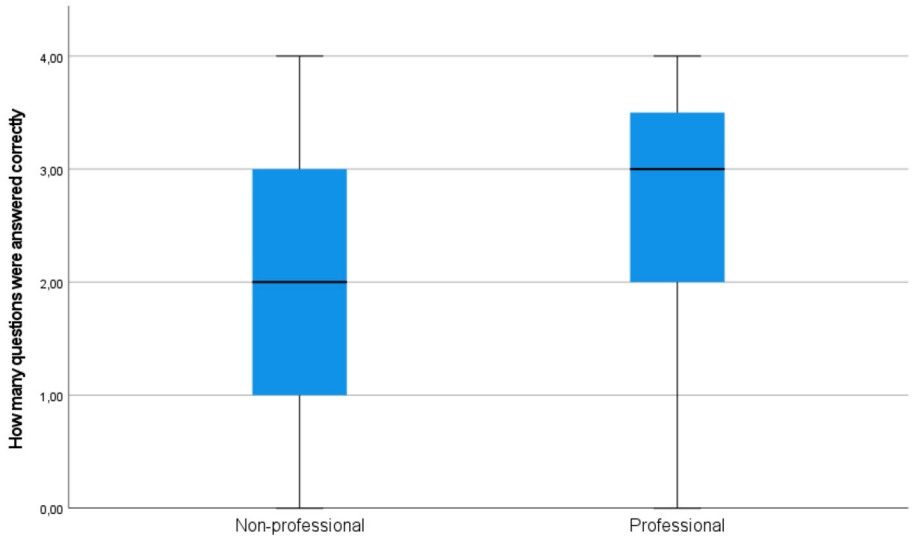


FIGURE 11
A simple box plot of text comprehension results in the groups of professionals and non-professional users.

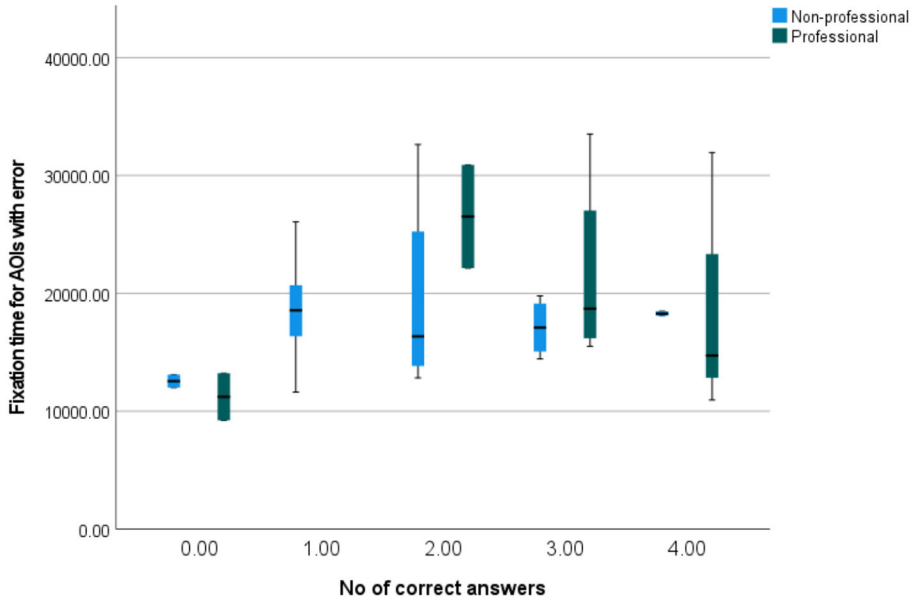


FIGURE 12
A simple box plot of fixation times for AOIs with errors by the number of correctly answered questions.

agreement with those obtained in a study by Hu et al. (2020) where subjects with low proficiency in English considered a raw machine-translated output quality lower than the post-edited text, i.e., one containing no errors. Although Hu et al.âĂŹs and our studies have different designs and purposes, it may be inferred that even non-professionals who may be expected to be ignorant of or care less about mistakes in the text are generally aware of drawbacks and notice them.

Some of our study results may also be to some extent comparable with those obtained in the investigation by Colman et al. (2021) where an increased number of eye fixations and increased gaze duration while reading machine translation segments were found in comparison with human translation (Colman et al., 2021), which may imply that less naturalistic and possibly erroneous text segments require more cognitive load. In our study, all respondents (both professionals and non-professionals) demonstrated increased values of all tested eye movement variables on areas of interest with errors compared with areas of interest without errors.

Other noteworthy findings to emerge from this study relate to the question whether there is a difference in the processing of raw machine-translated output between non-professional users with different levels of acceptability of machine-translated text. The

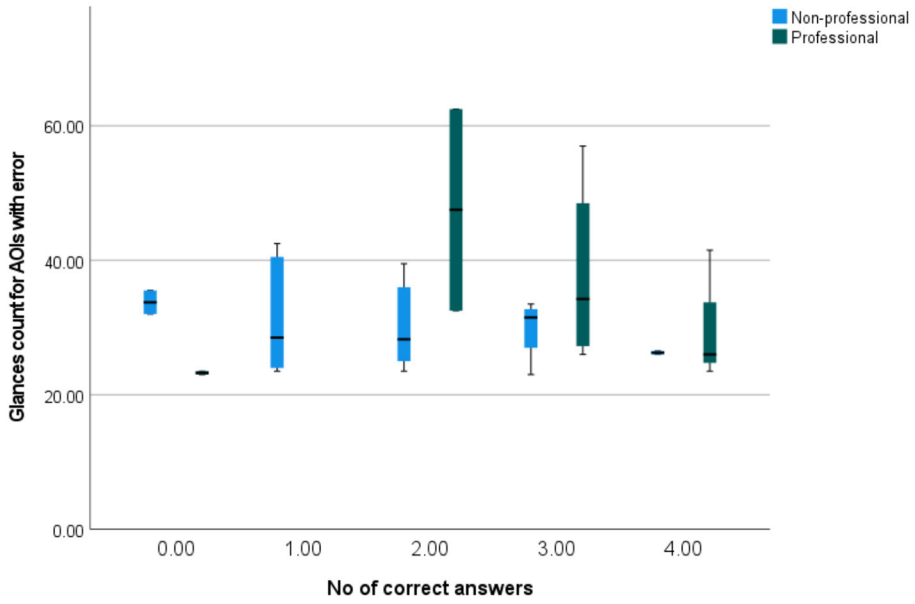


FIGURE 13
A simple box plot of glances counts for AOIs with errors by the number of correctly answered questions.

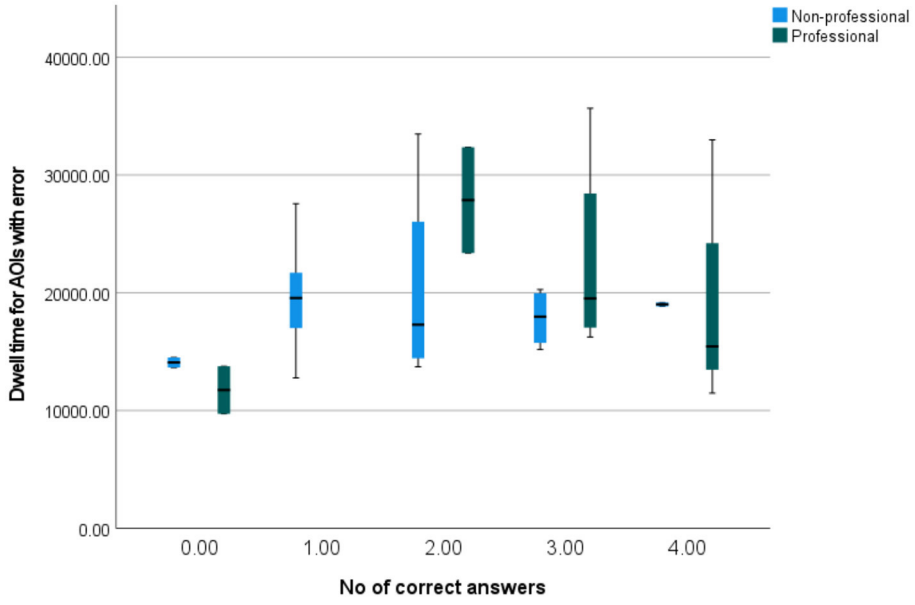


FIGURE 14
A simple box plot of dwell times for AOIs with errors by the number of correctly answered questions.

overall acceptability of machine-translated text was found to be higher for those non-professional users who spent less time/effort on areas of interest with errors, which might be an indication that the participants who did not notice or were more positive or tolerant about the mistakes were more positive about the machine-translated text in general. The text comprehension results revealed that the subjects in the group of professional translators who scored low on the comprehension questions, demonstrated a greater number of glance counts, which may imply that the professional background may influence the level of comprehension. These findings indirectly support the more positive attitudes toward raw machine-translated text in terms of its comprehension and trust by users with lower proficiency of language as reported by Rossetti et al. (2020).

However, with a relatively small sample size, caution must be applied while interpreting the results within the group of non-professional users of machine translation as the findings might be diverse depending on the subject's background, level of education, experience, language proficiency and other variables.

6. Conclusions

The study was aimed at determining the acceptability of raw machine translation texts in Lithuanian, a low-resource language. An eye tracking experiment measuring acceptability *via* the comparison between professional and non-professional users of machine translation and *via* the comparisons between the respondents who assessed the quality, satisfaction with and usability of the text differently (either lower or higher than average) revealed some insightful findings. There is a difference in the machine translation output cognitive processing and acceptability between professional and non-professional users. The professional users scored better in text comprehension compared with non-professional users. One of the possible reasons for that might be the experience of professional translators in dealing with badly written (perhaps also machine-translated) text. Professional users critically assess machine translation on all components of acceptability. Non-professional users—end-users with no linguistic background—have more positive attitudes toward machine translation quality, usability and satisfaction, which may imply possible risks if machine translation is used without critical awareness, judgement and revision. The lower professional users' satisfaction with the text, and overall acceptability, may suggest that they are likely to have higher expectations for the translated text.

The major general implication of these findings is the lower awareness of non-professional users regarding the machine translation output drawbacks and imperfections, which may result in a variety of misunderstandings that might go unnoticed and ignored, as well as risks and threats with undefined consequences.

The major limitation of this study is the small and uneven sample sizes of professional and non-professional users of machine translation. More equal sample sizes of different groups would help establishing a greater degree of accuracy on this matter. Besides, the differences within the non-professionals' group should be taken into consideration, as the results may be affected by various individual characteristics of subjects. Therefore, larger controlled trials could be focused more on the differences in educational backgrounds and language proficiency of subjects as well as the provided stimulus text variety or task description to give more definitive evidence regarding acceptability of machine translation.

A further limitation concerns imperfections of eye tracking equipment. To some extent they have been mitigated, but those mitigations can also be a cause of further limitations (for example, padding AOIs by about 1 character to all sides is a common way to mitigate imprecision of eye tracking leading to failures to notice the subject looking at the AOI, but it can result in including cases when the subject is looking at the area near the AOI).

Notwithstanding these limitations, the study provides a possibility to understand more deeply the readers' cognitive processing and the level of acceptability they exhibit toward machine-translated texts. Overall, the results of the study demonstrate diversified and contrasting views of the population and call for raising public awareness and machine translation literacy improvement.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Research Ethics Commission of Kaunas University of Technology. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

RK and JM: conceptualization. RK, JM, IP, and MP: methodology. RK, JM, IP, MP, and JH: investigation and writing—review and editing. IP and MP: data curation, visualization. RK, JM, and JH: writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding

This research had received funding from the Research Council of Lithuania (LMTLT, agreement No S-MOD-21-2).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Lykke Jakobsen, A. (2011). "The process of post-editing: a pilot study," in *Copenhagen Studies in Language* (Frederiksberg), 131–142.

Carl, M., Gutermuth, S., and Hansen-Schirra, S. (2015). "Chapter post-editing machine translation: efficiency, strategies, and revision processes in professional translation settings," in *Psycholinguistic and Cognitive Inquiries Into Translation and Interpreting* (Amsterdam: John Benjamins Publishing Company), 145–174.

Castilho, S. (2016). *Measuring acceptability of machine translated enterprise content* (Ph.D. thesis). Dublin: Dublin City University.

Castilho, S., and O'Brien, S. (2016). "Evaluating the impact of light post-editing on usability," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portoroz: European Language Resources Association, ELRA), 310–316.

Castilho, S., O'Brien, S., Alves, F., and O'Brien, M. (2014). "Does post-editing increase usability? a study with Brazilian Portuguese as target language," in *Proceedings of the 17th Annual conference of the European Association for Machine Translation* (Dubrovnik: European Association for Machine Translation), 183–190.

Castilho, S., and O'Brien, S. (2018). "Acceptability of machine-translated content: a multi-language evaluation by translators and end-users," in *Linguistica Antverpiensia, New Series - Themes in Translation Studies, Vol. 16* (Antwerp).

Colman, T., Fonteyne, M., Daems, J., and Macken, L. (2021). "It's all in the eyes: an eye tracking experiment to assess the readability of machine translated literature," in *31st Meeting of Computational Linguistics in The Netherlands (CLIN 31), Abstracts* (Ghent).

Daems, J., Vandepitte, S., Hartsuiker, R. J., and Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Front. Psychol.* 8, 01282. doi: 10.3389/fpsyg.2017.01282

Doherty, S. (2016). Translations| the impact of translation technologies on the process and product of translation. *Int. J. Commun.* 10, 947–969.

Doherty, S., and O'Brien, S. (2014). Assessing the usability of raw machine translated output: a user-centered study using eye tracking. *Int. J. Hum. Comput. Interact.* 30, 40–51. doi: 10.1080/10447318.2013.802199

Fereira, A, Gries, S. T., and Schwieter, J. W. (2021). "Assessing indicators of cognitive effort in professional translators: A study on language dominance and directionality," in *Translation, interpreting, cognition: The way out of the box*, ed Tra&Co Group (Berlin: Language Science Press), 115–143. doi: 10.5281/zenodo.4545041

García, I. (2010). Is machine translation ready yet? *Target* 22, 7–21. doi: 10.1075/target.22.1.02gar

Guerberof Arenas, A., Moorkens, J., and O'Brien, S. (2021). The impact of translation modality on user experience: an eye-tracking study of the microsoft word user interface. *Mach. Transl.* 35, 205–237. doi: 10.1007/s10590-021-09267-z

Hoi, H. T. (2020). Machine translation and its impact in our modern society. *Int. J. Sci. Technol. Res.* 9, 1918–1921.

Hu, K., O'Brien, S., and Kenny, D. (2020). A reception study of machine translated subtitles for MOOCs. *Perspectives* 28, 521–538. doi: 10.1080/0907676X.2019.1595069

Jakobsen, A. L., and Jensen, K. T. H. (2008). Eye movement behaviour across four different types of reading task. *Copenhagen Stud. Lang.* 36, 103–124.

Kasperė, R., Horbačauskienė, J., Motiejunienė, J., Liubinienė, V., Patašienė, I., and Patašius, M. (2021). Towards sustainable use of machine translation: usability and perceived quality from the end-user perspective. *Sustainability* 13, 3430. doi: 10.3390/su132313430

Kasperavičienė, R., Motiejūnienė, J., and Patašienė, I. (2020). Quality assessment of machine translation output. *Texto Livre* 13, 271–285. doi: 10.35699/1983-3652.2020.24399

Kasperė, R., and Motiejūnienė, L. (2021). "Eye-tracking experiments in human acceptability of machine translation to study societal impacts," in *Sustainable Multilingualism 2021: The 6th International Conference*, eds A. Daukšaite-Kolpakoviene and Ž. Tamašauskaite (June 4–5, 2021, Kaunas, Lithuania): book of abstracts (Kaunas: Vytautas Magnus University), 114.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (MQM): a framework for declaring and describing translation quality metrics. *Tradumàtica tecnol. trad.* 12, 455. doi: 10.5565/rev/tradumatica.77

McConkie, G. W., and Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Percept. Psychophys.* 17, 578–586. doi: 10.3758/BF03203972

Moorkens, J. (2018). "Chapter Eye-Tracking as a Measure of Cognitive Effort for Post-Editing of Machine Translation," in *Eye Tracking and Multidisciplinary Studies on Translation* (Amsterdam: John Benjamins Publishing Company), 55–69.

Moorkens, J., and O'Brien, S. (2015). "Post-editing evaluations: trade-offs between novice and professional participants," in *Proceedings of the 18th Annual Conference of the European Association for Machine Translation* (Antalya), 75–81.

MQM Commitee. (2022). *MQM (Multidimensional Quality Metrics): what is MQM?* Available online at: https://themqm.org/ (accessed October 10, 2022).

Nurminen, M., and Koponen, M. (2020). Machine translation and fair access to information. *Transl. Spaces* 9, 150–169. doi: 10.1075/ts.00025.nur

Ortega, J., Sánchez-Martínez, F., Turchi, M., and Negri, M. (2019). "Improving translations by combining fuzzy-match repair with automatic post-editing," in *Proceedings of Machine Translation Summit XVII: Research Track* (Dublin: European Association for Machine Translation), 256–266.

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., et al. (2019). *The ai index 2019 annual report*. Technical report, Stanford: AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

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

Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: a systematic review. *Lang. Resour. Eval.* doi: 10.1007/s10579-021-09537-5

Rossetti, A., O'Brien, S., and Cadwell, P. (2020). "Comprehension and trust in crises: investigating the impact of machine translation and post-editing," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (Lisboa: European Association for Machine Translation), 9–18. Available online at: https://aclanthology.org/2020.eamt-1.2

Rossi, C., and Carré, A. (2022). "How to choose a suitable neural machine translation solution: Evaluation of MT quality," in *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*, ed D. Kenny (Berlin: Language Science Press), 51–79. doi: 10.5281/zenodo.6759978

Rossi, C., and Chevrot, J.-P. (2019). Uses and perceptions of machine translation at the european commission. *J. Special. Transl.* 31, 177–200. Available online at: https://shs.hal.science/halshs-01893120/file/Rossi_and_Chevrot_article8.pdf

Schuster, M., Johnson, M., and Thorat, N. (2016). *Zero-shot Translation With Google's Multilingual Neural Machine Translation System*. Google AI blog. Available online at: https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html

Stasimioti, M., and Sosoni, V. (2021). "Chapter Investigating post-editing: a mixed-methods study with experienced and novice translators in the English-Greek language pair, ' *Translation, Interpreting, Cognition: The Way Out of the Box* (Berlin: Language Science Press).

Taivalkoski-Shilov, K., Toral, A., Hadley, J. L., and Teixeira, C. S. C., editors (2022). "Using technologies for creative-text translation," in *Routledge Advances in Translation and Interpreting Studies* (London: Routledge).

Taylor, R. M., Crichton, N., Moult, B., and Gibson, F. (2015). A prospective observational study of machine translation software to overcome the challenge of including ethnic diversity in healthcare research. *Nurs. Open* 2, 14–23. doi: 10.1002/nop2.13

Ueffing, N. (2018). "Automatic post-editing and machine translation quality estimation at eBay," in *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing* (Boston, MA: Association for Machine Translation in the Americas), 1–34.

Vardaro, J., Schaeffer, M., and Hansen-Schirra, S. (2019). Translation quality and error recognition in professional neural machine translation post-editing. *Informatics* 6, 41. doi: 10.3390/informatics6030041

Vieira, L. N. (2020). Machine translation in the news. *Transl. Spaces* 9, 98–122. doi: 10.1075/ts.00023.nun

Vieira, L. N., O'Hagan, M., and O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Inf. Commun. Soc.* 24, 1515–1532. doi: 10.1080/1369118X.2020.1776370

Yasuoka, M., and Bjorn, P. (2011). "Machine translation effect on communication: what makes it difficult to communicate through machine translation?" in *2011 Second International Conference on Culture and Computing* (Kyoto: IEEE).

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 15 February 2023
DOI 10.3389/fpsyg.2023.1107839

OPEN ACCESS

EDITED BY
Ernest Greene,
University of Southern California,
United States

REVIEWED BY
Jiangjie Chen,
Jiangnan University,
China
Alex Miklashevsky,
University of Potsdam,
Germany

*CORRESPONDENCE
Tanja Medved
✉ tanja.medved@ntf.uni-lj.si

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

RECEIVED 25 November 2022
ACCEPTED 25 January 2023
PUBLISHED 15 February 2023

CITATION
Medved T, Podlesek A and Možina K (2023) Influence of letter shape on readers' emotional experience, reading fluency, and text comprehension and memorisation. *Front. Psychol.* 14:1107839. doi: 10.3389/fpsyg.2023.1107839

Influence of letter shape on readers' emotional experience, reading fluency, and text comprehension and memorisation

Tanja Medved[1]*, Anja Podlesek[2] and Klementina Možina[1]

[1]Faculty of Natural Sciences and Engineering, University of Ljubljana, Ljubljana, Slovenia, [2]Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

Introduction: The amount of educational material delivered to pupils and students through digital screens is increasing. This method of delivering educational materials has become even more prevalent during the COVID-19 pandemic. To be as effective as possible, educational material must be properly designed not only in terms of content, but also in terms of form, e.g., the typeface. The present study investigated the effect of letter shape on readers' feelings of pleasantness during reading, reading fluency, and text comprehension and memorisation.

Methods: To find out whether age influences the effects of typeface shape on reading measures, we divided the participants into a group of less experienced readers (children) and more experienced readers (adults). Both groups read texts in eight different typefaces: four of them were round or in rounded shape, and four were angular or in pointed shape. With an eye-tracker, the reading speed and the number of regressive saccades were recorded as measures of reading fluency and changes in pupil size as an indicator of emotional response. After reading each text, the participants rated the pleasantness of the typeface, and their comprehension and memorisation of texts were checked by asking two questions about the text content.

Results: We found that compared to angular letters or letters in pointed shape, round letters or letters in round shape created more pleasant feelings for readers and lead to a faster reading speed. Children, as expected, read more slowly due to less reading experiences, but, interestingly, had a similar number of regressive saccades and did not comprehend or remember the text worse than university students.

Discussion: We concluded that softer typefaces of rounder shapes should be used in educational materials, as they make the reading process easier and thus support the learning process better for both younger and adult readers. The results of our study also showed that a comparison of findings of different studies may depend on the differences among the used letter shapes.

KEYWORDS

typeface shape, pleasantness, reading fluency, text comprehension, text memorisation, age differences

1. Introduction

The shape of letters and the general typographic design of a text affect the legibility of the text (Beier at al., 2017), the transparency of the presentation of information (Brath and Banissi, 2016) and, consequently, the fluency of reading (Gasser et al., 2005; Beier and Larson, 2013; Cacali, 2016; Bessemans, 2016a,b). The present study examined the effect of letter shape on readers' feelings of pleasantness during reading, pupil size and eye movements during reading, and text comprehension and memorisation.

1.1. Reading fluency

The concept of reading fluency combines accuracy and speed of reading with the ability to comprehend the content being read. Some definitions of reading fluency focus more on letter recognition and reading speed (Meyer and Felton, 1999), while others include content comprehension (Pikulski and Chard, 2005).

Many factors affect reading fluency. Reading fluency is affected by the shape or legibility of the typeface (Ali et al., 2013), type size (Mueller et al., 2014; Su et al., 2018), and the overall typographic design of the text (Koch, 2012). When reading on a screen, reading fluency is also affected by the screen resolution, as with a higher resolution, letters and their features can be displayed better (Bessemans, 2016a; Bigelow, 2019).

1.2. Text comprehension and memorisation

Several studies showed that the shape of letters and the text can influence the comprehension of the read content (Choi et al., 2018) and the actual memorisation of the read content (Lewis and Walker, 1989; Gasser et al., 2005). Poorer fluency results in poorer information processing and, consequently, poorer comprehension and memorisation of the text (Novemsky et al., 2007; Oppenheimer and Frank, 2008; Meyer et al., 2015; Bjork and Yue, 2016; Pieger et al., 2016; Rummer et al., 2016; Sanchez and Nayor, 2018; Dressler, 2019; Wu et al., 2019).

Studies examining how using a perceptually difficult-to-process typeface with an increased desirable difficulty designed specifically to reduce legibility, such as Sans Forgetica, found either no processing or memory benefit of such typefaces or even yielded a memory cost (Geller et al., 2020; Taylor et al., 2020; Wetzler et al., 2021; Cushing and Bodner, 2022; Maxwell et al., 2022). However, there is also a whole series of studies which showed that poorer fluency of the text or desired difficulty in the fluency of the text resulted in better processing of the text and consequently in better memorisation of the read content (Diemand-Yauman et al., 2011; Macdonald and Lavic, 2011; Bjork et al., 2013; Halin, 2016; Pieger et al., 2016).

Numerous studies demonstrated that reading fluency affects the learning process, more specifically short-term and long-term memory (Weissgerber and Reinhard, 2017), as well as metacognition (Yue et al., 2013; Ilic and Akbulut, 2019). Based on the shape of letters and the text, readers can predict how long it will take them to read the text and remember the content of the text (Beier and Larson, 2013; Price et al., 2016). Higher reading fluency should promote a positive attitude towards the text, consequently the feeling of better memorability of the text, and it should allow for better memorisation and comprehension of the text (Song and Schwarz, 2008; Mueller et al., 2013; Labro and Pocheptsova, 2016; Pieger et al., 2016; Mead and Hardesty, 2018). In contrast, poorer fluency should promote poorer attitudes toward the text and readers should assume that they will spend more time reading and memorising the text.

1.3. The role of emotions in the reading process

Emotions play a specific role in reading. The typographic design or the shape of the typeface has a great impact on the reader's mood, more specifically on their emotional response or feeling of pleasantness that the reader experiences when reading certain letterforms (Larson and Picard, 2005; Larson et al., 2006; Koch, 2012; Petit et al., 2015). The shape of letters and the text can suggest the nature and content of the text to the reader (Lewis and Walker, 1989; Ehsen and Lupton, 1998; Celhay et al., 2015; Bigelow, 2019; Davis, 2019; Raden and Qeis, 2019).

Several studies have shown that the perception of shapes, tastes and sounds evokes various feelings in humans, including the feeling of pleasantness (Childers and Jass, 2002; Brumberger, 2003; Mackiewicz, 2005; Shaikh et al., 2006; Bar and Neta, 2007; Tsonos and Kouroupetroglou, 2011; Amare and Manning, 2012; Crisinel et al., 2012; Ngo et al., 2013; Velasco et al., 2014, 2015a,b, 2016, 2018a; Salgado-Montejo et al., 2015; Jordan, 2017; Davis, 2019; Haenschen and Tamul, 2019). Round and rounded shapes, as well as symmetric shapes evoke more pleasant feelings than angular or pointed and asymmetric shapes (Bar and Neta, 2007; Ngo et al., 2013; Turoman et al., 2018; Velasco et al., 2018b).

We have not found a study that would examine how these features of human perception can be effectively used in the typographic design of educational materials, but based on the previous studies we can assume that round typefaces would evoke more pleasant feelings than angular ones.

1.4. The influence of letter shape on the reading process

The core of typographic design are typefaces, which can be grouped based on the shape of the main strokes, and the transitions between the strokes and the stroke ends (terminals, serifs). One group of typefaces contains round/rounded typefaces and the other group contains angular/pointed typefaces. A typical example of typefaces that could be classified in the round/rounded group based on their design features are typefaces that belong to the group of Venetian, Garalde and Transitional typefaces (McLean, 1997; Možina, 2003). Typefaces that could be classified in the angular/pointed group based on their design characteristics are typefaces that belong to the Didone, Slab-Serif and Sans Serif group (McLean, 1997; Možina, 2003).

Previous studies found that rounded, organic shapes of strokes and softer transitions between the strokes and stroke ends are perceived as more pleasing whereas the letters with more geometric stroke shapes, sharp transitions between strokes and finial stroke are found to be less pleasant (Spence and Deroy, 2012; Hyndman, 2016). The feeling of pleasure we experience when reading different typefaces influences motivation and concentration (Mano, 1997; Koch, 2012), memorisation and comprehension of a text (Mano, 1997). However, research addressing how the reader's emotional response to the shape of letters affects reading fluency is scarce.

1.5. The effect of age on reading

It has been shown that perception in reading also depends on the age of the reader. Children and adult readers differ in the level of development of cognitive and physiological abilities until the age of four, after which the ability to recognise letters should be the same in children and adults (Woods et al., 2005). However, studies reported that children from 4 to 11 years old react to different stimuli, e.g., colour, shape, taste, smell, differently from adult students (Gollely and Guichard, 2011). It has been discovered that reactions to the same stimuli are different also in younger adults (under 35 years old) and

older adults (over 60 years old) (Piqueras-Fiszman et al., 2011). In younger readers (aged 7 to 9 years), typefaces with serifs and a difference in stroke width were found to lead to more fluent reading, whereas sans serif typefaces that have no or minor difference in stroke width result in fewer reading errors (Wilkins et al., 2009). It is also claimed that a larger type size allows faster decoding of information and better memory, but only in children (age 9 to 12 years old), not in adult students (Abukaber and Lu, 2012). Children (from 7 to 12 years old) read letters that are heterogeneous in shape more easily (Wilkins et al., 2009; Abukaber and Lu, 2012); especially the heterogeneity in the shape of letters seems to greatly aid visually impaired children (age 5 to 10 years old) in reading (Bessemans, 2016b). In the study conducted by Katzir et al. (2013), the increased desirable difficulty of the typeface affected reading fluency, demonstrating positive effects in older children (11 years old), but negative effects in younger children (8 years old).

1.6. The aim of our study

Our study had two aims. The first aim was to determine how the shape of the typeface (round/rounded vs. angular/pointed) affects reading fluency, subjective reading experience, and reading performance. The second aim of our study was to investigate whether the effect of typeface shape is the same for younger, less experienced readers and for adult, more experienced readers.

We used an eye-movement tracking device as it provides objective measures of reading fluency (Piqueras-Fiszman et al., 2013; Franken et al., 2015). We monitored the reading speed and regressive saccades as measures of reading fluency. We also used this device to observe changes in pupil size, which should be indicative of the reader's emotional response (Hess and Polt, 1960; Margareth et al., 2008; Wang et al., 2018). Objective measures of emotional response to different shapes of typefaces were complemented with subjective ratings of feelings of pleasantness. Text comprehension and memorisation of what was read were also observed as indicators of reading performance.

2. Methods and materials

The studies involving human participants were reviewed and approved by the Ethics Commission of the Faculty of Arts, University of Ljubljana. An informed consent document to participate in this study was provided by the participants or their legal guardian/next of kin. All studies were performed in accordance with the Declaration of Helsinki.

2.1. Apparatus

To track eye movements, we used a Tobii X120 eye-tracking device and Tobii Studio 3.4.8 software (Tobii AB, Sweden). The device tracks eye movements by tracking the reflection of the image from the cornea. The corneal reflection is generated by infrared emitters on the front of the device that create IR light patterns that are then reflected off the cornea. The device contains a camera that is sensitive to IR light and monitors each movement and fixation of the eye based on the reflection of IR light from the cornea (Tobii Pro, 2017).

Before the measurements, each participant had 5 min to adapt to the lighting conditions in the test room and to perform a nine-point screen-based calibration of the device. We used an LCD screen with a resolution of 2,400 × 1900 pixels (pixel size 0.27 mm) and a refresh rate of 60 Hz.

2.2. Preliminary studies

Prior to the main study in which we investigated how the shape of different typefaces affects the pleasantness ratings and the reading speed, memorisation and understanding of a text, we conducted two preliminary studies. The purpose of the first preliminary study was to select eight texts comparable in cognitive load and the purpose of the second preliminary study was to select eight typefaces.

The measurements were done in a quiet room with walls painted with grey matte paint in accordance with the ISO 3664 standard (ISO 3664, 2009). The letters of the texts that the participants read on the screen were dark on a light background (text colour: #000, background colour: #eee) according to the ISO 12646 standard (ISO 12646, 2015). The participants were located at a distance of 60 cm +/− 1 cm from the screen, in line with the recommendations of the ISO 9241-303 standard (ISO 9241-303, 2011). Their movements were not restricted, but they were asked to remain at a fixed position.

2.2.1. First preliminary study

With the first preliminary study, we selected texts for the main study. Thirty-one students and employees of Faculty of Natural Sciences and Engineering at the University of Ljubljana participated in the study. Their mean age was 44.2 years ($SD = 7.4$), 22 were female and 9 were male. They were not paid for their participation in the study. They reported normal or corrected-to-normal vision.

We prepared 45 different texts in Slovenian (participants' native language) with contents of similar complexity. The texts were (i) sample texts published as a part of guidelines developed for teachers on how to evaluate the reading efficacy in children (Pečjak and Kramarič, 2018) and (ii) excerpts from a children's illustrated encyclopaedia about animals (Burnie, 2010). The selected texts had a meaningful beginning and end. They had a length of 457 to 510 characters without spaces ($SD = 13.55$). They appeared on the screen in 10 or 11 lines ($SD = 0.43$) in the Verdana typeface, type size 16 pixels. The text was displayed as an HTML document using the CSS programming language. In this way, we were able to ensure that the text was always displayed in exactly the same type size and position on the screen (i.e., in the centre of the screen).

After calibration, the 45 texts were presented in the same order to all the participants. Consecutive texts were invoked by a mouse click. For each text, we measured the reading speed and the number of fixations.

From the 45 texts, we selected 8 texts for the main study that showed highest reading speeds. They contained 471–510 characters ($M = 492$, $SD = 18$). The average reading speed of the selected 8 texts across the participants varied between 50.39 ms and 56.30 ms per character ($M = 52.45$ ms, $SD = 2.20$ ms). We also examined the number of fixations for each text as another indicator of reading fluency. The lower the number of fixations, the more fluently the participants read the text. The average number of fixations per character varied between 0.35 and 0.44 ($M = 0.38$, $SD = 0.03$). The texts seemed comparable in content complexity and suitable for fluent reading of the general population, including children, and contained no distracting factors such as overly long and demanding words and unclear content. The comparable

content difficulty, reading speed and relative number of fixations across the eight selected texts lead us to believe that the texts will result in a similar cognitive load when presented in the main study. The texts and their English translations can be found in the Supplementary Materials.

2.2.2. Second preliminary study

With the second preliminary study, we collected different typefaces for the main study. Fifty-five participants were included, 34 of whom were university students from the same institution as in the first preliminary study. They were between 19 and 26 years old, with the average age of 20.7 years ($SD = 1.3$). Twenty-three were female and 11 were male. The remaining 21 participants were second-triad primary school pupils aged 10 to 12 years, with the average age of 10.7 years ($SD = 0.6$). Ten of them were female and 11 were male. All participants had normal or corrected-to-normal vision and were not paid for their participation in the study.

We checked pleasantness of 15 different typefaces (i.e., Adobe Caslon Pro, American Typewriter, Anka, Arial Nova, Birch STD, Chaparral Pro, Comic Sans MS, Didot, Erlenmeyergraphy, FG April Trial, Matilda, Nogomet, Sans Forgetica, Times New Roman, Verdana). The participants read 15 pangrams on the screen. A pangram is a sentence or a portion of a text that uses all the letters of the alphabet and is typically difficult to read since the content of the sentences formed is unusual or senseless. Each pangram was displayed in a different typeface. The average length of a pangram was 36.9 characters without spaces ($SD = 3.7$). The pangrams were displayed in the centre of the screen. Consecutive texts were invoked by a mouse click.

Using a 5-level hedonic scale, participants rated how pleasant they found each typeface (1 – very unpleasant, 2 – unpleasant, 3 – neutral, 4 – pleasant, 5 – very pleasant). Based on the results, we selected eight final typefaces: four rated as most pleasant and four as least pleasant. The four most pleasing typefaces were Chaparral Pro, FG April Trial, Matilda and Times New Roman. These typefaces all had round/rounded shapes: the transitions between the strokes and the stroke ends (terminals, serifs) were soft, just like the transitions between the thick and thin strokes. Also, the shape of the bowls and counters was round and more convex, which is why we considered them as members of the group of round/rounded typefaces. The four least pleasing typefaces had the characteristics of angular/pointed shapes, i.e., Arial Nova, Nogomet, Sans Forgetica, Verdana. These typefaces were all sans serif typefaces, all of them had angular or pointed shaped stroke ends (terminals) and none or minor difference between the thick and thin strokes. The shape of the bowls and counters, especially on the left end right side of the bowl, was less convex and more straight, which is why we considered them as members of the group of angular/pointed typefaces. Figure 1 shows examples of all eight typefaces that we selected for use in the main study.

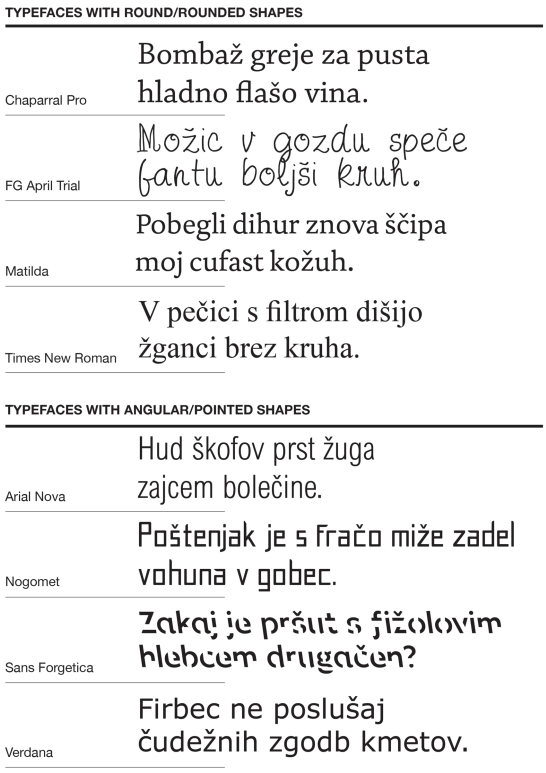


FIGURE 1
Eight typefaces selected for the main study.

2.3. Main study

2.3.1. Participants

Twenty university students (adult readers; 7 male, 13 female) aged between 18 and 26 ($M = 20.0$ years, $SD = 1.8$ years) and 15 children (pupils of grades 4 to 6 of primary school; 9 male, 6 female) aged 10 through 11 ($M = 10.7$ years, $SD = 0.5$ years) participated in the main study. All participants had normal or corrected-to-normal vision and were not paid for their participation.

2.3.2. Stimuli

For the main study, we used eight selected texts from the first preliminary study and eight selected typefaces from the second preliminary study. Each text was set in one of the typefaces. The texts in different typefaces are shown in the Supplementary Materials.

The size of the typeface was adjusted to achieve the most uniform x-height across typefaces possible, which varied between 0.17 and 0.20 degrees of visual angle; the average x-height was 0.19 degrees of visual angle ($SD = 0.016$). Due to the different shape of the letters, the number of lines of different texts varied between 10 and 11, and the average number of lines was 10.13 ($SD = 0.35$). In all cases, the leading (i.e., line spacing) was 140% of the type size.

2.3.3. Procedure

The main study was conducted under the same standardized conditions as in the first and second preliminary studies. The exception was the lighting in the room, which was now a bright light room, with artificial lighting.

To control for the effect of fatigue, each participant read the texts in a different order (the so-called Latin square). We measured the reading speed, number of saccades, length of fixations, and the size of the pupils for each text in all participants during the whole reading time.

After reading the text on the screen, participants answered two additional questions to check their understanding and remembering of the text content. Text comprehension was checked with a question about the text content. Each reader had three answers possible, from which they chose the one they thought was correct. Text memorisation was checked by presenting the readers with a sentence and asking them whether they had read that exact sentence in the text. They also rated

the pleasantness of the typeface with which the text was displayed, using a 5-point rating scale (1 – very unpleasant, 2 – unpleasant, 3 – neutral, 4 – pleasant, 5 – very pleasant).

2.3.4. Data analysis

We considered the rating of pleasantness as a subjective measure of emotional response to typefaces, and pupil size as an objective measure of such response. Pupil size should be enlarged when a person experiences or perceives something pleasant (Hess and Polt, 1960; Margareth et al., 2008; Wang et al., 2018). Although pupil size under controlled lighting conditions may reflect factors other than the reader's emotional response, such as surprise (Preuschoff et al., 2011) or cognitive load and metacognitive confidence (Gavas et al., 2018), we assume that these effects were minimized due to careful selection of texts in the preliminary study. We examined left pupil size (pupil diameter measured in millimeters). Pupil size changed during reading, but a careful examination of how it changed over time did not reveal specific patterns that could be generalized across different texts within a single participant or across different participants reading the same text. The 5-percent trimmed means of pupil diameter during the total time of reading a given text, which would eliminate potential outliers, were not significantly different from the uncorrected mean values (the difference was to the third decimal place), so we decided to use an uncorrected mean value of pupil diameter during the reading interval in further analyses.

Two objective indicators of reading fluency were analyzed, namely the number of regressive saccades and reading speed. Reading speed was determined by measuring the time spent per character (excluding spaces). Text comprehension and short-term text memory were used as measures of reading performance.

Data were analyzed using linear mixed modelling in the GAMLj module (Gallucci, 2020) for jamovi (The Jamovi Project, 2019).

To determine the extent to which pupil size actually reflects emotional response (typeface pleasantness), we first examined the relationship between subjective and objective indicators of emotional response to reading. Pupil size was used as an interval outcome variable, and pleasantness ratings centred within subjects were used as an interval predictor in the linear mixed model. The data were nested within participants. Participants were entered in the model as random intercepts and slopes.

Next, six different linear mixed models were developed. In each model, one of the six measures (pleasantness ratings, pupil size, number of regressive saccades, reading speed, text comprehension score, and text memorisation score) served as the outcome variable. Eight texts (level-1 units) were nested within 35 participants (level-2 units). Typeface shape was used as a level-1 predictor, i.e., as a within-subject factor-type variable with two levels describing the shape of the typeface (0 – round/rounded vs. 1 – angular/pointed typeface shape). Age was used as a level-2 predictor, i.e., a between-subject factor-type variable describing the participant (0 – child vs. 1 – university student). Three fixed effects were entered in the prediction model: the effect of typeface shape, the effect of age, and the interaction between age and typeface shape. To account for the inter-individual differences in the measured outcome variables, participants were entered in the model as random intercepts. Because we expected the effect of typeface shape to differ across participants, we also included the random slopes for typeface shape in the model. Equation 1 shows the model for predicting the outcome variable (Y′).

$$\begin{aligned} Y' = {} & b_0 + b_1 \bullet \text{Age} + b_2 \bullet \text{Typeface shape} \\ & + b_3 \bullet \text{Age} \times \text{Typeface shape} + (\text{Intercept}|\text{Participant}) \\ & + (\text{Typeface Shape}|\text{Participant}) \end{aligned} \quad (1)$$

To examine the effect of a factor (typeface shape or age) manipulation on each of the six outcome variables, we compared Bayes factors (BF) for different models. We used the default settings of the BayesFactor package (Morey and Rouder, 2022) to calculate the BFs. The package specifies the Jeffrey prior for the grand mean and error variance, uses the default setting for the multivariate Cauchy prior distributions (scale set to 0.5 and 1 for fixed effects and random effects, respectively), and does not explicitly model the correlation between random slopes and intercepts (van Doorn et al., 2021). There is a "lack of clarity and consensus about how to best conduct Bayesian model comparison when considering mixed effects" (van Doorn et al., 2021, p. 2). Because we assume that some inter-individual variability is intrinsically present in the level of outcome variables and in the effect of typeface shape, we decided to use the model without fixed effects but with random intercepts and slopes specific to subjects as *a reference model*. To test for a specific fixed effect, we compared the reference model with a model that included the fixed effect under study along with random intercepts and slopes for the participants. We first calculated Bayes factors for both the reference model (BF_r) and the fixed-effect model under test (BF_t). Both BFs compared the model to the Intercept (b0)-only model (model without random or fixed effects). We then calculated the BF_t/BF_r ratio. The ratio obtained (BF) greater than 1 indicated that the fixed-effects model was preferred, and BF less than 1 indicated that the reference model, i.e., the random-effects-only model, was preferred and that no notable fixed effect was present.

3. Results

The aim of our study was to examine the effect of typeface shape and age on reading. Table 1 shows the regression parameters for the fixed effects in the models tested. Large interindividual differences (large ICCs, i.e., intraclass correlation coefficients) were found in the eye-tracking measures—pupil size, number of regressive saccades, and reading time per character. ICCs were much lower for pleasantness ratings, text comprehension score and text memorisation score. For these variables, intrapersonal variability (differences between the eight typefaces) was much larger than interpersonal variability (differences between participants). However, on the legibility measures, intraindividual differences were much smaller than interindividual differences, suggesting that the reading skills of our participants were relatively diverse. Some were less fluent readers in general, i.e., across all eight texts, whereas the others were more fluent readers of all texts.

No interaction between age and typeface shape was observed on any of the measures examined, so we can next focus on the main effects of typeface shape and age on various reading measures.

3.1. Effect of typeface shape and reader age on pleasantness ratings and pupil size

First, we examined the relationship between the subjective and objective indicators of the emotional response to reading. We found that ratings of pleasantness predicted pupil size ($b = 0.01$, $\beta = 0.17$, $SE_b = 0.004$,

TABLE 1 Effect of typeface shape and age group on different reading parameters (pleasantness rating, pupil size, number of regressive saccades, reading speed, text comprehension and memorisation).

Corr. figure	Source of variability	*b*	SE_b	95% CI for *b* lower bound	95% CI for *b* upper bound	*t*	*df*	BF
3A	**Pleasantness rating**							
	ICC = 0.028, LRT(2) = 0.397, $p = 0.820$, BF for the full model = 889.50							
	Intercept	3.41	0.07	3.27	3.54	47.65	36.6	
	Typeface shape	−0.78	0.13	−1.05	−0.53	−5.96	118.3	5100.71
	Age	0.29	0.14	0.01	0.57	2.01	36.6	0.57
	Typeface shape × Age	0.23	0.26	−0.28	0.75	0.89	118.3	0.48
3B	**Pupil size (mm)**							
	ICC = 0.874, LRT(2) = 0.048, $p = 0.976$, BF for the full model = 0.08							
	Intercept	2.76	0.03	2.69	2.83	79.28	33.0	
	Typeface shape	−0.02	0.01	−0.04	−0.00	−2.11	238.5	0.66
	Age	−0.02	0.07	−0.15	0.12	−0.26	33.0	0.74
	Typeface shape × Age	−0.003	0.02	−0.04	0.03	−0.16	238.5	0.22
4A	**Reading time per character (ms)**							
	ICC = 0.802, LRT(2) = 13.3, $p < 0.001$, BF for the full model = 540.71)							
	Intercept	71.21	3.22	64.90	77.51	22.13	33.0	
	Typeface shape	4.85	1.72	1.48	8.22	2.82	33.0	5.20
	Age	30.98	6.43	18.37	43.59	4.82	33.0	408.44
	Typeface shape × Age	0.71	3.44	−6.03	7.45	0.21	33.0	0.29
4B	**Number of regressive saccades**							
	ICC = 0.587, LRT(2) = 0.412, $p = 0.814$, BF for the full model = 0.02							
	Intercept	86.48	7.94	70.93	102.0	10.90	33.0	
	Typeface shape	1.74	4.54	−7.16	10.6	0.38	207.6	0.18
	Age	8.47	15.87	−22.64	39.6	0.53	33.0	0.43
	Typeface shape × Age	−2.55	9.08	−20.35	15.3	−0.28	207.6	0.22
5A	**Text comprehension**							
	ICC = 0.135, LRT(2) = 10.00, $p = 0.007$, BF for the full model = 0.13							
	Intercept	0.87	0.03	0.82	0.92	32.66	35.2	
	Typeface shape	−0.09	0.04	−0.17	−0.00	−2.02	53.2	0.85
	Age	−0.09	0.05	−0.20	0.01	−1.76	35.2	0.67
	Typeface shape × Age	0.00	0.08	−0.16	0.17	0.05	53.2	0.23
5B	**Text memorisation**							
	ICC = 0.072, LRT(2) = 0.53, $p = 0.764$, BF for the full model = 0.02							
	Intercept	2.62	0.04	2.53	2.70	62.28	33.3	
	Typeface shape	−0.09	0.07	−0.22	0.04	−1.33	123.6	0.28
	Age	−0.08	0.08	−0.25	0.08	−0.97	33.3	0.28
	Typeface shape × Age	−0.05	0.13	−0.32	0.21	−0.40	123.6	0.23

Corr. figure = corresponding figure number. LRT shows whether including the random slope in the model (i.e., random effect of typeface shape on the outcome variable, in other words the variability of the typeface shape effect across participants) improves the fit of the model, with all other model parameters held constant. BF shows the Bayes factors for the tested models with fixed effects and random effects (random slopes and intercepts) against the reference models with random effects only. BF larger than 1 indicates that the model with fixed effects was preferred, that is that the examined fixed effect was present, and BF smaller than 1 indicates that the model with random effects only was preferred, i.e., that no notable fixed effect was present.

$t(111.6) = 3.32$), with strong support for the alternative hypothesis that the two variables are correlated (BF = 51.75). This suggests that readers responded emotionally to less or more pleasant typefaces. Pupil size was larger when reading typefaces were rated as more pleasant than when reading typefaces were rated as less pleasant (see also Figure 2).

Next, we examined the effects of typeface shape, age, and their interaction on pleasantness ratings and pupil size. Table 1 shows the results of linear mixed modelling and the Bayesian factors for each effect. BF values greater than 1 indicate evidence for the tested model, i.e., the model with both fixed and random effects, and values less

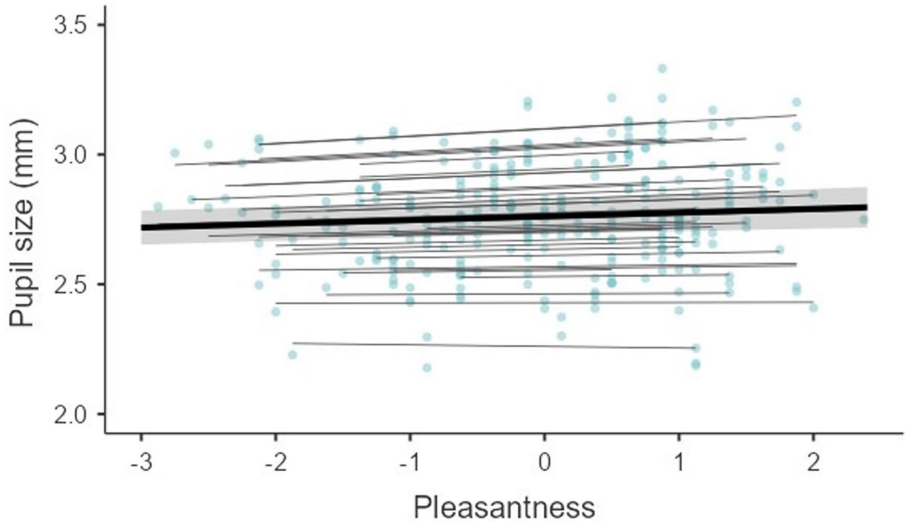


FIGURE 2
The relationship between pupil size and pleasantness ratings in different participants.

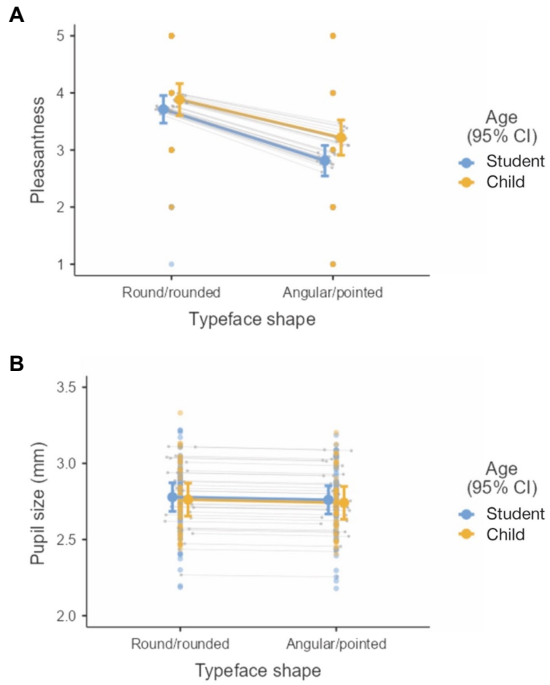


FIGURE 3
Effect of typeface shape and age group on measures of emotional response during reading: **(A)** Typeface pleasantness ratings and **(B)** pupil size. 95% confidence interval for means in different experimental conditions are shown. The same note also applies to Figures 4, 5.

than 1 indicate evidence for the reference model without fixed effects. In Table 1, we see that our data show very strong evidence for the effect of typeface shape on the pleasantness ratings. The extremely high BF value for the effect of typeface shape indicates that the model with fixed and random effects was preferred to the model with only random effects.

Figure 3A shows the pleasantness ratings and Figure 3B shows the pupil size under different experimental conditions. Typeface shape, as already mentioned, affected the ratings of pleasantness. In general, readers rated round/rounded typefaces as more pleasant than the angular/pointed ones. No such effect of typeface shape was observed in the pupil size data. Pupil size was only slightly larger for round/rounded typefaces than for angular/pointed typefaces (Figure 3B). Pleasantness ratings and pupil sizes were relatively similar in children and adults.

3.2. Effect of typeface shape and age group on reading speed and number of regressive saccades

Our data showed no evidence of a fixed effect of typeface shape on the number of regressive saccades; however, there was moderate evidence that typeface shape affected reading time per character (see Table 1; Figure 4A). The round/rounded typefaces had lower reading time per character than the angular/pointed typefaces. Thus, we can confirm that round/rounded typefaces allow for more fluent reading than angular/pointed typefaces.

An interesting discovery was that there was no fixed effect of age on the number of regressive saccades (see Table 1; Figure 4B). However, there was strong evidence for the effect of age on reading time per character (see Table 1; Figure 4A). Children read more slowly than adults.

3.3. Effect of typeface shape and reader age on text comprehension and text memorisation

Figures 5A,B show comprehension and memorisation scores under different experimental conditions, respectively. The analysis revealed no fixed effects of typeface shape, age, or their interaction on text comprehension or memorisation beyond the random effects. The BF values were in favour of the models with only random effects.

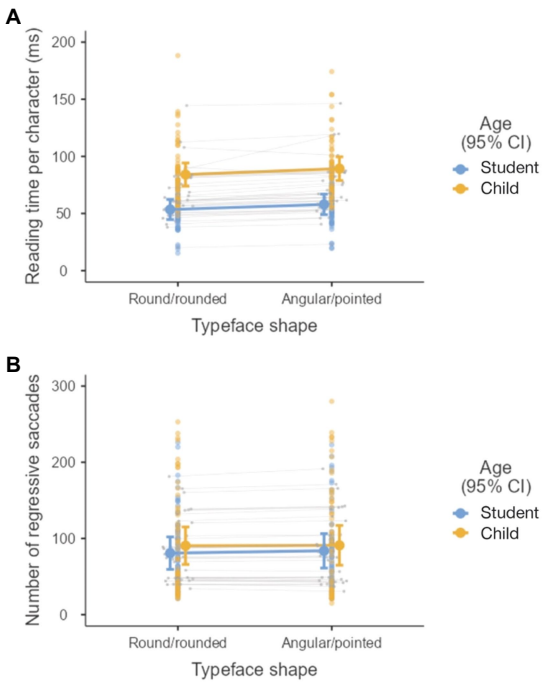


FIGURE 4
Effect of typeface shape and age group on measures of reading fluency: **(A)** Reading time per character and **(B)** number of regressive saccades.

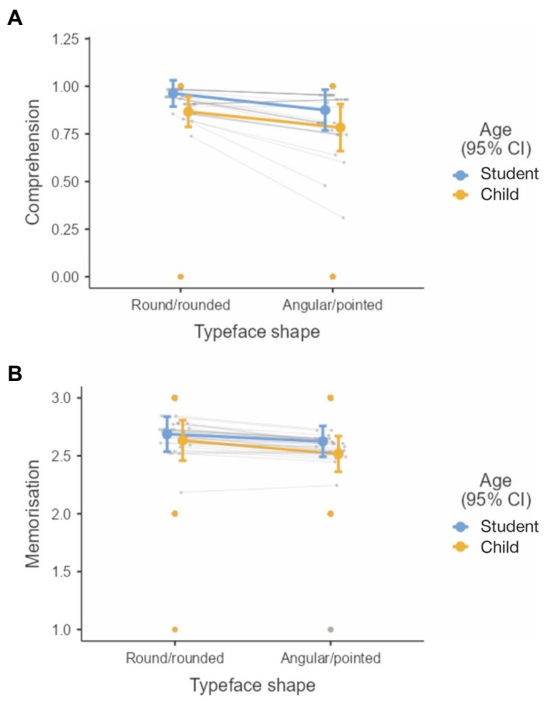


FIGURE 5
Effect of typeface shape and age group on reading performance measures: **(A)** Comprehension and **(B)** memorisation score.

4. Discussion with conclusion

The aim of our study was to (i) determine how the shape of the typeface (round/rounded vs. angular/pointed) affects the feelings of pleasantness of the typeface and pupil size, reading fluency (reading speed and number of regressive saccades), and reading performance (text comprehension and memorisation), and to (ii) examine whether the effect of the shape of the typeface is the same for younger (less experienced) and older (more experienced) readers.

With regard to the second aim of our study, the absence of the interaction between age and typeface shape in all models tested showed that the effect of the shape of the typeface was the same for both age groups. With regard to the first aim of our study, we can conclude that the only notable fixed effects were the main effect of typeface shape on pleasantness ratings and reading speed and the main effect of age on reading speed. Other measures were better explained by the regression model which included only random intercepts and slopes. There was a great deal of variability in the measures examined between participants, either in their average level of the measures or in the effect of typeface shape on the measures.

4.1. Effect of typeface shape on examined parameters of reading

The pleasantness of the typeface was tested with a hedonic scale in which readers rated how pleasant they found the typeface. Both children and adults found round/rounded typefaces more pleasing than angular/pointed typefaces (see Table 1; Figure 3A).

The effect of different typeface shape on subjective experience was also tested by measuring pupil size while reading different typefaces. The measured pupil size was slightly larger when reading round/rounded typefaces (this can also be seen in Figure 3B), which was also perceived as more pleasant by the readers. The rated typeface pleasantness correlated with pupil size (see Figure 2), supporting the assumption that the shape of the typeface influences the reader's emotional experience (Hess and Polt, 1960; Margareth et al., 2008; Wang et al., 2018). However, only the fixed effect of typeface shape on pleasantness ratings was convincing, whereas the effect of the typeface shape on pupil size was less remarkable. The analysis indicated that small differences in pupil size when reading round/rounded and angular/pointed typefaces could be a consequence of interindividual differences and could be attributed to random effects, i.e., to individual differences in pupil size and interindividual variability in the effect of typeface shape on pupil size. The fact that the effect of typeface shape on pupil size was smaller than effect of typeface shape on pleasantness ratings might indicate that factors other than the reader's emotional response, e.g., surprise (Preuschoff et al., 2011) or cognitive load (Gavas et al., 2018), influenced pupil size, although we tried to control for cognitive load by selecting texts with homogeneous difficulty.

We found that the shape of the typeface had an effect on one of the measures of reading fluency, i.e., reading speed. Readers read round/rounded typefaces faster than angular/pointed typefaces (see Table 1; Figure 4B). Typeface shape did not show notable effects on other measures of reading fluency and reading performance measures. It is possible that our comprehension and memory tests were not discriminative enough to detect differences between the two typeface shapes. Future studies should use psychometrically validated measures of memorisation (and comprehension) for the texts used in the study.

Based on the results of our study, we can conclude that the shape of the typeface can influence reading speed and feelings of pleasantness while reading. Round/rounded typeface shapes may be perceived as

more pleasant than angular/pointed shapes. Round/rounded typefaces also support reading fluency and allow readers to read faster.

4.2. Differences between age groups

Reading time varied by reader age – as expected, children read more slowly than adults, who tend to be more experienced readers (see Table 1; Figure 4B).

A somewhat surprising result was that the number of regressive saccades during the reading was not affected by age; that is, children did not have, on average, a higher number of regressive saccades than adults, as would be expected given their reading experience (see Table 1; Figure 4A). There were also no major differences between children and adults in text comprehension and memorisation. This can probably be explained by the fact that the texts used were not complex; they were easy to read and could be processed easily by both age groups. Future studies should examine how different reading parameters change with increasing text difficulty and whether age interacts with text difficulty in predicting reading performance and emotional and physiological responses during reading.

4.3. Limitations

Our study had several limitations. Even though we used texts of comparable difficulty, factors other than typeface shape may have influenced the results.

First, different participants might have responded differently to different texts. Their emotional response might depend on their specific interests (e.g., adults might respond differently to descriptions of animals than children). This could have increased the between-subject variability of the data.

Second, the typefaces we used differed in some characteristics that could affect reading parameters, such as typographic tonal density and overall character size: for example, we controlled for the x-height, but the different typefaces had different sizes of ascenders and descenders. As a result, the whiteness in the ascenders and descenders of the different typefaces was different, resulting in different line spacing, even though the leading was set to the same size (e.g., to 140%). Because of the different whiteness in ascenders and descenders, and because of the different counter shapes of the letters of different typefaces, the typographic tonal density value of texts in different typefaces will always be different, even if we unified the size of the x-height. Previous studies (Franken et al., 2015; Pušnik et al., 2016) have shown that factors such as these can affect reading speed and letter recognition. Future studies should investigate how manipulating a single feature of the typeface (e.g., only the shape of the strokes, while controlling for all other features, if possible) affects reading.

Third, the COVID-19 pandemic made it difficult to include larger samples, and the power of our complex statistical tests was low. Future studies should include larger samples.

Nevertheless, we believe that our results, although they should be considered preliminary, are quite informative because different measures of text processing were used, and although the fixed effects studied did not appear to be salient, all results pointed in the same direction – reading was more pleasant and fluent, and reading performance was minimally better with round/rounded typefaces compared to angular/pointed ones. Further studies will need to be conducted to provide more evidence, but our results suggest that it is important to consider typeface shape when examining reading or comparing findings from different studies.

4.4. Conclusion

Based on the results of our study, the use of round/rounded typefaces is recommended for the design of educational materials because readers or learners experience more pleasant feelings when reading than with angular/pointed typefaces. Using round/rounded typefaces also allows learners to read faster, which can have a positive impact on the learning process. The effect of typeface shape was similar in primary school pupils and university students, showing that the effect of typeface shape can be generalised across ages for simple texts. The typefaces with round/rounded shapes could be recommended for the design of educational materials used on the screen of a digital device for less experienced and more experienced readers. Such typefaces could make the learning process easier and more enjoyable.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Commission of the Faculty of Arts, University of Ljubljana. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

TM is a PhD student who has conducted research as part of her dissertation that she plans to publish in the indicated article. KM is her mentor and AP is her co-mentor. TM, KM, and AP jointly prepared a research plan for preliminary studies and for the final study. TM conducted all studies. TM reviewed and collected the literature. AP assisted TM in preparing data for statistical analysis. AP and TM jointly prepared the statistical analysis of the streams. All authors contributed to the article and approved the submitted version.

Funding

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding no. P5-0110 and no. P2-0213, and Infrastructural Centre RIC UL-NTF).

Acknowledgments

We would like to thank Ann Bessemans of Hasselt University for allowing us to use her original typeface Matilda for the purpose of our research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1107839/full#supplementary-material

References

Abukaber, A. A., and Lu, J. (2012). The optimum font size and type for students aged 9–12 reading Arabic characters on screen: a case study. *J. Phys. Conf. Ser.* 364, 1–14. doi: 10.1088/1742-6596/364/1/012115

Ali, A. Z. M., Wahid, R., Samsudin, K., and Dris, M. Z. (2013). Reading on the computer screen: does font type has effects on web text readability? *Int. Educ. Stud.* 6, 26–35. doi: 10.5539/ies.v6n3p26

Amare, N., and Manning, A. (2012). Seeing typeface personality: emotional responses to form as tone. Professional communication conference (IPCC). *2012 IEEE International Professional Communication Conference*, 1–9. doi: 10.1109/IPCC.2012.6408605

Bar, M., and Neta, M. (2007). Visual elements of subjective preference modulate amygdala activation. *Neuropsychologia* 45, 2191–2200. doi: 10.1016/j.neuropsychologia.2007.03.008

Bessemans, A. (2016a). Typefaces for children's reading. *TMG J. Media Hist.* 19, 1–9. doi: 10.18146/2213-7653.2016.268

Bessemans, A. (2016b). Matilda: a typeface for children with low vision. *Digit. Fonts Reading* 2016, 8–34. doi: 10.1142/9789814759540_0002

Beier, S., Sand, K., and Starrfelt, R. (2017). Legibility implications of embellished display typefaces. *Visible Lang.* 51, 112–133.

Beier, S., and Larson, K. (2013). How does typeface familiarity affect reading performance and reader preference? *Inf. Design J.* 20, 16–31. doi: 10.1075/idj.20.1.02bei

Bigelow, C. (2019). Typeface features and legibility research. *Vis. Res.* 165, 162–172. doi: 10.1016/j.visres.2019.05.003

Bjork, R. A., Dunlosky, J., and Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annu. Rev. Psychol.* 64, 417–444. doi: 10.1146/annurev-psych-113011-143823

Bjork, R. A., and Yue, C. L. (2016). Commentary: is disfluency desirable? *Metacogn. Learn.* 11, 133–137. doi: 10.1007/s11409-016-9156-8

Brath, R., and Banissi, E. (2016). Using typography to expand the design space of data visualization. *J Design Econ. Innov.* 2, 59–87. doi: 10.1016/j.sheji.2016.05.003

Brumberger, E. R. (2003). The rhetoric of typography: the persona of typeface and text. *Tech. Commun.* 50, 206–223.

Burnie, D. (2010). Ilustrirana enciklopedija živali [Illustrated Animal Encyclopedia]. Tržič: Učila International.

Cacali, E. (2016). The effect of font on vocabulary memorization. *Kwansei Gakuin Univ. Hum. Rev.* 21, 63–72.

Celhay, F., Boysselle, J., and Cohen, J. (2015). Food packages and communication through typeface design: the exoticism of exotypes. *Food Qual. Prefer.* 39, 167–175. doi: 10.1016/j.foodqual.2014.07.009

Childers, T. L., and Jass, J. (2002). All dressed up with something to say: effects of typeface semantic associations on brand perceptions and consumer memory. *J. Consum. Psychol.* 12, 93–106. doi: 10.1207/S15327663JCP1202_03

Choi, S., Jang, K. E., Lee, Y., Song, H., Cha, H., Lee, H. J., et al. (2018). Neural processing of lower-and upper-case text in second language learners of English: an fMRI study. *Lang. Cogn. Neurosci.* 33, 165–174. doi: 10.1080/23273798.2017.1384028

Crisinel, A. S., Jones, S., and Spence, C. (2012). The sweet taste of Maluma: cross modal associations between tastes and words. *Chemosens. Percept.* 5, 266–273. doi: 10.1007/s12078-012-9133-9

Cushing, C., and Bodner, G. E. (2022). Reading aloud improves proofreading (but using sans Forgetica font does not). *J. Appl. Res. Mem. Cogn.* 11, 427–436. doi: 10.1037/mac0000011

Davis, S. W. (2019). Say what? How the interplay of tweet readability and brand hedonism affects consumer engagement. *J. Bus. Res.* 100, 150–164. doi: 10.1016/j.jbusres.2019.01.071

Diemand-Yauman, C., Oppenheimer, D. M., and Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): effects of disfluency on educational outcomes. *Cognition* 118, 111–115. doi: 10.1016/j.cognition.2010.09.012

Dressler, E. (2019). *Understanding the Effect of Font Type on Reading Comprehension/Memory Under Time-Constrains*. Omaha: University of Nebraska at Omaha.

Ehsen, H., and Lupton, E. (1998). *Design Papers 5: Rhetorical Handbook: An Illustrated Manual for Graphic Designers*. Halifax, Nova Scotia, Design Division Nova Scotia College of Art and Design.

Franken, G., Podlesek, A., and Možina, K. (2015). Eye-tracking study of reading speed from LCD displays: influence of type style and type size. *J. Eye Mov. Res.* 8, 1–7. doi: 10.16910/jemr.8.1.3

Gallucci, M. (2020). GAMLj Suite for Jamovi. Available at: https://github.com/gamlj/gamlj. (Accessed October 02, 2022).

Gasser, M., Haffeman, J. B. M., and Tan, R. (2005). The influence of font type on information recall. *N. Am. J. Psychol.* 7, 181–188.

Gavas, R. D., Tripathy, S. R., Chatterjee, D., and Sinha, A. (2018). Cognitive load and metacognitive confidence extraction from pupillary response. *Cogn. Syst. Res.* 52, 325–334. doi: 10.1016/j.cogsys.2018.07.021

Geller, J., Davis, S. D., and Peterson, D. J. (2020). Sans Forgetica is not desirable for learning. *Memory* 28, 957–967. doi: 10.1080/09658211.2020.1797096

Gollely, M., and Guichard, N. (2011). The dilemma of flavor and color in the choice of packaging by children. *Young Consumers Insight Ideas Responsible Mark.* 12, 82–90. doi: 10.6007/IJARBSS/v4-i1/536

Haenschen, K., and Tamul, D. J. (2019). What's in a font?: ideological perceptions of typography. *Commun. Stud.* 71, 244–261. doi: 10.1080/10510974.2019.1692884

Halin, N. (2016). Distracted while Reading? Changing to a hard-to-read font shields against the effects of environmental noise and speech on text memory. *Front. Psychol.* 7, 1–11. doi: 10.3389/fpsyg.2016.01196

Hess, E. H., and Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science* 132:349. doi: 10.1126/science.132.3423.349

Hyndman, S. (2016). *Why Fonts Matter* London: Virgin Books, An Imprint of Ebury Publishing.

Ilic, U., and Akbulut, Y. (2019). Effect of disfluency on learning outcomes, metacognitive judgments and cognitive load in computer assisted learning environments. *Comput. Hum. Behav.* 99, 310–321. doi: 10.1016/j.chb.2019.06.001

ISO 12646 (2015). *Graphic technology — Displays for colour proofing — Characteristics*. Geneva, International Organization for Standardization.

ISO 3664 (2009). *Graphic Technology and Photography – Viewing Conditions* Geneva, International Organization for Standardization.

ISO 9241-303 (2011). *Ergonomics of Human-System Interaction — Part 303: Requirements for Electronic Visual Displays*. Geneva, International Organization for Standardization.

Jordan, T. R. (2017). What's in a typeface? Evidence of the existence of print personalities in Arabic. *Front. Psychol.* 8, 1–8. doi: 10.3389/fpsyg.2017.01229

Katzir, T., Hershko, S., and Halamish, V. (2013). The effect of font size on reading comprehension on second and fifth grade children: bigger is not always better. *PLoS One* 8, e74061–e74068. doi: 10.1371/journal.pone.0074061

Koch, E. B. (2012). Emotion in typographic design: an empirical examination. *Visible Lang.* 46, 206–228.

Labro, A. A., and Pocheptsova, A. (2016). Metacognition and consumer judgment: fluency is pleasant but disfluency ignites interest. *Curr. Opin. Psychol.* 10, 154–159. doi: 10.1016/j.copsyc.2016.01.008

Larson, K., Hazlett, R. L., Chaparro, B. S., and Picard, R. W. (2006). Measuring the Aesthetics of Reading. Human Computer Interaction Conference, People and Computers XX–Engage, Proceedings of HCI, 41–56.

Larson, K., and Picard, R. (2005). The Aesthetics of Reading. Available at: https://affect.media.mit.edu/pdfs/05.larson-picard.pdf (Accessed June 6, 2022).

Lewis, C., and Walker, P. (1989). Typographic influences on reading. *Br. J. Psychol.* 80, 241–257. doi: 10.1111/j.2044-8295.1989.tb02317.x

Macdonald, J. S. P., and Lavic, N. (2011). Visual perceptual load induces inattentional deafness. *Atten. Percept. Psychophys.* 73, 1780–1789. doi: 10.3758/s13414-011-0144-4

Mackiewicz, J. (2005). How to use five letterforms to gauge a typeface's personality: a research-driven method. *J. Tech. Writ. Commun.* 35, 291–315. doi: 10.2190/LQVL-EJ9Y-1LRX-7

Mano, H. (1997). Affect and persuasion: the influence of pleasantness and arousal on attitude formation and message elaboration. *Psychol. Mark.* 14, 315–335. doi: 10.1002/(SICI)1520-6793(199707)14:4<315::AID-MAR2>3.0.CO;2-C

Margareth, M. B., Miccoli, L., Escrig, M. A., and Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 602–607. doi: 10.1111/j.1469-8986.2008.00654.x

Maxwell, N. P., Perry, T., and Huff, M. J. (2022). Perceptually fluent features of study words do not inflate judgements of learning: evidence from font size, highlights, and sans Forgetica font type. *Metacogn. Learn.* 17, 293–319. doi: 10.1007/s11409-021-09284-6

McLean, R. (1997). *The Manual of Typography*. London: Thames and Hudson.

Mead, J. A., and Hardesty, D. M. (2018). Price font disfluency: anchoring effects on future Price expectations. *J. Retail.* 94, 102–112. doi: 10.1016/j.jretai.2017.09.003

Meyer, M. S., and Felton, R. H. (1999). Repeated Reading to enhance fluency: old approaches and new directions. *Ann. Dyslexia* 49, 283–306. doi: 10.1007/s11881-999-0027-8

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T., Ball, L. J., et al. (2015). Disfluent font don't help people solve math problems. *J. Exp. Psychol.* 144, e16–e30. doi: 10.1037/xge0000049

Morey, R., and Rouder, J. (2022). Bayes Factor: Computation of Bayes Factors for Common Designs. R Package Version 0.9.12–4.4 [Computer Software]. Available at: https://CRAN.R-project.org/package=BayesFactor (Accessed January 10, 2023).

Možina, K. (2003). *Knjižna Tipografija [Book Typography]*. Ljubljana: Univerza v Ljubljani.

Mueller, M. L., Dunlosky, J., Tauber, S. K., and Rhodes, M. G. (2014). The font-size effect on judgments of learning: does it exemplify fluency effects or reflect people's beliefs about memory? *J. Mem. Lang.* 70, 1–12. doi: 10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, K. S., and Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychon. Bull. Rev.* 20, 378–384. doi: 10.3758/s13423-012-0343-6

Ngo, M. K., Velasco, C., Salgado, A., Boehm, E., O'Neill, D., and Spence, C. (2013). Assessing crossmodal correspondences in exotic fruit juices: the case of shape and sound symbolism. *Food Qual. Prefer.* 28, 361–369. doi: 10.1016/j.foodqual.2012.10.004

Novemsky, N., Dhar, R., Schwarz, R., and Simonson, I. (2007). Preference fluency in choice. *J. Mark. Res.* 44, 347–356. doi: 10.1509/jmkr.44.3.347

Oppenheimer, D. M., and Frank, M. C. (2008). A rose in any other font would not smell as sweet: effects of perceptual fluency on categorization. *Cognition* 106, 1178–1194. doi: 10.1016/j.cognition.2007.05.010

Pečjak, S., and Kramarič, M. (2018). *Bralne Strategije. Primeri Besedil za 4. Razred [Reading Strategies. Text Examples for the 4th Grade of Primary School]*. Ljubljana: Rokus Klett.

Petit, O., Velasco, C., Cheok, A. D., and Spence, C. (2015). Consumer Sensory Neuroscience in the Context of Food Marketing. ACE 2015 Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology, 49, 1–4.

Pieger, E., Mangelkamp, C., and Bannert, M. (2016). Metacognitive judgments and disfluency – does disfluency lead to more accurate judgments, better control, and better performance? *Learn. Instr.* 44, 31–40. doi: 10.1016/J.LEARNINSTRUC.2016.01.012

Pikulski, J. J., and Chard, D. J. (2005). Fluency: bridge between decoding and reading comprehension. *Read. Teach.* 58, 510–519. doi: 10.1598/RT.58.6.2

Piqueras-Fiszman, B., Ares, G., and Varela, P. (2011). Semiotics and perception: do labels convey the same messages to older and younger consumers? *J. Sens. Stud.* 26, 197–208. doi: 10.1111/j.1745-459X.2011.00336.x

Piqueras-Fiszman, B., Velasco, C., Salgado-Montejo, A., and Spence, C. (2013). Using combined eye tracking and word association in order to assess novel packaging solutions: a case study involving jam jars. *Food Qual. Prefer.* 28, 328–338. doi: 10.1016/j.foodqual.2012.10.006

Preuschoff, K., Hart, B., and Einhauser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Front. Neurosci.* 5:115. doi: 10.3389/fnins.2011.00115

Price, J., McElroy, K., and Martin, N. J. (2016). The role of font size and font style in younger and older adults predicted and actual recall performance. *Aging Neuropsychol. Cognit.* 23, 366–388. doi: 10.1080/13825585.2015.1102194

Tobii Pro (2017). *Tobii Studio User's Manual (Version 3.4.8)* Stockholm: Tobii AB.

Pušnik, N., Podlesek, A., and Možina, K. (2016). Typeface comparison — does the x-height of lower-case letters increased to the size of upper-case letters speed up recognition? *Int. J. Ind. Ergon.* 54, 164–169. doi: 10.1016/j.ergon.2016.06.002

Raden, A. Z. M., and Qeis, M. I. (2019). Song and typography: expressing the lyrics visually through lyrical typography. *Int. J. Sci. Technol. Res.* 8, 61–64.

Rummer, R., Schweppe, J., and Schwede, A. (2016). Fortune is fickle: null-effects of disfluency on learning outcomes. *Metacogn. Learn.* 11, 57–70. doi: 10.1007/s11409-015-9151-5

Salgado-Montejo, A., Alvarado, J. A., Velasco, C., Salgado, C. J., Hasse, K., and Spence, C. (2015). The sweetest thing: the influence of angularity, symmetry, and the number of elements on shape-valence and shape-taste matches. *Front. Psychol.* 6, 1–17. doi: 10.3389/fpsyg.2015.01382

Sanchez, C. A., and Nayor, J. S. (2018). Disfluent presentations lead to the creation of more false memories. *PLoS One* 13, e0191735–e0191738. doi: 10.1371/journal.pone.0191735

Shaikh, A. D., Chaparro, B. S., and Fox, D. (2006). Perception of fonts: perceived personality traits and uses. *Usability News* 8.

Song, H., and Schwarz, N. (2008). If It's hard to read, It's hard to do–processing fluency affect effort prediction and motivation. *Psychol. Sci.* 19, 986–988. doi: 10.1111/j.1467-9280.2008.02189.x

Spence, C., and Deroy, O. (2012). Crossmodal correspondences: innate or learned? *i-Perception* 3, 316–318. doi: 10.1068/i0526ic

Su, N., Li, T., Zheng, J., Hu, X., Fan, T., and Luo, L. (2018). How font size affect judgments of learning: simultaneous mediating effect of item-specific beliefs about fluency and moderating effect of beliefs about font size and memory. *PLoS One* 13, 1–14. doi: 10.1371/journal.pone.0200888

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., and Garry, M. (2020). Disfluent difficulties are not desirable difficulties: the (lack of) effect of sans forgetica on memory. *Memory* 28, 850–857. doi: 10.1080/09658211.2020.1758726

The Jamovi Project. (2019). Jamovi [Computer Software]. Available at: https://www.jamovi.org (Accessed January 10, 2023).

Tsonos, D., and Kouroupetroglou, G. (2011). Modelling reader's emotional state response on document's typographic elements. *Adv. Hum. Comput. Interact.* 2011, 1–18. doi: 10.1155/2011/206983

Turoman, N., Velasco, C., Chen, Y., Huang, P., and Spence, C. (2018). Symmetry and its role in the crossmodal correspondence between shape and taste. *Atten. Percept. Psychophys.* 80, 738–751. doi: 10.3758/s13414-017-1463-x

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., et al. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychon. Bull. Rev.* 28, 813–826. doi: 10.3758/s13423-020-01798-5

Velasco, C., Beh, E. J., Le, T., and Marmolejo-Ramos, F. (2018a). The shapes associated with the concept of 'sweet and sour' foods. *Food Qual. Prefer.* 68, 250–257. doi: 10.1016/j.foodqual.2018.03.012

Velasco, C., Hyndman, S., and Spence, C. (2018b). The role of typeface curvilinearity on taste expectations and perception. *Int. J. Gastron. Food Sci.* 11, 63–74. doi: 10.1016/j.ijgfs.2017.11.007

Velasco, C., Salgado-Montejo, A., Marmolejo-Ramos, F., and Spence, C. (2014). Predictive packaging design: tasting shapes, typefaces, names, and sounds. *Food Qual. Prefer.* 34, 88–95. doi: 10.1016/j.foodqual.2013.12.005

Velasco, C., Woods, A. T., Deroy, O., and Spence, C. (2015a). Hedonic mediation of the crossmodal correspondence between taste and shape. *Food Qual. Prefer.* 41, 151–158. doi: 10.1016/j.foodqual.2014.11.010

Velasco, C., Woods, A. T., Hyndman, S., and Spence, C. (2015b). The taste of typeface. *i-Perception* 6, 1–10. doi: 10.1177/2041669515593040

Velasco, C., Woods, A. T., Marks, L. E., Cheok, A. D., and Spence, C. (2016). The semantic basis of taste-shape associations. *PeerJ* 4, 1–23. doi: 10.7287/PEERJ.PREPRINTS.1366

Wang, C. A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., and Munoz, D. P. (2018). Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task. *Front. Neurol.* 9, 1–13. doi: 10.3389/fneur.2018.01029

Weissgerber, S. C., and Reinhard, M. (2017). Is disfluency desirable for learning? *Learn. Instr.* 49, 199–217. doi: 10.1016/j.learninstruc.2017.02.004

Wetzler, E. L., Pyke, A. A., and Werner, A. (2021). Sans Forgetica is not the "font" of knowledge: disfluent fonts are not always desirable difficulties. *SAGE Open* 11:215824402110566. doi: 10.1177/21582440211056624

Wilkins, A., Cleave, R., Grayson, N., and Wilson, L. (2009). Typography for children may be inappropriately designed. *J. Res. Read.* 32, 402–412. doi: 10.1111/j.1467-9817.2009.01402.x

Woods, R. J., Davis, K., and Scharff, L. F. V. (2005). Effects of typeface and font size on legibility for children. *Am. J. Psychol. Res.* 1, 86–102.

Wu, R., Shah, E. D., and Kardes, F. R. (2019). "The struggle isn't real": how need for cognitive closure moderates inference from disfluency. *J. Bus. Res.* 109, 585–594. doi: 10.1016/j.jbusres.2019.03.042

Yue, R. L., Castel, A. D., and Bjork, R. A. (2013). When disfluency is––and is not––a desirable difficulty: the influence of typeface clarity on metacognitive judgments and memory. *Mem. Cogn.* 41, 229–241. doi: 10.3758/s13421-012-0255-8

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 June 2023
DOI 10.3389/fpsyg.2023.1076501

OPEN ACCESS

EDITED BY
Marijan Palmovic,
University of Zagreb, Croatia

REVIEWED BY
Susana Araújo,
University of Lisbon, Portugal
Dongchuan Yu,
Southeast University, China
Angela Jocelyn Fawcett,
Swansea University, United Kingdom

*CORRESPONDENCE
Christoforos Christoforou
✉ christoc@stjohns.edu
Timothy C. Papadopoulos
✉ papadopoulos.timothy@ucy.ac.cy

RECEIVED 21 October 2022
ACCEPTED 31 May 2023
PUBLISHED 20 June 2023

CITATION
Christoforou C, Theodorou M, Fella A and Papadopoulos TC (2023) RAN-related neural-congruency: a machine learning approach toward the study of the neural underpinnings of naming speed.
Front. Psychol. 14:1076501.
doi: 10.3389/fpsyg.2023.1076501

RAN-related neural-congruency: a machine learning approach toward the study of the neural underpinnings of naming speed

Christoforos Christoforou[1]*, Maria Theodorou[2], Argyro Fella[3] and Timothy C. Papadopoulos[4]*

[1]Division of Computer Science, Mathematics and Science, St. John's University, New York, NY, United States, [2]Independent Researcher, New York, NY, United States, [3]Department of Education, University of Nicosia, Nicosia, Cyprus, [4]Department of Psychology and Center for Applied Neuroscience, University of Cyprus, Nicosia, Cyprus

Objective: Naming speed, behaviorally measured via the serial Rapid automatized naming (RAN) test, is one of the most examined underlying cognitive factors of reading development and reading difficulties (RD). However, the unconstrained-reading format of serial RAN has made it challenging for traditional EEG analysis methods to extract neural components for studying the neural underpinnings of naming speed. The present study aims to explore a novel approach to isolate neural components during the serial RAN task that are (a) informative of group differences between children with dyslexia (DYS) and chronological age controls (CAC), (b) improve the power of analysis, and (c) are suitable for deciphering the neural underpinnings of naming speed.

Methods: We propose a novel machine-learning-based algorithm that extracts spatiotemporal neural components during serial RAN, termed RAN-related neural-congruency components. We demonstrate our approach on EEG and eye-tracking recordings from 60 children (30 DYS and 30 CAC), under phonologically or visually similar, and dissimilar control tasks.

Results: Results reveal significant differences in the RAN-related neural-congruency components between DYS and CAC groups in all four conditions.

Conclusion: Rapid automatized naming-related neural-congruency components capture the neural activity of cognitive processes associated with naming speed and are informative of group differences between children with dyslexia and typically developing children.

Significance: We propose the resulting RAN-related neural-components as a methodological framework to facilitate studying the neural underpinnings of naming speed and their association with reading performance and related difficulties.

KEYWORDS

EEG, fixation-related potential (FRP), neural-congruency, machine learning, dyslexia, rapid automatized naming (RAN), RAN-related neural-congruency, eyetracking

1. Introduction

Rapid Automatized Naming (RAN), broadly defined as the ability to name as fast as possible visually presented stimuli such as colors, objects, digits, and letters (Kirby et al., 2010), is one of the most examined underlying cognitive factors of reading development and reading difficulties (RD). Indeed, since the original work by Denckla (1972), there has been an ongoing effort to explain the rather complex relationship between RAN and reading (e.g., Kirby et al., 2010; Araújo et al., 2015) across different ages (e.g., Landerl and Wimmer, 2008; Moll et al., 2009) and ability (e.g., Wolf and Bowers, 1999; Papadopoulos et al., 2009a; Torppa et al., 2013) groups and languages (e.g., Georgiou et al., 2012; Moll et al., 2014; Papadopoulos et al., 2021), focusing on group and individual differences. This effort has been based on behavioral/cognitive and neuroimaging data evidence. With regard to the former, two research approaches have been used, a componential (e.g., Georgiou et al., 2014) and a correlational (e.g., Papadopoulos et al., 2016) approach. With regard to the latter, data derived from fMRI studies (e.g., Cummine et al., 2015; Al Dahhan et al., 2020), electroencephalography (EEG) methods (e.g., Bakos et al., 2020) or more recently Fixation-Related Potentials (FRPs; e.g., Christoforou et al., 2021a). Although the evidence shows that RAN predicts reading performance and that RAN taps into universal cognitive mechanisms involved in reading (Papadopoulos et al., 2021), little is known about which neural components of RAN could better distinguish children with reading difficulties from typically developing peers. Thus, the present study aims to take this research line further: to produce and test methods that isolate the most informative neural components for RAN, suitable for deciphering group or individual differences in naming speed.

Based on correlational data, behavioral or cognitive research has repeatedly confirmed that RAN relates to reading because both tasks require serial processing and lexical access (e.g., Georgiou et al., 2013; Logan and Schatschneider, 2014). Also, it has been shown that RAN exerts direct effects on reading fluency only when oral reading fluency is the outcome measure (Georgiou et al., 2013; van den Boer et al., 2014; Papadopoulos et al., 2016), suggesting that articulation is essential for the RAN-reading relationship. Correlational research has also concluded that universal cognitive mechanisms such as working memory, attention, and processing speed are distal "common cause" processes to the RAN-reading relationship (Papadopoulos et al., 2016). Indeed, it is well-established that processing speed partly mediates the RAN-reading relationship (e.g., Bowey et al., 2005; Georgiou et al., 2012; Liao et al., 2015). Also, working memory is necessary because of the effortful nature of cognitive control required to perform naming speed tasks successfully, as it also occurs with word reading (Jacobson et al., 2011) or reading comprehension (e.g., Leong et al., 2008; Kendeou et al., 2012). Likewise, for serial processing to occur successfully, attention must be disengaged from naming a current item and directed to the next (Altani et al., 2017). Recent studies using eye-tracking methodology have verified the influential role of attention on RAN performance (e.g., Jones et al., 2009; Kuperman et al., 2016). Finally, evidence shows that speech production planning processes are also involved before articulation (e.g., Araújo et al., 2021).

These findings are further validated through research examining the unique contribution of articulation and pause time and what these components share with cognitive mechanisms such as the above. Since oral reading fluency and rapid naming require articulation alongside processing speed, the unique contribution of articulation time is justified (Georgiou et al., 2012). In turn, attention shifting, required as the participants move from one stimulus to another in a short time, is encapsulated in pause time (Wolf and Bowers, 1999; Georgiou et al., 2014), providing quick access to phonological codes or semantics in long-term memory (Rijthoven et al., 2018). Developmental data corroborate this evidence, as the contribution of pause time for typical readers decreases with time as they rely on larger orthographic units to read fluently (e.g., Georgiou et al., 2014). In contrast, pause time continues to explain significant variance in children with reading difficulties because of the deficits in accessing phonological codes experienced by this ability group (Ziegler et al., 2003; Araújo et al., 2011). Likewise, other processes have also been investigated, including multi-element sequence processing, coordinating rapid serial eye movements, and speech production planning processes of successive items (e.g., Gordon and Hoedemaker, 2016; Henry et al., 2018). However, studying these dimensions of the RAN-reading relationship was beyond the scope of the present paper to further explore their contribution.

Neurocognitive research has verified such findings with adults or typically developing and same-age poor readers, based on neuroimaging data. For example, Cummine et al. (2015), using functional magnetic resonance imaging (fMRI) with an adult group of typical readers, reported that RAN and reading rely on highly similar neural regions and that the RAN–reading relationship is driven by motor/serial processing. Likewise, Al Dahhan et al. (2020), using fMRI and eye-tracking methods, concluded that compared to typically achieving readers, readers with reading difficulties performed poorer in naming speed tasks. They had more extended articulation and pause times, longer fixation durations, and more regressions, resulting in decreased performance. This deficient processing was also reflected in greater bilateral activation and recruited additional regions involved with memory, namely the amygdala and hippocampus. Moreover, when the RAN-letter stimuli were visually or phonologically similar, adult readers showed higher activation in the amygdala and hippocampus, irrespective of their group (dyslexics vs. controls).

Furthermore, studies using eye-tracking (e.g., Easson et al., 2020) or electroencephalography (EEG) methods (e.g., Bakos et al., 2020) have provided additional evidence. Their results have focused on the RAN's constituent components or the neurophysiological differences between children with reading difficulties and typically developing readers. For example, Easson et al. (2020) revealed significant contributions of fixation duration and saccade count to the prediction of naming speed performance. In addition, Bakos et al. (2020) showed that EEG activity differed between 10-year-olds with reading difficulties and their counterparts at around 300 ms after stimulus presentation. This difference was evident in the left-occipital-temporal P2 component and was statistically significantly correlated to RAN performance, albeit small $r(72) = 0.24$, $p < 0.04$.

More recently, Christoforou et al. (2021a,b) combined EEG and eye-tracking recordings to examine the underlying factors elicited during the serial Rapid-Automatized Naming (RAN) task that may differentiate between children with reading difficulties

and chronological age controls (CAC). In doing so, the authors extracted fixation-related potentials (FRPs) under phonologically similar (rime-confound) or visually similar (resembling lowercase letters) and dissimilar (non-confounding and discrete uppercase letters, respectively) RAN tasks. As a result, the authors reported significant differences in FRP amplitudes between RD and CAC groups under phonologically similar and non-confounding conditions. These differences were evident in a cluster emerging around 128–170 ms in the frontal and occipital channels and between 80–160 ms for the rime-non-confusable and the rime-confusable RAN-letter tasks, respectively. However, no differences were observed in the case of the visual conditions. Moreover, regression analysis showed that the average amplitude of the extracted components significantly predicted RAN performance.

That research investigating the RAN-reading relationship concludes that RAN is a proxy for reading because it exerts similar processes to the neural reading system in the brain's left hemisphere is not a surprise. This system includes a ventral stream that helps the reader recognize the words and their semantic meaning (Norton and Wolf, 2012) and a dorsal stream which connects sub-lexical phonological codes to orthographic representations (Pugh et al., 2001; Price, 2012). Deficits with the processing of grapheme-phoneme correspondence, in turn, are reflected in lower activation in the dorsal stream. Likewise, automatic visual word recognition deficits are reflected in lower activation in the ventral occipital-temporal system (Richlan et al., 2011). Consequently, when performing RAN tasks whose stimuli exhibit phonological or visual similarities, this network tends to suffer more (Al Dahhan et al., 2020; Christoforou et al., 2021a).

Despite these efforts to isolate the neural components for RAN, the findings about the different brain regions identified do not tell the complete story of the RAN-reading relationship. For example, the evidence does not tell us why group or individual differences exist or which are the most informative components that could help replicate such findings with groups of different ages, varying cognitive or linguistic abilities, or language. We argue that more advanced methods are needed to isolate the most informative components, explain group differences and improve the power of analysis. FRPs, for example, can be used as markers of ability, but we need more specific attributes to carry more information about group differences.

In recent years, machine learning approaches are becoming more prominent in analyzing EEG signals and studying neurocognitive processes. Machine learning allows an algorithm to isolate neural components that "optimally" characterize group differences under different conditions; therefore having the potential to detect more informative neural components than traditional EEG analysis methods (i.e., average ERPs and traditional frequency-band analysis). Several machine learning approaches have been proposed to overcome the methodological constraints of traditional EEG analysis methods. For example, single-trial correlation analysis (Christoforou et al., 2013) was developed to identify associations between continuous behavioral measures and concurrent neuronal activity. It was applied to exploring the neural underpinnings for the Stimulus Presentation Modality Effects in Traumatic-Brain-Injury treatment protocols. In the context of spatial cognition, a Common Spatial Pattern (CSP)-based single-trial analysis algorithm was proposed (Christoforou et al., 2018) to disambiguate the neural basis of two spatial-cognition processes, namely Perspective Taking and Mental Rotation. Machine-learning-based algorithms have been also proposed for decoding neural activity during complex interactions, such as consuming video and music context, toward studying user's preferences and affective state (Dmochowski et al., 2012; Christoforou et al., 2017; Christoforou and Theodorou, 2021), as well as in other decision making (Philiastides and Sajda, 2005). In the context of reading and reading disorders, machine learning algorithms were proposed for detecting informative neural components during performing a phoneme elision task (Christoforou et al., 2022a,b), and classifying dyslexic from non-dyslexic participants during resting EEG (Rezvani et al., 2019). However, most of the proposed machine-learning approaches assume some prior domain knowledge of the spatial and temporal characteristics of the sought EEG components. They also require experimenter-controlled time-locked events (i.e., stimulus onset), and are typically limited to within-participant comparisons because of the large inter-subject variability in the EEG signals (Christoforou et al., 2010). These methodological requirements do not hold in the case of the serial RAN task which makes their direct application to RAN ineffective.

In the present study, we explore a novel machine-learning-based approach to isolate neural components informative of group differences between children with dyslexia and controls during the serial RAN. Our approach overcomes many methodological challenges of traditional methods which enable us to extract differential spatiotemporal profiles of neural components among children with dyslexia and controls during RAN and in the absence of experimenter-controlled time-locked events. Our method first formulates an optimization problem for extracting EEG components based on the Neural-congruency hypothesis. This relates to the premise that neural activity elicited during a cognitive task is similar (i.e., congruent) among participants that have mastered the task but less congruent otherwise (Christoforou et al., 2021b; Christoforou and Theodorou, 2021). Subsequently, our approach optimally combines the resulting components to identify neural differences between children with dyslexia and controls. We demonstrate the ability of our approach to extract informative neural components on a real EEG dataset involving children with dyslexia and controls of ages 9 and 12 (i.e., 3rd and 6th grade). Moreover, we examine the predictive power of the resulting components under a set of phonological and visual confounding RAN tasks. Importantly, our proposed analysis approach serves as a novel methodological framework for studying the neural underpinnings of cognitive processing in children under the serial RAN, on which traditional analysis methods have proven inadequate.

2. Materials and methods

2.1. Experimental task and data collection

The data we used in this study were collected as part of a broader project aiming to identify the neural underpinnings of dyslexia in children. In this section, we briefly describe the key parameters of the RAN experimental task and the data collection

procedure relevant to our analysis; we refer to Christoforou et al. (2021a) for full details on the data collection apparatus.

2.1.1. Participants

Participants were recruited from Grades 3 and 6 from inner-city public elementary schools in Cyprus. A total of 60 children (36 boys, 24 girls, age range = 7.6 through 12.1 years) participated in the study; all children were native Greek speakers. Two groups were formed from this sample: a group of children with dyslexia (DYS) and a chronological-age control group (CAC), based on a stepwise group selection process (see Christoforou et al., 2021a) using a lenient cutoff threshold on their reading fluency scores. Particularly, thirty Grade 3 and Grade 6 children (19 males; mean age = 9.6, SD = 1.5) who scored at least one standard deviation below their respective age group mean on the reading fluency tasks (word reading fluency and nonword reading fluency; ERS-AB; Papadopoulos et al., 2009c) and within the average range on verbal (Vocabulary Wechsler Intelligence Scale for Children—Third Edition; Greek standardization: Georgas et al., 1997) and non-verbal ability tasks (Nonverbal Matrices from the Cognitive Assessment System; Niglieri and DAS, 1997; Greek standardization: Papadopoulos et al., 2009b) were included in the DYS group. Another group of 30 children (17 males; mean age = 9.92 years, SD = 1.62) were randomly chosen from the same classes and were matched to the DYS group on chronological age and gender. Groups did not differ in age, $F(1,58) = 0.22$, ns, gender, $\chi 2$ (1, $N = 60$) = 0.28, ns, and the verbal and non-verbal ability measures, Wilks $\lambda = 0.98$, $F(2,57) = 0.70$, ns. Parental consent and school consent were obtained before to each assessment. The study was carried out per the Cyprus National Bioethics Committee recommendations (EEBK/EP/2011/10). It also received approval from the Ministry of Education and Culture, Cyprus (#7.15.01.27/17).

2.1.2. Serial RAN task

A computerized version of the serial Rapid Automatized Naming task was adapted from the work of Jones et al. (2008) to allow for simultaneous recordings of EEG and eye-tracking measurements during the experiment. The RAN task comprises four letter-matrix stimuli each encapsulating one of four conditions that differed by the degree of visual and phonological confusability among letters. In particular, the conditions encoded by the stimuli were *rime-confusable* (Condition 1), *rime non-confusable* (Condition 2), *visual confusable* (Condition 3) and *visual-non-confusable* (Condition 4). In the *rime-confusable* condition, pairs of letters that are phonologically confusable in the Greek alphabet (i.e., β-θ, ε-υ; beta-theta, epsilon-upsilon) were presented adjoining each other. In the *rime non-confusable* conditions, the pairs were disjoined (i.e., β–ε, β–υ, θ–ε, θ–υ, beta-epsilon, beta-upsilon, theta-epsilon, theta-upsilon). In the *visual-confusable* condition, pairs of letters that are visually confusable in the Greek alphabet (i.e., ζ-ξ, ρ-φ; zeta-xi, rho-phi), were presented adjoining each other. The visual similarity was removed in the *visual-non-confusable* condition by using the corresponding capital form of the letters (i.e., Z-Ξ, P-Φ). Each letter-matrix stimulus was organized in five rows and ten columns. Participants were shown the corresponding letter-matrix for each condition and asked to name each letter aloud, reading from left to right and from top to bottom, as fast and as accurately as possible. Before each conditioned stimulus, a fixation cross was displayed on the screen to prime participants to focus their eye-gaze at the center. The experimenter monitored the participants during the experiment and pressed the SPACE bar button the moment the participants name aloud the last letter of the letter-matrix. The experimenter also controlled the transition from one condition to the other. A schematic representation of the experimental task and example stimuli is shown in **Figure 1**.

2.1.3. EEG and eye-tracking data collection during RAN

All participants had to perform the serial RAN task while simultaneous eye-tracking and EEG measurements were collected during the session. Eye-tracking data were collected using the EyeLink 1000 Plus eye-tracker (SR Research, Kanata, ON, Canada) at a 1,000 Hz sampling rate. Eye fixations and saccade events were automatically detected and recorded by the EyeLink parser along with the raw gaze data. The stimuli were presented on a Dell Precision T5500 workstation with an ASUS VG-236 monitor (1,920 × 1,080, 120 Hz, 52 × 29 cm) at a viewing distance of 60 cm. A chin rest was used to maintain the participant's head proper positioning and to improve measurement stability. A nine-point calibration session was performed prior to experiment to establish a correct mapping to screen coordinates. EEG data were collected using a BioSemi Active-two system (BioSemi, Amsterdam, Netherlands) at a sampling rate of 256 Hz. Before to the experimental session, a 64-electrode cap was fitted to the participants, following the 10/20 system. The DC offset of all sensors was kept below 20 mV using electro gel. To align the stimulus presentation time to the EEG and eye-tracking signal streams, event markers were sent to each device at the beginning and ending of each condition. Specifically, the event markers were sent to the trigger channel of the EEG amplifier via parallel port TTL signals and to the eye-tracking recorder via direct ethernet event logs. Eye-tracking data were collected for each participant in a separate data files; specifically an EDF file format (default of the eye-link trackers), and the EEG data in a BDF file format.

2.2. EEG and eye-tracking data pre-processing

2.2.1. EEG data pre-processing

All EEG data preprocessing was implemented using custom python code and the MNE python library. Preprocessing of the EEG signals was performed separately on the recordings of each participant. First the raw, continuous EEG data and the corresponding trigger channel were loaded from the data file (i.e., BDF data format file) using MNE library functions. Once loaded, all EEG data channels were re-referenced to the average channel. A 0.5 Hz high pass filter was used to remove DC drifts and a notch filter at 50 Hz and 100 HZ was used to reduce the power-line noise interferences. Markers in the EEG trigger channel were used to identify the timestamp of the beginning and end of each stimulus trial (i.e., each condition). Four EEG sub-segments were generated for each condition, each segment spanning 2 s before each stimulus onset to 2 s after the trial

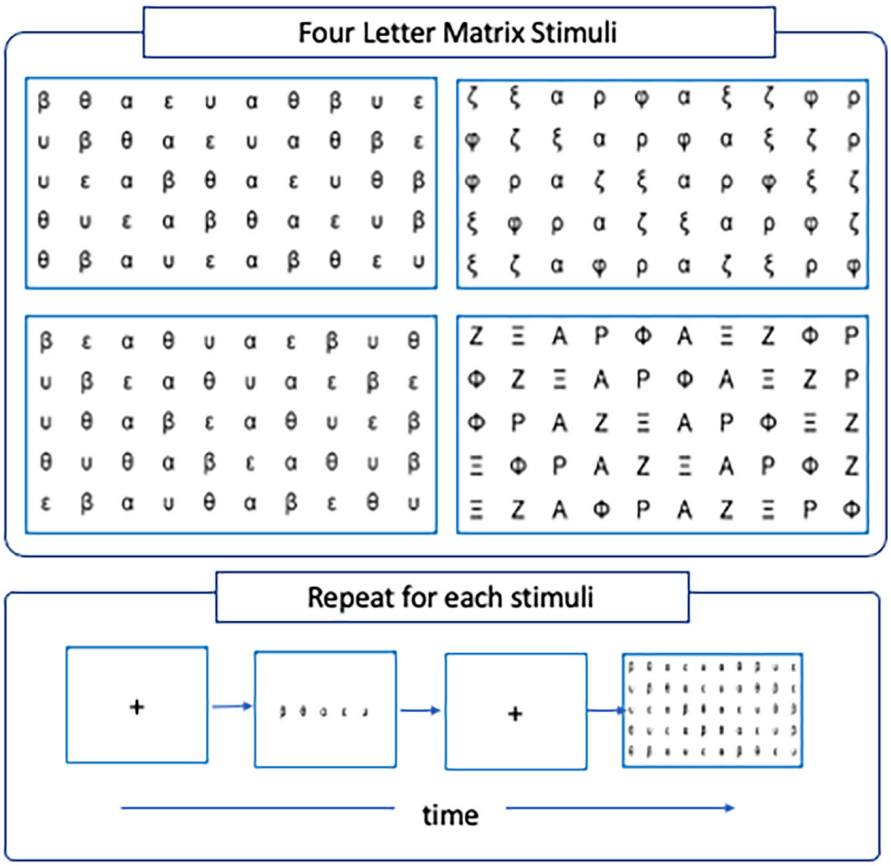


FIGURE 1
(Top) The four letter-matrix stimuli used in the serial RAN encapsulating the four experimental conditions, *Rime-confusable* ***(top-left)***, *Rime-non-confusable* ***(bottom-left)***, *Visual-confusable* ***(top-right)***, and *Visual-non-confusable* ***(bottom-right)***. ***(Bottom)*** Schematic representation of the serial-RAN trials, repeated for each one of the letter-matrix stimuli.

conclusion (i.e., after the participant finished reading all letters in the letter-matrix). The baseline amplitude of each segment (i.e., activity from −200 ms to zero) was subtracted from each segment. After basic EEG pre-processing, we have an EEG segment $EEG_{p,c}$ for each participant p and condition *c*, each representing the entire EEG recording of that participant reading the entire letter-matrix of that conditions. It is important to note that the duration of each EEG segment varies from condition-to-condition and from participant to participant, as each participant took a different time to complete the reading of each matrix.

2.2.2. Eye-tracking data preprocessing

Preprocessing of the eye-tracking data was also performed separately for each participant. The eye-fixation data and the corresponding event logs were loaded using the PyGaze Analyzer python library. Information on the event logs was used to determine the timestamp of the beginning and ending of each stimulus trial. Each eye-fixation data point comprised an absolute timestamp, the x-y screen coordinates of the fixation and durations in milliseconds. The set of eye-fixation points was grouped into four subsets, one for each of the four conditions. Each fixations subset comprised those fixations whose timestamp fell within the time window spanning the beginning and end of the stimulus presentation of that condition. The timestamp of the stimulus onset event of each condition (as recorded in the eye-tracking data) was subtracted from the timestamp of each fixation within that condition's fixation subset to achieve temporal alignment between the fixation data to the EEG data. Therefore, each fixation's timestamp is now relative to the onset of the stimulus.

2.3. Generating single-trial fixation-related potentials

A particular challenge in analyzing EEG signals obtained during the serial RAN is the lack of experimenter-controlled, time-locked trials necessary to extract Event-related Potentials. As such, we opted to explore the neural activity time-locked to the onset of eye fixations; this activity is referred to as single-trial Fixation-related Potential (sFRP). To extract the sFRP, we integrate information from eye-tracking and EEG measurements. In particular, the onset of each eye-fixation's timestamp in the dataset is used as a temporal marker to epoch the EEG segments. More precisely, given an EEG data segment $EEG_{p,c}$ (i.e., the segment generated during the EEG pre-processing step above, which represents the EEG response of participant *p* while reading the entire letter-matrix of condition *c*), and given the corresponding set of fixations $FIX_{p,c}$, we epoch the $EEG_{p,c}$ between −200 ms

to 500 ms of the onset time of each $f \in FIX_{p,c}$ and subtract the baseline amplitude of each epoch. This procedure results in a new set defined as

$$FRP_{p,c} = \{sFRP_i\}_{i=1}^{|FIX_{p,c}|}$$

where $sFRP_i$ corresponds to the EEG epoch on the onset time of the i-th fixation of $FIX_{p,c}$. We note that the two sets have the same cardinality. The generation of all sFRPs was implemented using custom python code.

2.4. Reading-related neural-congruency components

Our objective was to isolate neural components in the extracted fixation-related potentials that were likely modulated by RAN tasks and were informative of differences between CAC and DYS. Our approach was motivated by the hypothesis that the neural activity of participants that have developed adequate reading skills would exhibit neural activation patterns congruent with other participants with adequately developed reading skills. While contrarily, participants who experienced reading difficulties would have neural responses that deviated from such stereotypical patterns. Toward this objective, we formulated an optimization procedure to isolate neural components congruent among participants with sufficiently developed reading skills and explore those components as potential differentiation neuro-markers between CAC and DYS. Here, we provide details of our approach to isolate such reading-related neural-congruency components.

We seek to identify components (i.e., spatial projections of the fixation-related potential) that capture neural activity that maximally correlates among a group of children with adequately developed reading skills (i.e., CAC group). For this, we formulate an optimization problem as follows: for a group of S participants, $\mathcal{S} = \{s_1, s_2, \ldots, s_S\}$, *where* $s_i \in \mathbb{Z}^+$ denotes a participants index, representing a CAC group, we define the *between-subject* and *within-subject* cross-covariance matrices was :

$$R_b = \frac{1}{S(S-1)} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \left(1 - \delta_{ij}\right) \mathbf{R}_{ij}$$

$$\mathbf{R}_w = \frac{1}{S} \sum_{i \in \mathcal{S}} R_{ii}$$

where

$$\mathbf{R}_{ij} = \frac{1}{K} \mathbf{X}_i \mathbf{X}_\mathbf{j}^\mathbf{T}$$

where K is a normalizing scalar, δ_{ij} is the Kronecker delta [1], and $X_s \in \mathbb{R}^{D \times S.F}$ is the horizontally concatenated matrix comprised of the fixation-related potential of a participant s during a given condition (i.e., reading of a letter-matrix stimulus), defined as:

$$\mathbf{X}_s = [\, sFRP_1, \, sFRP_2, sFRP_3, \quad \ldots, sFRP_F]$$

For a spatial projection vector $\mathbf{w} \in \mathbb{R}^D$, the average Pearson Product Moment Correlation Coefficient between the fixation-related potentials, projected onto vector $\mathbf{w}$, across every pair of participants in the group is then defined as:

$$\rho = \frac{\mathbf{w}^T \mathbf{R}_b \mathbf{w}}{(\mathbf{w}^T \mathbf{R}_w \mathbf{w})}$$

We consider ρ as a measure of the degree of congruency in reading-related neural activity (projected onto component $\mathbf{w}$) among participants with adequately developed reading skills. As such, we seek to find the component $\mathbf{w}$ that maximized ρ. Taking the derivative of ρ with respect to $\mathbf{w}$ and setting it to zero, we get the solution of the optimization given as the eigenvectors to the generalized eigenvalue problem:

$$\left(R_w^{-1} R_b\right) \mathbf{w}_k = \lambda_k \mathbf{w}_k \tag{1}$$

where $\mathbf{w_k}$ is the k-th eigenvector of the matrix $\left(R_w^{-1} R_b\right)$ and corresponds to the component (i.e., spatial projection vector) that captures the k-th largest correlation in neural activity, and λ_k is the corresponding eigenvalue and denotes the strength of the correlation. We note that since $\left(R_w^{-1} R_b\right)$ is a $D \times D$ matrix (D being the number of channels), there are D solutions to the optimization problem (i.e. the D eigenvectors), each identifying a component at different correlation strength, with the first eigenvector (i.e., $k = 1$) having the strongest correlation, and subsequent components appearing in descending order of correlation strength. As such, the vector $\mathbf{w}_1$ defines a component (spatial projection) where neural activity is most strongly correlated among participants in the adequately developed reading skills, $\mathbf{w}_2$ defines the component where neural-activity exhibits the second strongest correlation among the groups, and so on.

To determine the reading-related neural-congruency (RRNC) score of an individual participant s with respect to the kth component, we measure the correlation between the fixation-related potentials of the subject to the fixation-related potentials of each subject in the group, after projecting both onto the component $\mathbf{w}_k$. Formally, we define the reading-related neural congruency score for a participant s and a component $\mathbf{w}_k$ as:

$$RRNC_{s,k} = \frac{\mathbf{w}_k^T \mathbf{R}_s^b \mathbf{w}_k}{\mathbf{w}_k^T \mathbf{R}_s^b \mathbf{w}_k}$$

where

$$\mathbf{R}_s^b = \frac{1}{S} \sum_{i \in \mathcal{S}} R_{si} + R_{is}, \; \mathbf{R}_s^w = \frac{1}{S} \sum_{i \in \mathcal{S}} R_{ss} + R_{ii},$$

We calculate RRNC scores separately for each condition (i.e., Rime-confound, Rime-non-confound, Visual-confound, Visual-non-confound). Moreover, to avoid training bias during the component extractions, data from the subject to be tested were excluded from the component extractions step. Finally, we define the neural-congruency feature vector of each participant s as:

$$\mathbf{c}_s = \left[RRNC_{s,1}, RRNC_{s,2}, \ldots RRNC_{s,K}\right]^\mathbf{T} \tag{2}$$

Each $RRNC_{s,k}$ score measures the strength of the congruency in neural activity between participant s and the CAC group, with respect to the k-th component. Therefore, the neural-congruency feature vector $\mathbf{c}_s$ encapsulates the neural-congruency of participant s, across all K components. In essence, the feature vector $\mathbf{c}_s$ characterizes the overall congruency observed in the participant's neural activity for each extracted component. We note that the

1 Kronecker delta is defined as: $\delta_{ij} = 1$ if $i = j$; delta $\delta_{ij} = 0$ if delta $i \neq j$

10.3389/fpsyg.2023.1076501

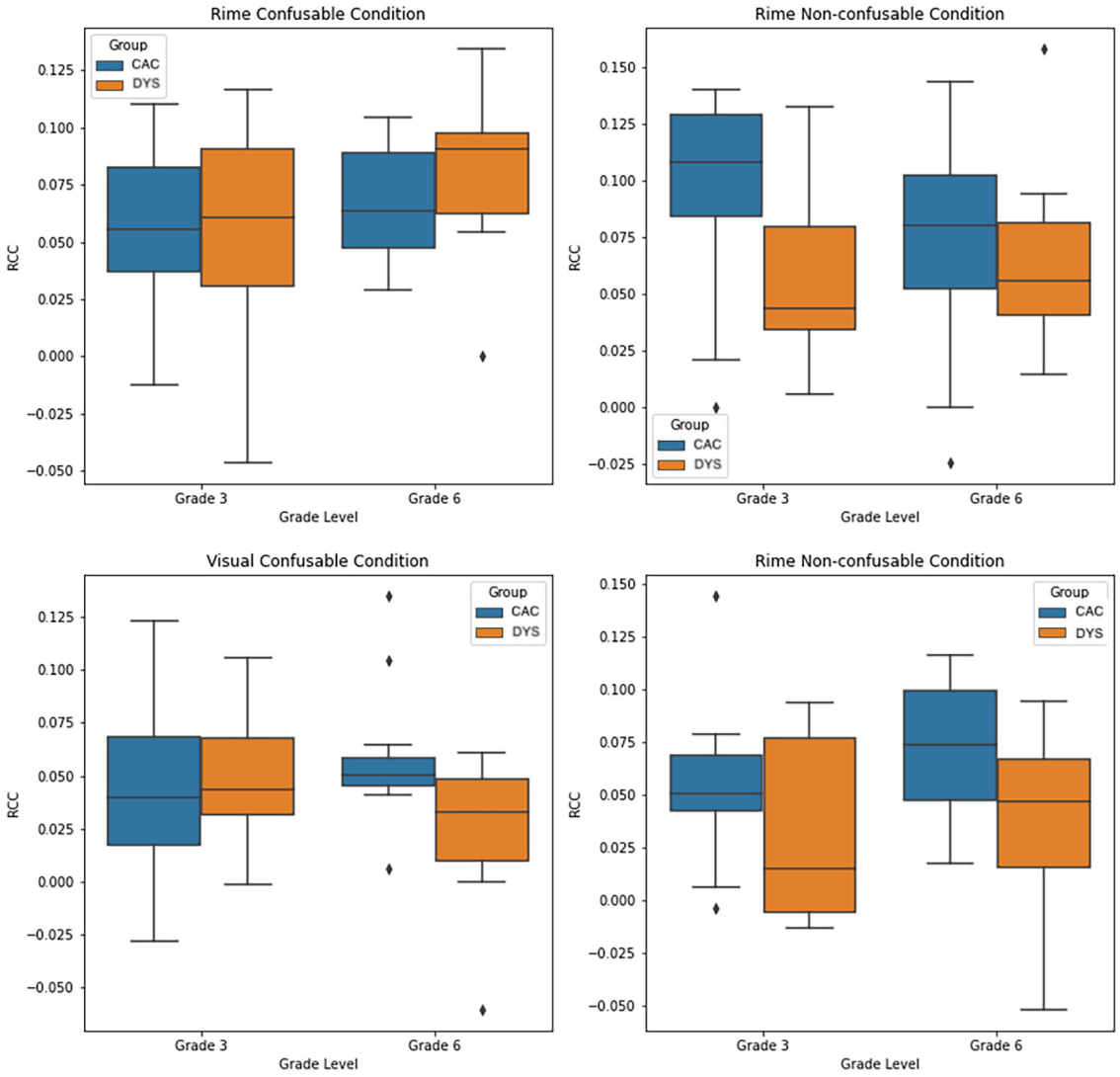


FIGURE 2
Box-plot showing the distribution of the Cumulative RAN-related Neural Congruency scores for each experiment condition, and each group.

dimensions of the vector $\mathbf{c}_s$ are indexed in descending order, according to the lambda score of each extracted component.

2.5. Aggregation of RAN-related neural-congruency components

Our goal was to explore whether information captured in the feature vector of neural-congruency components $\mathbf{c_s}$ is predictive of differentiating between DYS and CAC groups during the RAN task. Toward this objective, we considered two approaches for aggregating the RRNC scores into determining markers of dyslexia. The two approaches are detailed below.

2.5.1. Cumulative RAN-related neural-congruency metric

The first approach defines a neural metric by simply summing the RRNC score corresponding to the first $\widehat{K}$ components in the neural-congruency feature vector (i.e., those with the highest variance). Formally, given a feature vector $\mathbf{c}_s$ of a participant s (as defined in eq. 2), we define the Cumulative RAN-related Neural-congruency metric (C-RRNN) as follows:

$$C - RRNN_s = \sum_{k=1}^{\hat{K}} \mathbf{c}_s(k)$$

where the index k denotes the k-th element of the vector. The value of $\widehat{K} = 3$ was selected by identifying the 'knee' in the plot of eigenvalues of equation 1, and was fixed across the calculation of C-RRNN of all participants.

2.5.2. LASSO-weighted RAN-related neural-congruency metric

The second approach learns a classifier that optimally weights the contribution of each of the identified neural congruency components to best differentiate between typically developing children and children with dyslexia. In particular, we employed a sparse Logistic Regression classifier with LASSO

regularization, using the $K = 10$ components with the highest lambda values as independent variables, and an individual's group (DYS or CAC) as the dependent variable. We opted to use a sparse classifier because it minimizes the number of non-zero parameters, thus, favoring selecting a small subset of meaningful neural-congruency components. The classifier's prediction output corresponds to an optimally weighted-sum of the individual neural-congruency scores that maximizes the differentiation between DYS and CAC. Moreover, since the optimal weights are calculated using the LASSO regularizations, we refer to the resulting prediction score as LASSO-weighed RAN-related Neural-congruency metric. In our analysis, a separate classifier was trained on each of the four conditions (i.e., Rime-confusable; Rime Non-confusable; Visual-confusable; Visual Non-confusable) using a leave-one-participant-out cross-validation procedure to avoid training biases.

2.6. Spatiotemporal profiles of RAN-related neural-congruency

Given the solutions to the generalized eigenvalue problem, the temporal profile of each component was calculated as the product of each component $\widehat{\mathbf{w}}_k$, with each single-trial response and taking the grant-average response of the projected components. Moreover, the topographical profile (i.e., the forward model) of each component was calculated as follows:

$$\mathbf{a_k} = \frac{\mathbf{R_w}\widehat{\mathbf{w}}_\mathbf{k}}{\widehat{\mathbf{w}}_\mathbf{k}^\mathrm{T}\mathbf{R_w}\widehat{\mathbf{w}}_\mathbf{k}}$$

The forward model captures the covariance between each component's activity as measured by each electrode.

2.7. Statistical analysis

To avoid training bias, all model parameters, including the extracted neural-congruency components and classifier weights, are trained using a leave-one-participant-out cross-validation procedure. The classifier's generalization performance is calculated as the area under the Receiver Operator Characteristic curve (AUC) on cross-validated scores. A permutation test is used to determine statistical significance levels over AUC scores (10,000 repetitions). A two-way ANOVA was used for group and grade comparisons, with the neural congruency metrics as the dependent variable.

3. Results

3.1. Group comparisons using the RAN-related neural-congruency metrics

For each participant, Cumulative Neural-Congruency scores were calculated as the sum of the three components with the highest lambda values, corresponding to the components whose projection has the highest correlation. To avoid training biases during the neural-congruency component identification, data from the participant for whom the Neural-Congruency score was to be calculated was excluded from the component identification step. Neural-Congruency scores were obtained and analyzed separately for each of the four conditions (i.e., Rime-confusable; Rime Non-confusable; Visual-confusable; Visual Non-confusable). For each condition, a separate two-way ANOVA was performed to analyze the effect of group (i.e., CAC vs. DYS) and grade (grade 3 vs grade 6) on the Neural-Congruency scores. A two-way ANOVA on Rime-confusable Cumulative RAN-related Neural-Congruency scores shows a significant main effect of grade ($p < 0.04$), with grade 6 group showing higher neural-congruency than the grade 3 group. The analysis also releveled there was not a significant interaction effect between the group and grade, $F(1,52) = 0.740$, $p = 0.39$, nor a statistically significant effect for group ($p = 0.19$). A two-way ANOVA on Rime non-confusable Neural-Congruency scores revealed a significant main effect of group ($p < 0.01$). Participants in the control group showed a higher Cumulative RAN-Related Neural-Congruency scores than the participants with dyslexia group. There was no main effect of grade ($p = 0.61$), and there was no statistically significant effect observed between group and grade $F(1,52) = 1.67$, $p = 0.20$. On the Visual Confusable task, a two-way ANOVA revealed a statistically significant interaction effect between the group and grade $F(1,52) = 4.22$, $p = 0.04$. The analysis also revealed there was not a statistically significant main effect of either the group or the grade ($p > 0.05$). Finally, the Visual non-confusable task revealed a main effect of the group ($p < 0.01$), with participants in the control group showing a higher Neural-congruency scores than the children with dyslexia group (DYS). No grade or interaction effect between grade and groups were observed during the Visual non-confusable. **Figure 2**, shows the box plots for the four two-way ANOVA for each condition. Moreover, a two-factor repeated measures ANOVA was performed to compare the effect of modality condition (rime vs visual) and confusability (confusable vs non-confusable). The analysis showed a statistically significant difference in RAN-related Neural-Congruency scores between modality conditions ($p < 0.001$) with the Rime modality exhibiting higher scores.

3.2. Lasso-weighted RAN-related neural-congruency as a significant predictor of group differences across conditions

We aimed to further explore the characteristics of the underlying neural activity captured by the neural-congruency components and determine the degree to which each neural-congruency component contributes toward inferring an individual's group (i.e., DYS or CAC). We hypothesized that a weighted aggregation of individual neural-congruency components would carry predictive information about the participant's condition. We employed a sparse logistic regression classifier using the ten components with the highest lambda values as independent variables, and an individual's group (DYS or CAC) as the dependent variable. The classifier was modeled and trained according to the procedure described in Section "2.5.2. LASSO-weighted RAN-related neural-congruency metric." The statistical

significance levels over AUC scores were established using a permutation test (10,000 repetitions). In all four conditions, the cross-validated AUC scores of the classifiers show that the LASSO-weighted RAN-related Neural-congruency metric predicted an individual's group. The prediction accuracy for all conditions, the Rime-confusable condition (AUC = 0.86, $p < 0.00001$), Rime Non-confusable condition (AUC = 0.86, $p < 0.00001$) the Visual Non-confusable (AUC = 0.81, $p < 0.00001$), and the Visual Confusable condition (AUC = 0.73, $p < 0.005$) was high and statistically significant. The ROC curves and corresponding AUC scores for each condition and the 95th-percentile envelop of the ROC curve under the null-distribution are depicted in **Figure 3**. The boxplot in **Figure 4** shows the distribution of the lasso-weighted RAN-related neural-congruency for each grade and group and experimental condition.

3.3. Spatio-temporal profile of neural-congruency components

The spatio-temporal profile of each RAN-related Neural Congruency for all conditions is shown in **Figure 5** and the **Supplementary material**. Each spatio-temporal profile comprises the "Forward model," which shows the spatial distribution of the correlated neural activity captured by the corresponding component, and the temporal profile–the time course of the FRP's neural activity when projected onto that neural-congruency component. Visual inspection of the temporal profile provides insights into timeframe differences between groups and condition intensify. Similarly, visual inspection of the forward model alludes to potential brain areas from which the underlying neural activity originated from. The weights associated with each channel in the forward model capture the electrical coupling of the correlated components. The components are ordered based on their corresponding lambda scores, with component #1 reflecting the highest lambda, and component #10 the smallest lambda value. The LASSO-coefficients associated with each component are shown over each component's forward model and indicate the weight used to aggregate the neural-congruency components.

3.4. Naming speed behavioral data analysis

Analysis of the behavioral data obtained during the experiment has been previously reported by Christoforou et al. (2021a) and it is outside the scope of this papers. However, to provide the context on our results on neurophysiological data, we include a summary of the behavioral data analysis of this experiment. A MANOVA analysis was performed on the behavioral data, with the naming speed performance time for each of the four RAN tasks as dependent measures and Group (2) as a fixed factor. The main group effects was significant, Wilks' Lambda = 0.754, $F(4,55) = 4.48$, $p < 0.01$, $\eta^2 = 0.20$. Subsequent univariate analyses demonstrated that the group's main effect was significant for all individual measures after Bonferroni adjustments (**Supplementary Table 1**). The DYS group performed significantly poorer than the chronological age controls in all naming speed measures.

4. Discussion

Methodological difficulties in using traditional neurophysiological techniques to investigate the neural underpinning of dyslexia during serial RAN have hindered the development of studies in that direction (Bakos et al., 2020; Christoforou et al., 2021a). To help alleviate this problem, we proposed a novel computational approach for identifying neural components elicited during the serial RAN task. We also explore the component's contribution to characterizing the underlying neural differences between children with dyslexia and chronologically age controls in four experimental conditions. Specifically, we formulated an optimization problem to extract spatiotemporal components from EEG measures that maximize the correlation between single-trial fixation-related potentials during serial RAN. We treated RAN as an additional variable to the diagnosis of reading difficulties, explaining the shared variance of the disorder. Based on the resulting components, we defined the per-subject neural-congruency scores that indicate the degree to which each participant engaged in neural processes relevant to the RAN task. Results show that the neural-congruency components capture the neural activity of cognitive processes associated with reading and are informative of group differences between children with dyslexia and typically developing children. Moreover, our results provide insights into the spatial and temporal characteristics of the underlying mental process involved in the naming speed and points to which potential neuro-cognitive mechanisms differentiate between children with and without dyslexia. These findings are robust given the careful matching of the participating groups based on their verbal and non-verbal ability and demographic variables. Furthermore, the study findings contribute to the relevant research because previous evidence has overlooked the contribution of neurophysiological measurements during serial RAN tasks and their relation to behavioral measures (i.e., naming speed) that together might explain reading performance and related difficulties.

4.1. Differences between DYS and CAC in the cumulative RAN-related neural-congruency components

Regarding the contribution of the Cumulative RAN-related neural-congruency components in differentiating between children with and without dyslexia, the results revealed significant differences between the groups in the both non-confusable conditions (i.e., Rime non-confusable, and Visual non-confusable). Specifically, on the one hand, cumulative RAN-related neural-congruency scores in typically developing children were significantly higher than their counterparts in the DYS groups. This finding denotes an increase in the synchronicity in neural activity in the CAC groups, which suggests that the CAC group has developed a more consistent stereotypical response in neural activations when engaging processes associated with the execution of the serial RAN task. Breznitz and colleagues (Breznitz, 2001, 2003; Breznitz and Misra, 2003) have proposed the 'synchronization hypothesis' to describe this phenomenon. According to this hypothesis, accurate information integration in

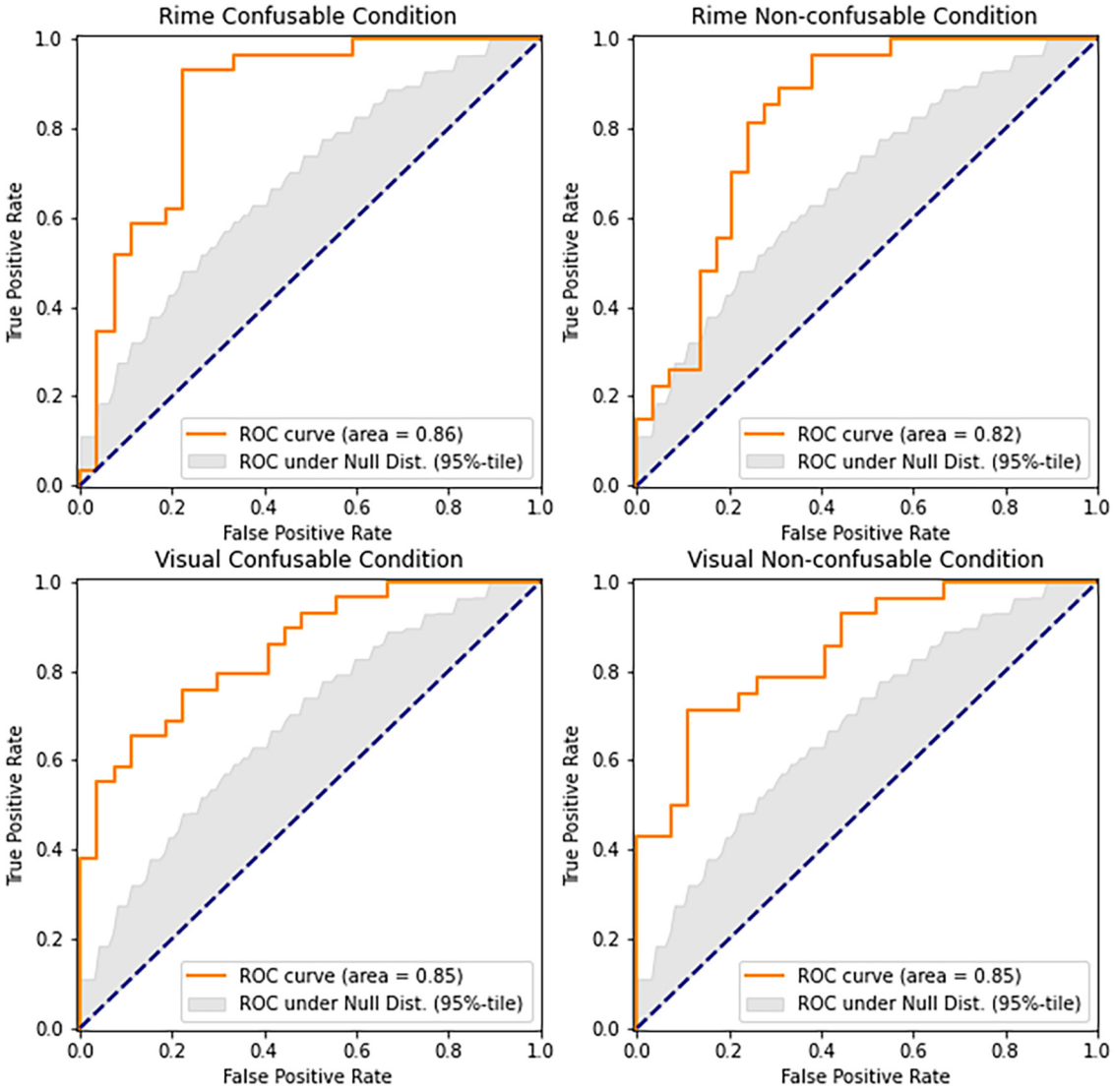


FIGURE 3
Shows the ROC curves of the predictions based on the LASSO-weighted RAN-related Neural-congruency components. The gray area denotes the ROC score under the null hypothesis (i.e., neural-congruency scores between DYS and CAC groups are indistinguishable). All four graphs show that the LASSO-weighted RAN-related Neural-congruency scores carry significant predictive information as the condition (i.e., DYS or CAC) of the participants.

decoding words can occur only when the modalities and brain systems are synchronized. This synchronization, therefore, requires that the processing speed and the accuracy with which content information is processed and transferred within and between the various activated neural systems are readily available. Our findings further confirm this hypothesis that this synchronicity in neural activity is evident in typically developing readers but to a lesser degree in children with dyslexia.

Indeed, on the other hand, the consistency in the synchronicity of the responses diminishes across the DYS groups, suggesting a lack of regularity in the processing the letter stimuli in the serial RAN. Visual inspection of the time course of the three neural-congruency components used in calculating the Cumulative RAN-related neural-congruency score suggests that the difference in congruency appears between 100 ms–200 ms following the fixation onset. The analysis also shows that the effect that emerged in the non-confusable tasks is not present in the confusable tasks (i.e., Rime-confusable, and Visual-confusable). That is, there were no significant differences observed in the Cumulative scores between DYS and CAC. We interpret these effects in the context of neural efficiency theory. Specifically, we argue that the CAC group has developed efficient mechanisms for recognizing, decoding, and reading letters as captured by the consistency in the neural responses across participants. In contrast, the DYS group does not show the same regularity in neural responses, suggesting the corresponding mechanisms for recognizing, decoding, and reading letters are less fine-tuned in the DYS group. This latter finding has additive value to Breznitz's Asynchrony Theory (Breznitz, 2008) which proposes that dyslexia is an outcome of the failure to synchronize the various brain entities activated during reading-related processes. Nevertheless, this asynchrony is also evident for children with dyslexia and their

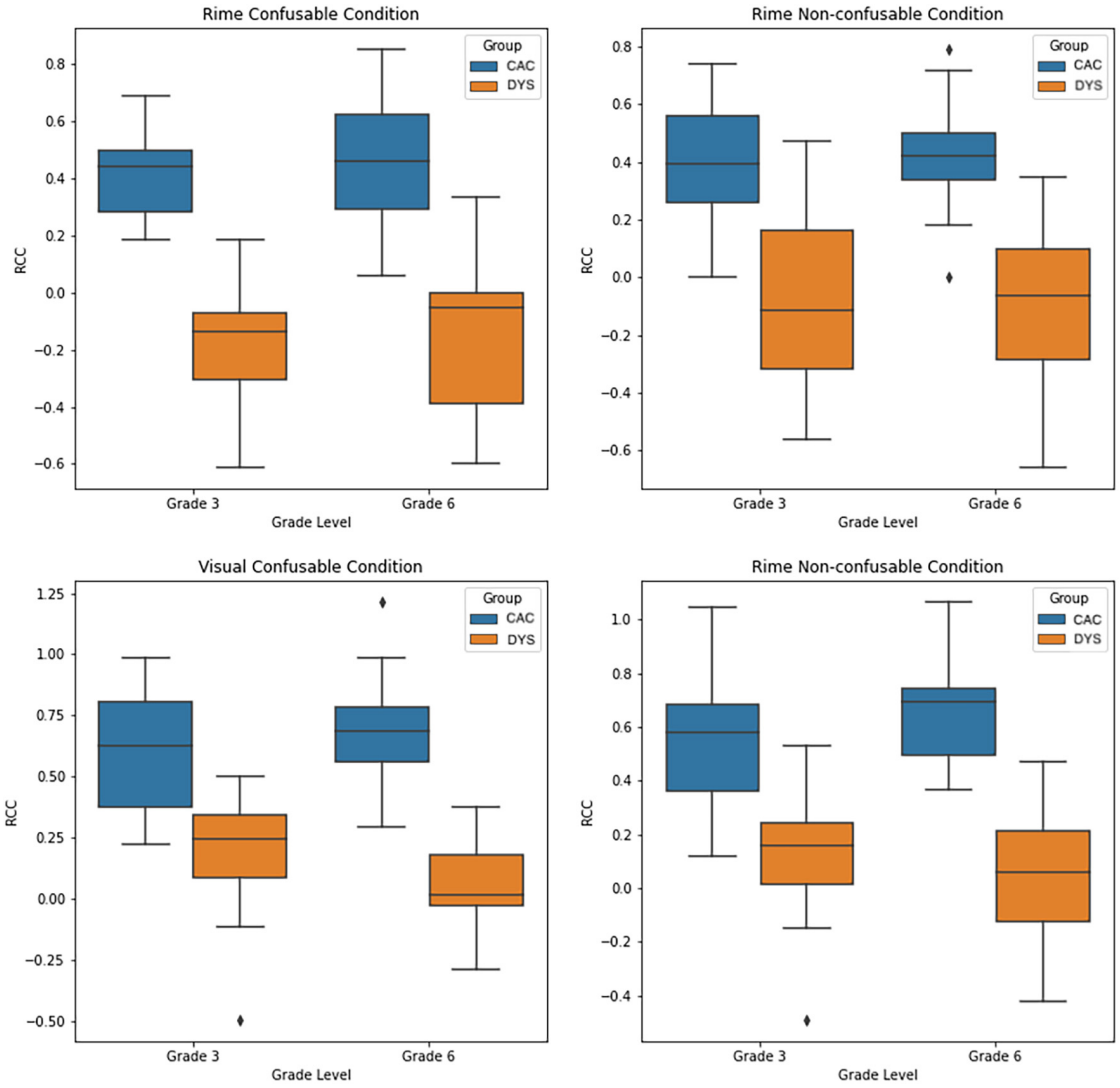


FIGURE 4
Box-plot showing the distribution of the LASSO-weighted RAN-related Neural Congruency scores for each experiment condition, and each group.

typically developing counterparts when the stimulus' complexity increases – as occur in the rime or visual confusability. This finding underscores the need for additional neural resources to resolve the stimulus's confounding elements (rime or visual). To that end, the Cumulative RAN-related neural congruency components, fail to capture a consistent neural-response across either of the groups.

4.2. Differences between DYS and CAC is the lasso-aggregate neural-congruency components

Although the simple aggregation of the top three neural-congruency scores revealed significant differences between groups under the non-confounding conditions, we hypothesized that an optimally weighted sum over all ten neural-congruency components would capture additional differences in neural activations between groups. Indeed, the sparse logistic regression classification revealed that an optimally weighted aggregation over the neural-congruency components differentiates between CAC and DYS across both confounding and non-confounding conditions. These results suggest that components beyond the three with the highest lambda values do capture neural activity relevant to the task. Moreover, the weights associated with each component differ in both amplitude and sign (i.e., they can contribute either positively or negatively to the sum). This finding suggests that the neural activations captured by each individual component might appear with different intensity and polarity in each group. For example, one component that might capture activity associated with character-disambiguation might exhibit stronger synchronicity in one group (i.e., contributing positively to the sum); however, another component might exhibit stronger synchronicity in the other group. Thus, simple aggregation of components might result in cancelling out this effect. Overall, our classification model suggests that individual components must be considered when analyzing neural activations.

10.3389/fpsyg.2023.1076501

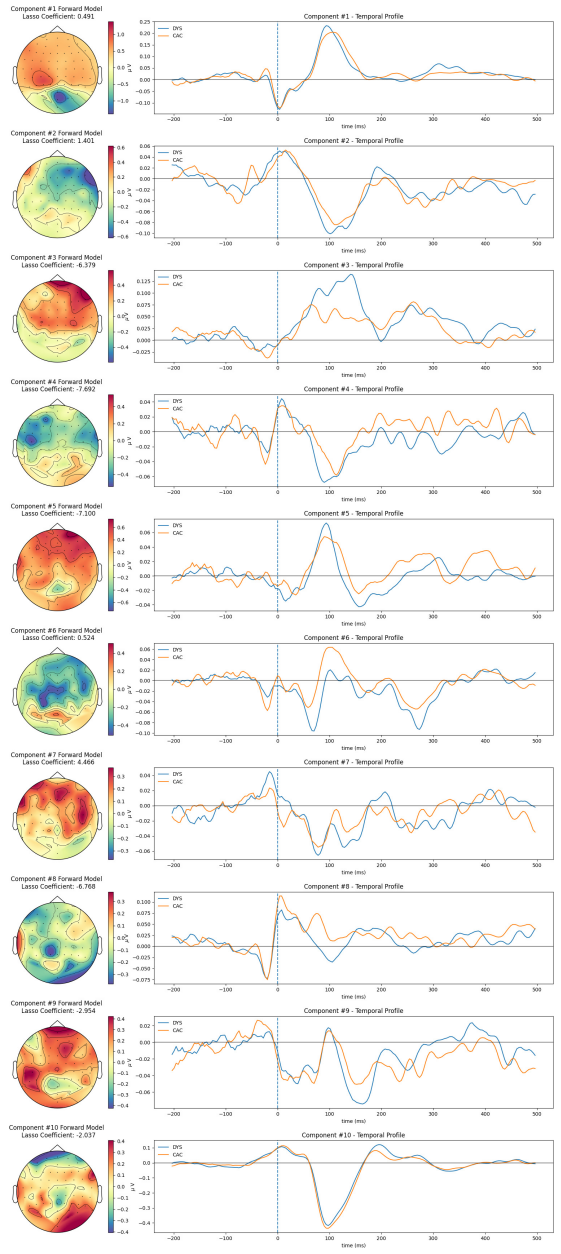


FIGURE 5
Spatiotemporal profiles of RAN-related Neural-congruency component for the rime-confusable condition; The corresponding plots for all conditions are provided in the **Supplementary material**.

4.3. Interpretation of the spatiotemporal profiles

Spatial (i.e., the forward model) and temporal profiles of the extracted components are depicted in **Figure 4** for each condition. The forward model of the neural-congruency component #1 (i.e., the one with the highest lambda score) exhibits a similar topography across all four conditions; moreover, their corresponding temporal profiles show the neural activity is most strongly modulated at a time window of around 100 ms. The similarity suggests that component #1 captures neural activity common to all four conditions, albeit at different intensity levels. At the very least, this finding confirms previous evidence showing that processing complex features of textual stimuli is reflected in the electrophysiological responses around 100 ms after stimulus presentation (Hauk et al., 2006).

The temporal profile of components #10 shows similarity in waveform across the four conditions and a peak amplitude at around 100 ms following the fixations onset. Similarities in the spatial profiles are observed among some of the remaining components as well, although the indexing/ordering of those components varies among conditions. Such similarities suggest that the matching neural-congruency components likely capture neural activity originating from the same underlying source. The variation in the indexing is expected since the ranking of the components is established independently for each condition and depends on the relative strength of all the neural-congruency components in that condition. Interestingly, several projected temporal profiles in each condition display a stereotypical response consistent with neural activations often observed in traditional grant average Event-related Potential analysis (i.e., N/P100 and N170). For example, the temporal profile of component #9 shows an N170 waveform response with visible differences in amplitude at around 170 ms between CAC and DYS. Moreover, the forward model topography of this component shows it to emerge more strongly in electrodes over the left posterior-occipital regions. In the literature, the difference in N170 over the left-posterior occipital region is regarded as an electrophysiological marker of visual expertise (Varga et al., 2020). Particularly, children with a lower letter knowledge, as pre-readers, have shown reduced N170 amplitudes and delayed N170 latency compared to typical readers during letter-string presentations (Maurer et al., 2005). Therefore, component #9, extracted by our method, can be interpreted as potentially indicating a visual precursor to literacy resulting from familiarity with letter strings. Furthermore, negativity components at 170 ms have been associated with attention modulation (i.e., Kropotov, 2016), and in turn, attention represents a known latent common cognitive factor of RAN and reading (Papadopoulos et al., 2016). Therefore, part of the neural activity captured by the component could also reflect those distal processes. Moreover, several forward model scalp plots display topographies that often arise as the "forward problem" solutions to single-source dipole models, indicating that those components likely capture source-localized neural activity of different neural processes. Finally, components with high lasso coefficients (i.e., either positive or negative) capture underlying neural activity that more strongly differentiate between DYS and CAC. Thus, group differences among groups appear in the underlying sources modeled by those components.

Further studies could investigate these components in more detail and draw additional conclusions about the neural and cognitive processes that contribute to these differentiations. For example, an interesting expansion of these findings relates to examining the validity of the suggested neural-congruency components analysis against other more conventional EEG analyses. Given that the literature has only recently started to investigate the important information that the FRPs can provide to studying the RAN-reading relationship or other similar correlates of reading performance (e.g., Christoforou et al., 2021b; Fella et al., 2022), the present findings are considered a promising beginning of this quest. Another interesting expansion would be to examine

the relationship of the extracted neural congruency components to eye-tracking-based metrics during RAN, such as the recently proposed entropy-based gaze time-series analysis on RAN (Wang et al., 2022).

In conclusion, the RAN-related neural congruency component, identified by our proposed method, carry information on the neural basis of naming speed that differentiates between children with dyslexia and their typically developing counterparts. The topographies of the resulting components suggest that each component likely captures source-localized neural activity corresponding to distinct neural processes. Moreover, neural differences appear to be distributed across several RAN-related Neural-congruency components but at different intensity levels. Therefore, optimally combining the RAN-related components using machine learning enhanced the power of analysis in identifying differences in both the confusable and non-confusable conditions, which have been missed by simple aggregation of the RAN-related Components. Our findings also support the Neural-congruency hypothesis (Christoforou et al., 2021b; Christoforou and Theodorou, 2021), indicating that neural activity elicited during cognitive tasks is more congruent among participants that have mastered the cognitive skills but less congruent otherwise. Finally, our proposed approach opens up new research directions in studying the neural underpinnings of naming speed and their association with reading performance and reading difficulties. For instance, until recently, evidence concluded that individuals with reading difficulties show increased reliance on inferior frontal regions of the reading network and right-hemisphere posterior regions (e.g., Richlan et al., 2011; Norton et al., 2015). The present findings show that several brain areas contribute to the execution of naming tasks which show similarities with reading tasks. Although simple at the surface level, RAN tasks are multi-componential, as is reading (Papadopoulos et al., 2016). Thus, we believe that the present method succeeded in better defining the properties of other processes (including those of reading) that RAN carries and how these are critical in determining naming speed's influential role on reading performance. At the very least, capturing the RAN performance in the form of neural components helps us better understand the process involved in performing RAN tasks and explore some reasons for poor performance. Next, provided that the generated components will be further systematically tested against behavioral performance measures, the likelihood of deciphering issues relevant to the significant similarity of the RAN and reading measures, such as seriality, is possible (see Altani et al., 2020). Thus, in future work, we plan to further explore the spatiotemporal characteristics and brain sources of the RAN-related Neural-congruency components and the relationship between the neural underpinnings of naming speed - as captured by the Neural-congruency components- and reading difficulties, as well ex explore eye-tr.

Data availability statement

The datasets presented in this article are not readily available because further analysis of the data is currently in progress. Requests to access the datasets should be directed to CC, christoc@stjohns.edu.

Ethics statement

The studies involving human participants were reviewed and approved by the Cyprus National Bioethics Committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

CC conceived the original methodology, designed and developed the computational framework, carried out the implementation of the methods, data pre-processing, and analysis, and took the lead in writing the manuscript. MT contributed to the method's implementation and designed the figures. CC, AF, and TP contributed to the RAN experiment and study design and data collection. TP contributed to drafting and editing the manuscript and to the interpretation of the results. All authors discussed the results and commented on the manuscript.

Funding

Dissemination of this work was supported by the Research Promotion and Innovation Foundation Programmes for Research, Technology Development and Innovation "RESTART 2016-2020" (Complementary/0916/0187, PI: TP).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1076501/full#supplementary-material

References

Al Dahhan, N. Z., Kirby, J. R., Brien, D. C., Gupta, R., Harrison, A., and Munoz, D. P. (2020). Understanding the biological basis of dyslexia at a neural systems level. *Brain Commun.* 2:fcaa173. doi: 10.1093/braincomms/fcaa173

Altani, A., Protopapas, A., and Georgiou, G. K. (2017). The contribution of executive functions to naming digits, objects, and words. *Read. Writ.* 30, 121–141. doi: 10.1007/s11145-016-9666-4

Altani, A., Protopapas, A., Katopodi, K., and Georgiou, G. K. (2020). Tracking the serial advantage in the naming rate of multiple over isolated stimulus displays. *Read. Writ.* 33, 349–375. doi: 10.1007/s11145-019-09962-7

Araújo, S., Huettig, F., and Meyer, A. S. (2021). What underlies the deficit in Rapid Automatized Naming (RAN) in Adults with Dyslexia? Evidence from eye movements. *Sci. Stud. Read.* 25, 534–549. doi: 10.1080/10888438.2020.1867863

Araújo, S., Inácio, F., Francisco, A., Faísca, L., Petersson, K. M., and Reis, A. (2011). Component processes subserving rapid automatized naming in dyslexic and non-dyslexic readers. *Dyslexia* 17, 242–255. doi: 10.1002/dys.433

Araújo, S., Reis, A., Petersson, K. M., and Faísca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *J. Educ. Psychol.* 107, 868–883. doi: 10.1037/edu0000006

Bakos, S., Mehlhase, H., Landerl, K., Bartling, J., Schulte-Körne, G., and Moll, K. (2020). Naming processes in reading and spelling disorders: An electrophysiological investigation. *Clin. Neurophysiol.* 131, 351–360. doi: 10.1016/j.clinph.2019.11.017

Bowey, J. A., McGuigan, M., and Ruschena, A. (2005). On the association between serial naming speed for letters and digits and word-reading skill: Towards a developmental account. *J. Res. Read.* 28, 400–422. doi: 10.1111/j.1467-9817.2005.00278.x

Breznitz, Z. (2001). "The determinants of reading fluency: A comparison of dyslexic and average readers," in *Dyslexia, Fluency and the Brain*, ed. M. Wolf (Timonium, MD: York Press), 245–276.

Breznitz, Z. (2003). Speed of phonological and orthographic processing as factors in dyslexia: Electrophysiological evidence. *Genetic Soc. Gen. Psychol. Monogr.* 129, 183–206.

Breznitz, Z. (2008). "The origin of dyslexia: The asynchrony phenomenon," in *The SAGE Handbook of Dyslexia*, eds G. Reid, A. Fawcett, F. Manis, and L. Siegel (London: SAGE Publications Ltd), 11–30. doi: 10.4135/9780857020987.n2

Breznitz, Z., and Misra, M. (2003). Speed of processing of the visual-orthographic and auditory-phonological systems in adult dyslexies: The contribution of "asynchrony" to word recognition deficits. *Brain Lang.* 85, 486–502. doi: 10.1016/S0093-934X(03)00071-3

Christoforou, C., Constantinidou, F., Shoshilou, P., and Simos, P. (2013). Single-trial linear correlation analysis: Application to characterization of stimulus modality effect. *Front. Comput. Neurosci.* 7:15. doi: 10.3389/fncom.2013.00015

Christoforou, C., Fella, A., Leppänen, P. H. T., Georgiou, G. K., and Papadopoulos, T. C. (2021a). Fixation-related potentials in naming speed: A combined EEG and eye-tracking study on children with dyslexia. *Clin. Neurophysiol.* 132, 2798–2807. doi: 10.1016/j.clinph.2021.08.013

Christoforou, C., Haralick, R. M., Sajda, P., and Parra, L. C. (2010). "The bilinear brain: Towards Subject-invariant Analysis," in *Proceedings of the 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, (Guilin), 1–6. doi: 10.1109/ISCCSP.2010.5463377

Christoforou, C., Hatzipanayioti, A., and Avraamides, M. (2018). Perspective-taking vs mental rotation: CSP-based single-trial analysis for cognitive process disambiguation. *Brain Informat.* 11309, 109–119. doi: 10.1007/978-3-030-05587-5_11

Christoforou, C., Papadopoulos, T. C., Constantinidou, F., and Theodorou, M. (2017). Your brain on the movies: A computational approach for predicting box-office performance from viewers brain responses to movie trailers. *Front. Neuriinfomat.* 11:72. doi: 10.3389/fninf.2017.00072

Christoforou, C., Papadopoulos, T. C., and Theodorou, M. (2021b). "Single-trial FRPs: A machine learning approach towards the study of the neural underpinnings of reading disorders," in *Proceedings of the International FLAIRS Conference Proceedings*, (Hutchinson Island). doi: 10.32473/flairs.v34i1.128446

Christoforou, C., Papadopoulos, T. C., and Theodorou, M. (2022a). "Machine Learning approach for studying the neural underpinnings of dyslexia on a phonological awareness task," in *Proceedings of the International FLAIRS Conference*, (Hutchinson Island). doi: 10.32473/flairs.v35i.130576

Christoforou, C., Papadopoulos, T. C., and Theodorou, M. (2022b). Towards the study of the neural underpinnings of dyslexia during final-phoneme elision: A machine learning approach. *Brain Informat.* 13406, 74–85. doi: 10.1007/978-3-031-15037-1_7

Christoforou, C., and Theodorou, M. (2021). "Towards EEG-based emotion recognition during video viewing: Neural-congruency explains user's emotion experienced in music videos," in *Proceedings of the International FLAIRS Conference*, (Hutchinson Island). doi: 10.32473/flairs.v34i1.128458

Cummine, J., Chouinard, B., Szepesvari, E., and Georgiou, G. K. (2015). An examination of the rapid automatized naming-reading relationship using functional magnetic resonance imaging. *Neuroscience* 305, 49–66. doi: 10.1016/j.neuroscience.2015.07.071

Denckla, M. B. (1972). Colour-naming defects in dyslexic boys. *Cortex* 8, 164–176. doi: 10.1016/S0010-9452(72)80016-9

Dmochowski, J. P., Sajda, P., Dias, J., and Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention. *Front. Hum. Neurosci.* 6:112. doi: 10.3389/fnhum.2012.00112

Easson, K., Al Dahhan, N. Z., Brien, D. C., Kirby, J. R., and Munoz, D. P. (2020). Developmental trends of visual processing of letters and objects using naming speed tasks. *Front. Hum. Neurosci.* 14:562712. doi: 10.3389/fnhum.2020.562712

Fella, A., Christoforou, C., Loizou, M., and Papadopoulos, T. C. (2022). Investigating the relationship between phonological awareness and reading using Event-Related Potentials. *Psychology* 27, 79–97.

Georgas, D. D., Paraskevopoulos, I. N., Bezevegis, I. G., and Giannitsas, N. D. (1997). *Standardization in Greek of the Intelligence Scale for Children*, 3rd Edn. Athens: Ellinika Grammata.

Georgiou, G. K., Papadopoulos, T. C., Fella, A., and Parrila, R. K. (2012). Rapid naming speed components and reading development in a consistent orthography. *J. Exp. Child Psychol.* 112, 1–17. doi: 10.1016/j.jecp.2011.11.006

Georgiou, G. K., Papadopoulos, T. C., and Kaizer, E. (2014). Different RAN components relate to reading at different points in time. *Read. Writ.* 27, 1379–1394. doi: 10.1007/s11145-014-9496-1

Georgiou, G. K., Parrila, R. K., Cui, Y., and Papadopoulos, T. C. (2013). Why is rapid automatized naming related to reading? *J. Exp. Child Psychol.* 115, 218–225. doi: 10.1016/j.jecp.2012.10.015

Gordon, P. C., and Hoedemaker, R. S. (2016). Effective scheduling of looking and talking during rapid automatized naming. *J. Exp. Psychol.* 42, 742–760. doi: 10.1037/xhp0000171

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., and Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage* 30, 1383–1400. doi: 10.1016/j.neuroimage.2005.11.048

Henry, R., Van Dyke, J. A., and Kuperman, V. (2018). Oculomotor planning in RAN and reading: A strong test of the visual scanning hypothesis. *Read. Writ.* 31, 1619–1643. doi: 10.1007/s11145-018-9856-3

Jacobson, L. A., Ryan, M., Martin, R. B., Ewen, J., Mostofsky, S. H., Denckla, M. B., et al. (2011). Working memory influences processing speed and reading fluency in ADHD. *Child Neuropsychol.* 17, 209–224. doi: 10.1080/09297049.2010.532204

Jones, M. W., Branigan, H. P., and Kelly, M. L. (2009). Dyslexic and nondyslexic reading fluency: Rapid automatized naming and the importance of continuous lists. *Psychon. Bull. Rev.* 16, 567–572. doi: 10.3758/PBR.16.3.567

Jones, M. W., Obregón, M., Louise Kelly, M., and Branigan, H. P. (2008). Elucidating the component processes involved in dyslexic and non-dyslexic reading fluency: An eye-tracking study. *Cognition* 109, 389–407. doi: 10.1016/j.cognition.2008.10.005

Kendeou, P., Papadopoulos, T. C., and Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learn. Instruct.* 22, 354–367. doi: 10.1016/j.learninstruc.2012.02.001

Kirby, R., Georgiou, G. K., Martinussen, R., and Parrila, R. K. (2010). Naming speed and reading: A review of the empirical and theoretical literature. *Read. Res. Q.* 45, 341–362. doi: 10.1598/RRQ.45.3.4

Kropotov, J. D. (2016). "Chapter 3.1 - sensory systems and attention modulation," in *Functional Neuromarkers for Psychiatry*, ed. J. D. Kropotov (Cambridge, MA: Academic Press), 137–169. doi: 10.1016/B978-0-12-410513-3.00011-5

Kuperman, V., Van Dyke, J. A., and Henry, R. (2016). Eye-movement control in RAN and reading. *Sci. Stud. Read.* 20, 173–188. doi: 10.1080/10888438.2015.1128435

Landerl, K., and Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *J. Educ. Psychol.* 100, 150–161. doi: 10.1037/0022-0663.100.1.150

Leong, C. K., Tse, S. K., Loh, K. Y., and Hau, K. T. (2008). Text comprehension in Chinese children: Relative contribution of verbal working memory, pseudoword reading, rapid automatized naming, and onset-rime phonological segmentation. *J. Educ. Psychol.* 100, 135–149. doi: 10.1037/0022-0663.100.1.135

Liao, C. H., Deng, C., Hamilton, J., Lee, C. S., Wei, W., and Georgiou, G. K. (2015). The role of rapid naming in reading development and dyslexia in Chinese. *J. Exp. Child Psychol.* 130, 106–122. doi: 10.1016/j.jecp.2014.10.002

Logan, J. A. R., and Schatschneider, C. (2014). Component processes in reading: Shared and unique variance in serial and isolated naming speed. *Read. Writ.* 27, 905–922. doi: 10.1007/s11145-013-9475-y

Maurer, U., Brem, S., Bucher, K., and Brandeis, D. (2005). Emerging neurophysiological specialization for letter strings. *J. Cognit. Neurosci.* 17, 1532–1552. doi: 10.1162/089892905774597218

Moll, K., Fussenegger, B., Willburger, E., and Landerl, K. (2009). RAN is not a measure of orthographic processing. Evidence from the asymmetric German orthography. *Sci. Stud. Read.* 13, 1–25. doi: 10.1080/10888430802631684

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., et al. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learn. Instruct.* 29, 65–77. doi: 10.1016/j.learninstruc.2013.09.003

Niglieri, J. A., and Das, J. P. (1997). *Das-Naglieri cognitive system*. Nashville, TN: Riverside Publishing.

Norton, E. S., Beach, S. D., and Gabrieli, J. D. (2015). Neurobiology of dyslexia. *Curr. Opin. Neurobiol.* 30, 73–78. doi: 10.1016/j.conb.2014.09.007

Norton, E. S., and Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annu. Rev. Psychol.* 63, 427–452. doi: 10.1146/annurev-psych-120710-100431

Papadopoulos, T. C., Csépe, V., Aro, M., Caravolas, M., Diakidoy, I. A., and Olive, T. (2021). Methodological issues in literacy research across languages: Evidence from alphabetic orthographies. *Read. Res. Q.* 56, S351–S370. doi: 10.1002/rrq.407

Papadopoulos, T. C., Georgiou, G., and Kendeou, P. (2009a). Investigating the double-deficit hypothesis in Greek: Findings from a longitudinal study. *J. Learn. Disabil.* 42, 528–547. doi: 10.1177/0022219409338745

Papadopoulos, T. C., Georgiou, G. K., Kendeou, P., and Spanoudis, G. (2009b). *Standardization in Greek of the Das-Naglieri Cognitive Assessment System*. Cyprus: University of Cyprus.

Papadopoulos, T. C., Spanoudis, G., and Georgiou, G. K. (2016). How is RAN related to reading fluency? A comprehensive examination of the prominent theoretical accounts. *Front. Psychol.* 7:1217. doi: 10.3389/fpsyg.2016.01217

Papadopoulos, T. C., Spanoudis, G., and Kendeou, P. (2009c). *Early Reading Skills Assessment Battery (ERS-AB)*. Cyprus: University of Cyprus.

Philiastides, M. G., and Sajda, P. (2005). Temporal characterization of the neural correlates of perceptual decision making in human brain. *Cereb. Cortex* 16, 509–518. doi: 10.1093/cercor/bhi130

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage* 62, 816–847. doi: 10.1016/j.neuroimage.2012.04.062

Pugh, K. R., Mencl, W. E., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., et al. (2001). Neurobiological studies of reading and reading disability. *J. Commun. Disord.* 34, 479–492. doi: 10.1016/S0021-9924(01)00060-0

Rezvani, Z., Zare, M., Žarić, G., Bonte, M., Tijms, J., Van der Molen, M. W., et al. (2019). Machine learning Classification of Dyslexic Children based on EEG local network features. *bioRxiv[Preprint]*. doi: 10.1101/569996

Richlan, F., Kronbichler, M., and Wimmer, H. (2011). Meta-analyzing brain dysfunctions in dyslexic children and adults. *Neuroimage* 56, 1735–1742. doi: 10.1016/j.neuroimage.2011.02.040

Rijthoven, R., Kleemans, T., Segers, E., and Verhoeven, L. (2018). Beyond the phonological deficit: Semantics contributes indirectly to decoding efficiency in children with dyslexia. *Dyslexia* 24, 309–321. doi: 10.1002/dys.1597

Torppa, M., Parrila, R. K., Niemi, P., Lerkkanen, M. K., Poikkeus, A. M., and Nurmi, J. E. (2013). The double deficit hypothesis in the transparent Finnish orthography: A longitudinal study from kindergarten to grade 2. *Read. Writ.* 26, 1353–1380. doi: 10.1007/s11145-012-9423-2

van den Boer, M., van Bergen, E., and de Jong, P. F. (2014). Underlying skills of oral and silent reading. *J. Exp. Child Psychol.* 128, 138–151. doi: 10.1016/j.jecp.2014.07.008

Varga, V., Tóth, D., and Csépe, V. (2020). Orthographic-phonological mapping and the emergence of visual expertise for print: A developmental event-related potential study. *Child Dev.* 91, e1–e13. doi: 10.1111/cdev.13159

Wang, H., Liu, F., Dong, Y., and Yu, D. (2022). Entropy of eye movement during rapid automatized naming. *Front. Hum. Neurosci.* 16:945406. doi: 10.3389/fnhum.2022.945406

Wolf, M., and Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *J. Educ. Psychol.* 91, 415–438. doi: 10.1037/0022-0663.91.3.415

Ziegler, J. C., Perry, C., Jacobs, A. M., Ma-Wyatt, A., Ladner, D., and Schulte-Ko¨rne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *J. Exp. Child Psychol.* 86, 169–193. doi: 10.1016/S0022-0965(03)00139-5

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 29 June 2023
DOI 10.3389/fpsyg.2023.1136488

OPEN ACCESS

EDITED BY
Marijan Palmovic,
University of Zagreb, Croatia

REVIEWED BY
Xiaolu Wang,
Zhejiang University City College, China
Qiaoyun Liao,
Shanghai International Studies University, China

*CORRESPONDENCE
Jingxin Wang
✉ wjxpsy@126.com

RECEIVED 03 January 2023
ACCEPTED 12 June 2023
PUBLISHED 29 June 2023

CITATION
Chang M, Zhang K, Sun Y, Li S and Wang J (2023) The graded predictive pre-activation in Chinese sentence reading: evidence from eye movements.
Front. Psychol. 14:1136488.
doi: 10.3389/fpsyg.2023.1136488

The graded predictive pre-activation in Chinese sentence reading: evidence from eye movements

Min Chang[1], Kuo Zhang[2], Yue Sun[3,4], Sha Li[5] and Jingxin Wang[3,4]*

[1]School of Education Science, Nantong University, Nantong, China, [2]Department of Social Psychology, Nankai University, Tianjin, China, [3]Faculty of Psychology, Tianjin Normal University, Tianjin, China, [4]Academy of Psychology and Behavior, Tianjin Normal University, Tianjin, China, [5]School of Psychology, Fujian Normal University, Fuzhou, China

Previous research has revealed that graded pre-activation rather than specific lexical prediction is more likely to be the mechanism for the word predictability effect in English. However, whether graded pre-activation underlies the predictability effect in Chinese reading is unknown. Accordingly, the present study tested the generality of the graded pre-activation account in Chinese reading. We manipulated the contextual constraint of sentences and the predictability of target words as independent variables. Readers' eye movement behaviors were recorded via an eye tracker. We examined whether processing an unpredictable word in a solid constraining context incurs a prediction error cost when this unpredictable word has a predictable alternative. The results showed no cues of prediction error cost on the early eye movement measures, supported by the Bayes Factor analyses. The current research indicates that graded predictive pre-activation underlies the predictability effect in Chinese reading.

KEYWORDS

lexical predictability, contextual constraint, graded pre-activation, Chinese reading, eye movement

Introduction

Prediction is a fundamental principle of language processing (Clark, 2013). Efficient language comprehension depends on two streams of information, i.e., the top-down expectation and the bottom-up conceptual input. In speech comprehension, listeners could predict the content at the end of other speakers' turns to make efficient turn-taking using statistical regularities information in speech (Scott et al., 2009). In reading comprehension, readers could make use of contextual predictability information to facilitate word identification and semantic integration (for a review see Staub, 2015). A word's predictability, as measured by the word's cloze value, i.e., the proportion of participants who give this word in a non-speeded sentence completion task (Taylor, 1953), has been shown to influence reading times and saccadic behavior in reading tasks using the eye-tracking method of English, German, and Chinese (Rayner and Well, 1996; Kliegl et al., 2004; Rayner et al., 2005; Wang et al., 2010; Staub, 2015; Liu et al., 2018; Zhao et al., 2019; Chang et al., 2020a,b). Specifically, predictable words are easier to read, receive fewer and shorter fixations, and elicit longer progressive or incoming saccade length than unpredictable words, i.e., the word predictability effect.

However, the mechanisms for the predictability effect in Chinese reading have not been investigated previously. Thus, the present study aims to determine how prediction occurs, i.e., the mechanism of word predictability effect in Chinese reading.

Two competing theoretical accounts explain the mechanisms for predictability effects, each of which has different predictions for processing unexpected words (Luke and Christianson, 2016; for a review see Staub, 2015). First, the word prediction could be defined as an "all-or-none" process in which readers may maintain specific, discrete predictions of upcoming perceptual input, also termed *lexical prediction* (also see Delong et al., 2014). According to this *lexical prediction account*, strong constraining sentences support expectations for predictable words with much facilitation. Reading can be facilitative when readers encounter predictable words but slow down when readers encounter unpredictable words in a sufficient constraining context, i.e., producing the prediction error cost (Kutas et al., 2011; Luke and Christianson, 2016). For example, readers would predict the most probable word *gift* in the constraining sentence "*Today was Annie's birthday, her mother bought her a-*." This predictable word *gift* would be processed quickly as it matches readers' expectations. On contrary, readers might be surprised when encountering an unexpected word like *book*, then they would spend more time reading this unexpected word (i.e., prediction error cost) as they must suppress the activated *gift*. While a neutral constraining sentence like "When Annie went home, her mother brought her a-" provides little contextual information to readers. Thus, processing the unpredictable words would rarely incur prediction error cost as no predictable word is pre-activated. Therefore, according to the *lexical prediction account*, the comparison of processing unpredictable words between the constraining context and the neutral context would cause a prediction error cost.

Second, prediction in language comprehension could also involve *graded pre-activation* so readers make diffuse, cost-free, and ubiquitous pre-activation of likely upcoming input (Luke and Christianson, 2016; for reviews, see Staub, 2015; Kuperberg and Jaeger, 2016). Compared to the *lexical prediction account*, the key prediction of this account is that processing the unpredictable word would not incur a prediction error cost when the expected word is another more possible alternative in a strong constraining sentence. Because not only the predictable word but also the unpredictable word would be pre-activated before the perceptual input is encountered. In the neutral context of the above example, readers would pre-activate a set of words that suit the context, like *book*, *hat*, *skirt*, and *guitar*. Please notice that these words mentioned above are nouns, which could be pre-activated at syntactic or semantic representation even if the word identities are not. Readers may not be able to predict *gift*, but they can be confident that the upcoming word will be a noun or something that could be carried. Thus, even if people do not predict specific words, they could predict some aspects of future stimuli (Pickering and Gambi, 2018). Therefore, according to the *graded pre-activation account*, the comparison of processing unpredictable words between the constraining context and the neutral context would not cause a prediction error cost.

The graded pre-activation account has been well-demonstrated in English reading (for a review see Kuperberg and Jaeger, 2016), as evidenced by the reliable correlation between word predictability (measured as word surprizal or cloze probability) and processing times (Monsalve et al., 2012; Smith and Levy, 2013; Goodkind and Bicknell, 2018), N400 amplitude (Delong et al., 2005; Frank et al., 2015), or neural activity (Henderson et al., 2016). Specifically, the word predictability was inversely correlated with reading times (e.g., gaze duration in Goodkind and Bicknell, 2018), N400 amplitudes of words (Delong et al., 2005), and changes in brain activation levels in the temporal, parietal, occipital, cingulate, and frontal regions (Carter et al., 2019). In addition, Luke and Christianson (2016) conducted a large-scale survey that provided cloze values for words in the Provo Corpus. Their results showed that most words had a more-expected competitor but with no misprediction error cost. Even if the word identity was rarely predicted, its semantic and morphosyntactic information was predictable. These findings support the *graded prediction account* but not the specific *lexical prediction account*. The null prediction error cost (as the key opinion of graded pre-activation account) also has been demonstrated by Frisson et al. (2017) using a controlled-experimental design with an eye-tracking method using a corpus study with high ecological validity.

Frisson et al. (2017) jointly manipulated the contextual constraint of sentences and the cloze probability of target words to explore the cognitive mechanism of predictability effects in English. They compared the processing of the same unpredictable word (e.g., *chair*) in the constraining context (e.g., "The young nervous paratrooper jumped out of the *plane/chair* when he heard the shots") and the neutral context (e.g., "The tired movie maker was sleeping in the *plane/chair* when he was woken up by a scream") to test the prediction error cost. Also, the cloze values for unpredictable words in the constraining and neutral sentences were comparable. Their results showed significant word predictability effects and contextual constraint effects, but null prediction error cost in the early or later eye movement measures. This study firstly provided evidence from the controlled experimental design for the absence of a prediction error cost and further supported that the *graded pre-activation* but not the *lexical prediction account* underlies the mechanism of word predictability effects.

Notably, the null prediction error cost in constraining sentences might be due to the priming effect from the pre-target word area. The richer information preceding the target words might facilitate automatic priming to the target words in the strong constraining sentences but not the neutral sentences (see Kuperberg and Jaeger, 2016). Although whether there is an interference from the priming effect in predictive processing is unclear, it is recommendable to control the pre-target region to investigate the predictive processing, especially in Chinese such visually denser scripts.

For Chinese reading, there have been several studies investigating how the word predictability affects eye movement behaviors or interplays with other linguistic factors (Rayner et al., 2005; Wang et al., 2010; Liu et al., 2018; Zhao et al., 2019; Chang et al., 2020a,b). However, studies of Chinese to date have yet to investigate the mechanism of word predictability effects. Whether prediction error cost exists in Chinese reading is still being determined. Chinese scripts lack morphosyntactic information, which readers use as cues for prediction. Moreover, parafoveal processing is more efficient in Chinese than English (Vasilev and Angele, 2017). Thus, readers might heavily rely on bottom-up perceptual processing in Chinese reading. Such Chinese script

characteristics might make it hard to produce a specific word prediction in Chinese reading. Therefore, predictive processing might rely on graded pre-activation rather than lexical prediction. The present study aimed to provide experimental evidence for the *graded pre-activation account* in Chinese reading.

Accordingly, the present study was a follow-up to a previous study (Frisson et al., 2017) but further made more rigid control of the pre-target context. There is no explicit visual marker in Chinese to demarcate work boundaries (Li et al., 2015). Characters, the component of words, are created from differing numbers of strokes. These characteristics, therefore, bring about the increased visual density in this language and lead to deeper parafoveal pre-processing, as demonstrated by the well-established semantic preview effect in Chinese, which is equivocal in English (Zhou et al., 2013; Rayner et al., 2014). The different content immediately before the target words might influence the processing of target words differently (Reichle et al., 2003). Moreover, early eye-tracking studies have found that transitional probabilities (i.e., the statistical likelihood that word N will follow word N-1) between word N-1 and word N influence fixation times on word N (McDonald and Shillcock, 2003; Frisson et al., 2005; Wang et al., 2010). Hence, it is necessary to control the influence of the pre-target region across conditions.

Given the above considerations, we manipulated the contextual constraint and word predictability to address the question using a natural sentence reading task, consistent with Frisson et al. (2017). However, we went further by constructing compound sentences, with the first half-sentences controlling contextual constraint and the second half-sentences having identical content at least three characters before the target words to control the possible priming effect or pre-target influence on the target words. We obtained the contextual constraint effect, word predictability effect, and the prediction error cost by three comparisons: (1) constraining context-unpredictable (CU) vs. constraining context-predictable (CP), testing the word predictability effect; (2) neutral context–predictable (NP) vs. constraining context-predictable (CP), testing the contextual constraining effect, and (3) constraining context-unpredictable (CU) vs. neutral context-unpredictable word (NU), testing the prediction error cost. According to the *lexical prediction account*, unpredictable word processing in the constraining context would result in extra prediction error cost but not in the neutral context. Thus, we compared CU and NU to evaluate the prediction error cost, as Frisson et al. (2017).

We expected to find the typical word predictability effect, i.e., predictable words yielding shorter reading times than unpredictable words. We also expected the significant contextual constraint effect, i.e., the strong constraining sentences but not the neutral sentences make target words read faster. The contextual effects and the standard word predictability effects in the first-pass reading measures demonstrated that we manipulated the two factors successfully. However, the two effects mentioned above are not key evidences to our hypothesis. The prediction error cost (CU vs. NU) is the primary evidence for distinguishing the two accounts. Specifically, if readers spent longer time on reading unpredictable word in CU than in NU (i.e., significant prediction error cost), then the result supported the *lexical prediction account*, otherwise (null prediction error cost) *supported the graded pre-activation account*.

Materials and methods

Ethics approval

The study was approved by the research ethics committee at the Tianjin Normal University and conducted according to the Declaration of Helsinki principles.

Participant

Forty-four undergraduates aged 18–26 years (M = 20.5 years, 34 female) from the author's university participated in the eye-tracking experiment for remuneration. The participant number was the same as Frisson et al. (2017). All were native Chinese readers, screened for normal acuity (more excellent than 20/40 in Snellen values) using a Tumbling E eye chart (Taylor, 1978), and naive to the purpose of the experiment. Informed consent was obtained from all individual participants in the study.

Design and stimuli

We constructed 48 sets of sentence frames, a number larger than Frisson et al. (2017). The experiment used a within-subjects design with the factors of sentence constraint (Constraining, Neutral) and word predictability (Predictable, Unpredictable) as independent variables. See **Table 1**, each sentence frame had a strong constraining sentence and a neutral sentence. The first half-sentence was manipulated to control the contextual constraint; predictable or unpredictable target words were inserted in the middle of the second half-sentence. At least three characters before target words were identical in the constraining and neutral conditions (excluding only five sets of sentences). As stated in the introduction, we conducted three comparisons to obtain the contextual constraint effect, word predictability effect, and prediction error cost. The most crucial comparison was the third one, i.e., constraining context-unpredictable word (CU) vs. neutral context-unpredictable word (NU), testing the prediction error cost. The significant prediction error cost indicates that an unexpected word in a constraining context with a predictable alternative will incur a processing cost, which supports the *lexical prediction account*.

In the cloze test, students were given the sentences truncated immediately before the target word and asked to provide the next word in the sentences. Twenty-two college students who did not participate in the experiment completed the cloze test. A predictable or unpredictable word was embedded in the constraining context (labeled CP and CU, respectively, see **Table 1**). The same two words were embedded in the corresponding neutral context and embedded in the constraining context. Given that the two target words, such as model/girl in the neutral context, were the same as targets in the constraining context, we labeled them as NP and NU, following Frisson et al. (2017). Please note that NP and NU were unpredictable because the neutral context did not provide strong word constraints. The mean cloze probability of the target words in the four conditions (CP, CU, NP, and NU) were 0.75 (SD = 0.16), 0.02 (SD = 0.04), 0.05 (SD = 0.06), and

TABLE 1 An example stimulus.

Condition	The first half-sentence	The second half-sentence
Constraining context–predictable (CP)	T台上的刘雯小姐优雅地款款走来，	这位走向世界的中国**模特**让外国友人看到了东方之美。
Constraining context-unpredictable (CU)	T台上的刘雯小姐优雅地款款走来，	这位走向世界的中国**女孩**让外国友人看到了东方之美。
Neutral context-predictable (NP)	舞台上的邓琦小姐散发着优雅知性的气质，	这位走向世界的中国**模特**让外国友人看到了东方之美。
Neutral context -unpredictable (NU)	舞台上的邓琦小姐散发着优雅知性的气质，	这位走向世界的中国**女孩**让外国友人看到了东方之美。

Target words are shown in bold. The constraining sentence translates as "Miss Liu Wen on the runway comes gracefully, and this Chinese *model/girl* who is famous around the world shows foreign friends the beauty of the East." The neutral sentence translates as "Miss Deng Qi on the stage exudes an elegant and intellectual temperament; this Chinese *model/girl* who is famous around the world shows foreign friends the beauty of the East." Please note that the target word, such as *模特(model)* in the first condition of neutral context, was the same as in the CP condition. Thus we labeled it as NP. The NP and NU did not differ in predictability, word frequency, and complexity.

0.04 (SD = 0.08), respectively. In the constraining context, t-tests showed that the cloze values for CP were significantly higher than for CU [$t(94) = 30$, $p < 0.001$]. In the neutral context, the two unpredictable targets had comparable cloze values [$t(94) = 1.07$, $p = 0.288$]. Importantly, the cloze values for the same unpredictable word (such as *girl*) in constraining and neutral contexts were comparable [$t(94) = 1.52$, $p = 0.13$].

The two target words in one sentence frame were matched for word frequency [Cai and Brysbaert, 2010; Predictable: M = 64/million, SD = 80; Unpredictable: M = 44/million, SD = 104; $t(94) = 1.06$, $p = 0.291$] and the whole word complexity in strokes [Predictable: M = 17.41, SD = 5.11; Unpredictable: M = 15.88, SD = 4.97; $t(94) = 1.50$, $p = 0.137$]. Forty participants evaluated sentences naturalness (using a 7-point scale, ranging from 1 = entirely unnatural to 7 = entirely natural). The average ratings were 5.41 (SD = 0.74), 5.31 (SD = 0.71), 5.32 (SD = 0.66), and 5.20 (SD = 0.7) for each conditions, respectively. The ANOVA analysis showed that the four conditions were comparable in naturalness [$F_{(3,188)} = 0.85$, $p = 0.468$].

We adopted a counterbalanced design in which the experimental sentences were divided into four lists, and one version of each sentence frame was in one list. Each participant read one list with equal numbers of sentences in each condition. Each list also included 40 filler sentences and began with six practice sentences. Eleven participants were randomly allocated to each list.

Apparatus and procedure

An SR Eyelink 1000 plus eye tracker tracked right-eye movements during binocular viewing at 1000 Hz. Stimuli were displayed in Song 32-point font as black-on-white text on a high-resolution (1920 × 1080 pixels) monitor with a fresh rate of 60 Hz. At 65 cm viewing distance, each character subtended 1° and so was of normal size for reading.

Participant took part individually and was instructed to read normally and for comprehension. At the start of the experiment, a 3-point horizontal calibration procedure was performed across the same line as each sentence presentation (ensuring 0.30° or better spatial accuracy for all participants). Calibration accuracy was checked before each trial and the eye-tracker recalibrated as required to maintain high spatial accuracy. At the start of each trial, a fixation square equal in size to one character was presented on the left side of the screen. Once the participant fixated on this location, the first half-sentence was presented with the first character replacing the square. Participant pressed the space key once they finished reading the first half-sentence. Then the same fixation square was presented again at the same position and disappeared once the participant fixated it, then the second half-sentence was presented. Participant pressed a response key once they finished reading the second half-sentence. This was replaced by a comprehension question requiring a yes/no button-press response on 25% of trials. The experiment lasted approximately 30 min for each participant.

Data analysis

Accuracy for answering comprehension questions was high for all participants (M = 84%, SD = 0.06, range = [73%, 95%]). We output the data of the second half-sentences and thus removed the data based on the second half-sentences. Following standard procedures, short (< 80 ms) and long (> 1200 ms) fixations were removed. Trials with head-movement, tracking-loss, or error were excluded, which affected seven trials (0.3%), as were trials for sentences receiving fewer than six fixations, which affected 99 trials (4.7%). In total, 5% of trials (106) were removed. The remaining data were analyzed by linear mixed-effects models (LMEs; Baayen et al., 2008) for continuous variables and generalized mixed-effects models for binomial variables, using the lme4 package (Version 1.1-21; Bates et al., 2015) in R (R Development Core Team, 2016). For all measures, models with the maximum random-effects structure were used (Barr et al., 2013), with the three comparisons as fixed factors and participant and stimuli as crossed random effects. If models did not converge, the random-effects structure was reduced by first trimming this for stimuli. Log-transformed fixation-time effects are reported alongside untransformed means. Following convention, t/z values > 1.96 were considered significant.

Results

We expected significant word predictability effects and contextual constraint effects on the early eye movement measures and explored whether unpredictable words in constraining sentences incur processing costs on early word identification or later semantic integration. Thus, consistent with Frisson et al. (2017), we reported four measures of first-pass reading for the target words, i.e., the word-skipping (SKIP, probability of not fixating a word during first-pass reading), first-fixation duration (FFD, duration of the first fixation on a word during first-pass reading), single-fixation duration (SFD, duration of the first fixation on a word receiving only one first pass fixation), gaze duration (GD, sum of all first pass fixations on a word). We also reported three measures concerning later semantic integration, i.e., regressions-out rate (RO, probability of first-pass regression from a word),

TABLE 2 Means and standard errors for target word measures ($M \pm SE$).

Measures	Constraining		Neutral	
	Predictable	Unpredictable	Predictable	Unpredictable
Skipping (%)	31 (2)	27 (2)	25 (2)	24 (2)
FFD (ms)	236 (5)	250 (5)	237 (4)	249 (5)
SFD (ms)	235 (5)	246 (5)	238 (4)	245 (5)
GD (ms)	251 (6)	272 (7)	264 (6)	274 (7)
RPD (ms)	307 (14)	333 (12)	307 (11)	361 (16)
RO (%)	10 (2)	14 (2)	11 (2)	16 (2)
TRT (ms)	348 (11)	371 (11)	351 (10)	384 (12)

regression path duration (RPD, the sum of all fixation durations beginning with the initial fixation on the target word and ending when the eyes exited the word to the right, including time spent rereading earlier words and time spent rereading the word itself) and total reading time (TRT, sum of all fixations on a target word). Target word means were shown in **Table 2**, and statistical effects were summarized in **Table 3**.

Word predictability effect and contextual constraining effect

We observed significant word predictability effects (CP vs. CU) and contextual constraining effects (CP vs. NP) on the first pass reading measures (see **Figure 1**). The word predictability effects, significant on FFD, SFD, and GD, were due to longer reading times for CU than CP conditions (FFD: $b = 0.06$, CI = [0.02, 0.11], $SE = 0.02$, $t = 2.63$; SFD: $b = 0.05$, CI = [0.01, 0.1], $SE = 0.02$, $t = 2.2$; GD: $b = 0.08$, CI = [0.03, 0.13], $SE = 0.03$, $t = 2.89$).[1] The comparison between CP and NP revealed significant contextual constraining effects on the early skipping rate ($b = -0.34$, CI = [−0.63, −0.05], $SE = 0.15$, $z = -2.3$) and gaze duration ($b = 0.06$, CI = [0.01, 0.11], $SE = 0.03$, $t = 2.16$). Readers made more skipping and shorter first-pass fixation durations on the target word. The clear word predictability and contextual constraining effects indicated that we manipulated the two factors successfully.

Prediction error cost

Most crucially, the prediction error cost was not significant on all the measures ($|z/t|\ s < 1.3$), i.e., an unexpected word did not incur processing cost in the constraining context with a predictable alternative, compared to the same target word in the neutral context.

We conducted Bayes factors analyses (Kass and Raftery, 1995) to determine the strength of the evidence for the null prediction error cost on the first-pass fixation time measures. The analyses were conducted using the lmBF function within the BayesFactor package (Version 0.9.12-4.2; Morey et al., 2015; R Development Core Team, 2016). Analyses were conducted with scaling factor for g-priors set to 0.5, using 10,000 Monte Carlo iterations. We first computed the Bayes Factor for a model with a fixed effect of prediction error cost (CU vs. NU) and random participant and item intercepts of FFD, SFD, and GD, i.e., BF_1. Then we computed Bayes Factor for a model with only random participant and item intercepts, i.e., BF_0. The critical value was the ratio of BF_1 and BF_0, i.e., BF_{10}, it is itself a Bayes Factor comparing the model with an effect of prediction error cost and participant and item intercepts, to a model with the only participant and item intercepts. According to Vandekerckhove et al. (2015), Bayes Factors ($BF_{10} < 1/3$) were taken to provide moderate to strong evidence for the null model. Thus, the present results (FFD, $BF_{10} = 0.11$; SFD, $BF_{10} = 0.03$; GD, $BF_{10} = 0.27$) provided moderate to strong evidence for the null model, i.e., the null prediction error cost.

Discussion

In the present experiment, we manipulated the contextual constraint of sentences and word predictability to investigate whether there is a prediction error cost in Chinese reading. We tested the prediction error cost by comparing the processing of unpredictable words between constraining contexts and neutral contexts (i.e., CU vs. NU). The results showed significant contextual effects and standard word predictability effects in the early stage of word processing, with shorter reading times (FFD, SFD, and GD) for more predictable words, which is in line with previous findings from Chinese studies (Rayner et al., 2005; Wang et al., 2010; Liu et al., 2018; Zhao et al., 2019; Chang et al., 2020a,b). Importantly, no significant prediction error cost was observed across a wide range of eye movements, i.e., the reading is not disruptive if the readers encounter the unpredictable word in a strong constraining sentence with a predictable alternative, supported by the Bayes factor analyses. This result resonated with findings from English studies (Frisson et al., 2005, 2017; Luke and Christianson, 2016). In particular, the findings suggested that readers make diffuse and graded pre-activation of likely upcoming input.

The current experiment adopted a similar design as Frisson et al. (2017). The key comparison between unpredictable words in the constraining and neutral sentences showed no prediction error cost on the fixation duration measures both for Frisson et al. and the

1 The word predictability effect was also significant on RPD while we did not mention it in the Results and Discussion. As the results on RPD, RO, and TRT might represent a mixture of predictability effect and semantic integrative effect. We want to obtain the clear and genuine predictability effect. Following the tradition of eye movement research, however, we reported these later eye movement measures in the table which could be accessible for other researchers for meta-analysis. Thus, we did not mention and discuss these later eye movement measures in sections "Results and Discussion."

TABLE 3 Summary of statistical effects (continuous variables were log-transformed).

Measures	Comparison	*b*	CI	*SE*	*t/z*	*p*
SKIP	Intercept	−1.1	[−1.36, −0.86]	0.12	−8.93	<0.001
	Predictability	−0.2	[−0.48, 0.09]	0.15	−1.34	0.182
	Constraint	−0.34	[−0.63, −0.05]	0.15	−2.3	0.022*
	Prediction error cost	−0.19	[−0.49, 0.10]	0.15	−1.29	0.198
FFD	Intercept	5.43	[5.38, 5.47]	0.02	236.02	<0.001
	Predictability	0.06	[0.02, 0.11]	0.02	2.63	0.009*
	Constraint	0.02	[−0.03,0.06]	0.02	0.73	0.464
	Prediction error cost	−0.02	[−0.06,0.03]	0.02	−0.68	0.494
SFD	Intercept	5.42	[5.37, 5.47]	0.02	228.39	<0.001
	Predictability	0.05	[0.01, 0.10]	0.02	2.2	0.028*
	Constraint	0.03	[−0.02, 0.08]	0.02	1.32	0.187
	Prediction error cost	−0.01	[−0.06, 0.03]	0.02	−0.58	0.565
GD	Intercept	5.48	[5.43, 5.54]	0.03	209.57	<0.001
	Predictability	0.08	[0.03, 0.13]	0.03	2.89	0.004*
	Constraint	0.06	[0.01, 0.11]	0.03	2.16	0.031*
	Prediction error cost	−0.01	[−0.06, 0.04]	0.03	−0.26	0.793
RPD	Intercept	5.61	[5.54, 5.68]	0.03	169.17	<0.001
	Predictability	0.11	[0.04, 0.19]	0.04	3.11	0.002*
	Constraint	0.05	[−0.02,0.12]	0.04	1.28	0.202
	Prediction error cost	0.03	[−0.04, 0.10]	0.04	0.94	0.346
RO	Intercept	−2.05	[−2.32, −1.81]	0.12	−16.37	<0.001
	Predictability	0.44	[−0.03, 0.92]	0.24	1.85	0.064
	Constraint	0.15	[−0.34, 0.64]	0.25	0.6	0.546
	Prediction error cost	0.18	[−0.22, 0.60]	0.21	0.89	0.372
TRT	Intercept	5.72	[5.66, 5.79]	0.03	171.51	<0.001
	Predictability	0.05	[−0.01, 0.12]	0.04	1.55	0.121
	Constraint	0.02	[−0.05, 0.09]	0.03	0.67	0.506
	Prediction error cost	0.03	[−0.04, 0.09]	0.03	0.76	0.448

Asterisks indicate significant effects where $t/z > 1.96$. CI = 95% confidence Interval.

present study. This is what we and Frisson et al. (2017) have found in common, indicating that the *lexical prediction account* would not seem able to account for the predictability effect both in English and Chinese. Notably, the present study differed from Frisson et al. (2017) on the numerical trend. They found a numerical trend in the opposite direction, i.e., the processing advantage for unpredictable words in constraining sentences compared to neutral sentences. Although this processing benefit did not reach significance on reading time measures, this trend was significant in the first pass regression rate ($z = -2.03$). The significant benefit of unpredictable words in constraining sentences might be due to the semantic priming effect or the transitional probability effect, i.e., the statistical likelihood that a word preceding the target might influence target word processing.

Like Frisson et al. (2017) study, the present study provided clear and strong evidence for null prediction error cost ($t/z < 1.29$). Unlike Frisson et al. (2017) we did not find significant benefits for unpredictable words in constraining sentences when controlling the pre-target region, providing stronger support for *graded pre-activation account*. The characteristics of the Chinese language could explain this. Chinese lacks overt cues (markers for number, gender, the tense of verbs, and case) to syntactic structure, which a reader utilizes to produce predictions about upcoming stimuli in English (see Kuperberg and Jaeger, 2016 for a review). Furthermore, the word predictability is lower in Chinese than in English, as shown by the comparison between cloze probability reported by Pan et al. (2021) in Beijing Sentence Corpus (BSC) and that by Luke and Christianson (2016) in Provo Corpus. The grand mean of cloze scores for the words in BSC is 0.07, far less than that reported in Luke and Christianson ($M = 0.13$). Thus, the sentence constraint in Chinese may be weaker than that in English. It is reasonable that we found more consistent results on the several eye movement measures.

The findings are consistent with the multi-representational hierarchical generative architecture, which views prediction as a graded and probabilistic phenomenon

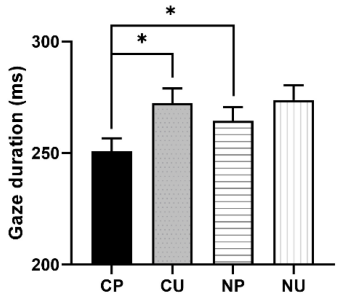


FIGURE 1
Context-predictable (CP), CU, NP, and NU represent constraining context with predictable word, constraining context with unpredictable word, neutral context with predictable word, and neutral context with unpredictable word, respectively. The contrast between CP and NP represents the contextual constraining effect; the contrast between CP and CU represents the word predictability effect; the contrast between CU and NU represents the prediction error cost. Figure describes the gaze duration in each condition. Asterisks indicate significant effect where $t > 1.96$.

(Kuperberg and Jaeger, 2016). Also, this architecture suggests distinguishing between predictive pre-activation and pre-activation through priming. The present study attempted to control interference from the priming effect across conditions by constructing compound sentences in which the first half-sentences controlled the contextual constraint and the second half-sentences were identical at least three characters before the target words. Thus, the content of the pre-target region was identical in the constraining and neutral sentences. The null prediction error cost on the first pass reading measures and the later eye movement measures suggest that encountering an unexpected word in a constraining sentence does not interrupt early lexical identification and later semantic integration. Readers pre-activate not only one specific item but a range of possible words. The present study confirmed the graded pre-activation mechanism of predictive processing in Chinese reading.

Limitations and future directions

The study had one limitation. The number of participants in the cloze task might influence the cloze value of words. There is a positive correlation between the number of participants and the precision of word's cloze value. Our present study recruited 22 participants for the cloze task. Although we successfully balanced the cloze values between CU and NU, however, the sample size might be not big enough provide a precise cloze value of a word.

Thus, future studies could recruit as many participants as possible to obtain more precise word cloze value. Besides, cross-linguistic studies are highly needed to explore how linguistic characteristics (e.g., word space, word length, and complexity) influence predictive language processing. In addition, to improve the external validity, studies about predictive language comprehension of special readers (e.g., non-native speakers, children with dyslexia, and older adults) are needed. These studies will inform us of the mechanism of reading difficulty for non-native speakers, children with dyslexia, and older adults.

Conclusion

In summary, we conducted an eye-tracking experiment to investigate whether processing an unpredictable word incurs prediction error cost when there is a predictable alternative. The null prediction error cost supports that the graded pre-activation account underlies the word predictability effect in Chinese reading.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the research Ethics Committee at Tianjin Normal University and conducted according to the Declaration of Helsinki principles. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MC and JW designed the experiment and wrote the manuscript. MC and YS experimented and analyzed the data. KZ and SL provided good suggestions. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by grants from the National Natural Science Foundation of China to JW (81771823) and Fujian Social Science Planning Project under Grant to SL (FJ2020C071).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

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

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Cai, Q., and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One* 5:e10729. doi: 10.1371/journal.pone.0010729

Carter, B. T., Foster, B., Muncy, N. M., and Luke, S. G. (2019). Linguistic networks associated with lexical, semantic and syntactic predictability in reading: a fixation-related fMRI study. *NeuroImage* 189, 224–240. doi: 10.1016/j.neuroimage.2019.01.018

Chang, M., Hao, L., Zhao, S., Li, L., Paterson, K. B., and Wang, J. (2020a). Flexible parafoveal encoding of character order supports word predictability effects in Chinese reading: evidence from eye movements. *Attent. Percept. Psychophys.* 82, 2793–2801. doi: 10.3758/s13414-020-02050-x

Chang, M., Zhang, K., Hao, L., Zhao, S., McGowan, V. A., Warrington, K. L., et al. (2020b). Word predictability depends on parafoveal preview validity in Chinese reading. *Vis. Cogn.* 28, 33–40. doi: 10.1080/13506285.2020.1714825

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Delong, K. A., Troyer, M., and Kutas, M. (2014). Pre-processing in sentence comprehension: sensitivity to likely upcoming meaning and structure. *Lang. Linguistics Compass* 8, 631–645. doi: 10.1111/lnc3.12093

Delong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121. doi: 10.1038/nn1504

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* 140, 1–11. doi: 10.1016/j.bandl.2014.10.006

Frisson, S., Harvey, D. R., and Staub, A. (2017). No prediction error cost in reading: evidence from eye movements. *J. Mem. Lang.* 95, 200–214. doi: 10.1016/j.jml.2017.04.007

Frisson, S., Rayner, K., and Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *J. Exp. Psychol.* 31, 862–877. doi: 10.1037/0278-7393.31.5.862

Goodkind, A., and Bicknell, K. (2018). "Predictive power of word surprisal for reading times is a linear function of language model quality," in *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL)*, Salt Lake City, UT, 10–18. doi: 10.18653/v1/w18-0102

Henderson, J. M., Choi, W., Lowder, M. W., and Ferreira, F. (2016). Language structure in the brain: a fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage* 132, 293–300. doi: 10.1016/j.neuroimage.2016.02.050

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* 16, 262–284. doi: 10.1080/09541440340000213

Kuperberg, G. R., and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31, 32–59. doi: 10.1080/23273798.2015.1102299

Kutas, M., DeLong, K. A., and Smith, N. J. (2011). "A look around at what lies ahead: prediction and predictability in language processing," in *Predictions in the brain: using our past to generate a future*, ed. M. Bar (Oxford: Oxford University Press), 190–207. doi: 10.1093/acprof:oso/9780195395518.003.0065

Li, X., Zang, C., Liversedge, S. P., and Pollatsek, A. (2015). "The role of words in Chinese reading," in *The Oxford handbook of reading*, eds A. Pollatsek and R. Treiman (New York, NY: Oxford University Press), 232–244.

Liu, Y., Guo, S., Yu, L., and Reichle, E. D. (2018). Word predictability affects saccade length in Chinese reading: an evaluation of the dynamic-adjustment model. *Psychon. Bull. Rev.* 25, 1891–1899. doi: 10.3758/s13423-017-1357-x

Luke, S. G., and Christianson, K. (2016). Limits on lexical prediction during reading. *Cogn. Psychol.* 88, 22–60. doi: 10.1016/j.cogpsych.2016.06.002

McDonald, S. A., and Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vis. Res.* 43, 1735–1751. doi: 10.1016/S0042-6989(03)00237-2

Monsalve, I. F., Frank, S. L., and Vigliocco, G. (2012). "Lexical surprisal as a general predictor of reading time," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, 398–408.

Morey, R. D., Rouder, J. N., Jamil, T., and Morey, M. R. D. (2015). *Package 'bayesfactor'*. Available online at: https://cran.r-project.org/web/packages/BayesFactor/ (accessed June 10, 2015).

Pan, J., Yan, M., Richter, E. M., Shu, H., and Kliegl, R. (2021). The Beijing sentence corpus: a Chinese sentence corpus with eye movement data and predictability norms. *Behav. Res. Methods* 54, 1989–2000. doi: 10.3758/s13428-021-01730-2

Pickering, M. J., and Gambi, C. (2018). Predicting while comprehending language: a theory and review. *Psychol. Bull.* 144, 1002–1044. doi: 10.1037/bul0000158

R Development Core Team (2016). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rayner, K., Li, X., Juhasz, B. J., and Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychon. Bull. Rev.* 12, 1089–1093. doi: 10.3758/BF03206448

Rayner, K., Schotter, E. R., and Drieghe, D. (2014). Lack of semantic parafoveal preview benefit in reading revisited. *Psychon. Bull. Rev.* 21, 1067–1072. doi: 10.3758/s13423-014-0582-9

Rayner, K., and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: a further examination. *Psychon. Bull. Rev.* 3, 504–509. doi: 10.3758/BF03214555

Reichle, E. D., Rayner, K., and Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: comparisons to other models. *Behav. Brain Sci.* 26, 445–476. doi: 10.1017/S0140525X03000104

Scott, S. K., Mcgettigan, C., and Eisner, F. (2009). A little more conversation, a little less action - candidate roles for motor cortex in speech perception. *Nat. Neurosci.* 10, 295–302. doi: 10.1038/nrn2603

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: critical review and theoretical interpretation. *Lang. Linguistics Compass* 9, 311–327. doi: 10.1111/lnc3.12151

Taylor, H. R. (1978). Applying new design principles to the construction of an illiterate E chart. *Am. J. Optom. Physiol. Opt.* 55, 348–351. doi: 10.1097/00006324-197805000-00008

Taylor, W. L. (1953). "Cloze Procedure": a new tool for measuring readability. *J. Q.* 30, 415–433. doi: 10.1177/107769905303000401

Vandekerckhove, J., Matzke, D., and Wagenmakers, E. J. (2015). "Model comparison and the principle of parsimony," in *The Oxford handbook of computational and mathematical psychology*, eds J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Oxford: Oxford University Press), 300–319.

Vasilev, M. R., and Angele, B. (2017). Parafoveal preview effects from word N + 1 and word N + 2 during reading: a critical review and Bayesian meta-analysis. *Psychon. Bull. Rev.* 24, 666–689. doi: 10.3758/s13423-016-1147-x

Wang, H.-C., Pomplun, M., Chen, M., Ko, H., and Rayner, K. (2010). Estimating the effect of word predictability on eye movements in Chinese reading using latent semantic analysis and transitional probability. *Q. J. Exp. Psychol.* 63, 1374–1386. doi: 10.1080/17470210903380814

Zhao, S., Li, L., Chang, M., Xu, Q., Zhang, K., Wang, J., et al. (2019). Older adults make greater use of word predictability in Chinese reading. *Psychol. Aging* 34, 780–790. doi: 10.1037/pag0000382

Zhou, W., Kliegl, R., and Yan, M. (2013). A validation of parafoveal semantic information extraction in reading Chinese. *J. Res. Read.* 36(Suppl.1), S51–S63. doi: 10.1111/j.1467-9817.2013.01556.x

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 29 June 2023
DOI 10.3389/fcomm.2023.1176986

OPEN ACCESS

EDITED BY
Marijan Palmovic,
University of Zagreb, Croatia

REVIEWED BY
Kristina Cergol,
University of Zagreb, Croatia
Ying Zhao,
Inner Mongolia University of Science and Technology, China

*CORRESPONDENCE
Rurik Tywoniw
✉ rtywoniw@illinois.edu

RECEIVED 01 March 2023
ACCEPTED 15 May 2023
PUBLISHED 29 June 2023

CITATION
Tywoniw R (2023) Compensatory effects of individual differences, language proficiency, and reading behavior: an eye-tracking study of second language reading assessment. *Front. Commun.* 8:1176986. doi: 10.3389/fcomm.2023.1176986

Compensatory effects of individual differences, language proficiency, and reading behavior: an eye-tracking study of second language reading assessment

Rurik Tywoniw*

Department of Linguistics, University of Illinois at Urbana-Champaign, Champaign, IL, United States

Reading in a second language (L2) is a complex process that incorporates linguistic knowledge and literacy abilities, as well as strategic competence to approach different types of reading tasks depending on reading goals. However, much of the previous research was limited to correlational studies and focused on the relative contribution of broad categories of L2 proficiency and first-language (L1) literacy to L2 reading comprehension. However, investigations into L2 reading performance can benefit from advances in real-time, concurrent data collection methodologies such as eye-tracking. This study utilized eye-tracking methods to examined L2 reading comprehension of 102 readers across three different reading tasks [Cloze reading, Multiple-choice (MC) quiz, and reading-to-summarize], comparing the comprehension scores to L2 proficiency, individual differences (reasoning, working memory, motivation) and reading behavior (eye-tracking metrics related to attention to reading texts and tasks, length of fixations). Results indicate that the score on each task could be modeled each using a different mix of predictors, with the cloze task being most strongly predicted and the MC task being least predicted. The Summary task was in-between, but with a highly interpretable model. Interactions between fixation duration and cognitive abilities were found, showing how efficient fixation is generally important for comprehension, but the impact can be compensated for with motivation and reasoning ability.

KEYWORDS

second language reading, language assessment, eye-tracking, English for academic purposes (EAP), language learning

Introduction

For multilingual readers and language learners, reading comprehension ability has been conceptualized as a product of language proficiency: learners reach a threshold of reading ability and can then transfer first language (L1) literacy skills (including comprehension monitoring, activating strategies, and integration of information across pieces of texts) into their second language (L2) reading (Koda, 1988, 1990). Features of reading comprehension that are not related to linguistic proficiency are often overlooked for multilingual readers. However, for many academic language learners in the modern era, advanced reading skills many develop uniquely for an L2 which is the primary language of academic engagement. As such, it remains unclear how features of reading comprehension processes which play a role in monolingual readers' comprehension, such as real-time reading behavior and individual differences, contribute to reading comprehension for multilingual readers. This

lack of understanding poses a threat to L2 reading assessment validity. Bachman and Palmer (1996) in their test-authenticity argument, state that use of a language test is justified when we can "demonstrate that performance on language tests corresponds to language us in specific domains other than the language test itself" (p. 23). To better understand the factors which influence reading comprehension performance for multilingual academic readers, it is necessary to compare factors of language ability, individual differences, and real-time reading, as well as comparing these factors' influence on performance on varied reading tasks which may elicit different skills and abilities.

In this study, three measures of reading comprehension were analyzed: multiple-choice questions (MC), cloze tasks, and summary tasks. Completion of these tasks was analyzed under the lens of text-reading behavior. Task differences were examined using eye-movement behavior (eye-tracking) variables which were compared with score (described with more specificity in the methods section). Scores were predicted with statistical modeling using eye-movement metrics, L2 proficiency and individual difference variables: reading speed, working memory, reasoning, and motivation. This research will help the field of reading comprehension assessment further understand the cognitive and construct validity of these assessment tasks. Additionally, this research will shed light on how the influences on reading ability (individual differences, language proficiency, and real-time reading behavior) interact with each other and can be used to compensate for weaknesses.

Literature review

L2 reading and reading assessment

The validity of L2 reading tests hinges on how well tests target different aspects of the reading process. Models of reading often include both lower-order and higher-order skills. Key aspects of lower-level reading processes are grapho-phonemic processing, morphological awareness, word recognition, and syntactic parsing, with each lower-level process facilitating the recognition of words on the page (Perfetti, 2007). Much of the lower-order skills in L2 reading are developed alongside general L2 proficiency. Higher-level processing is seen as having two levels (Kintsch, 1998; Grabe, 2009): a text base comprehension level, where a reader creates a model of ideas and propositional content found in a text, and a situation model level, where the overall meaning of a text is constructed by the reader through connecting propositions and relating content to background knowledge and reading context. L2 research has been more agnostic regarding the development of higher-order skills, believing much of this to be the recipient of L1 literacy transfer (Koda, 1988).

In general, L2 reading scholars have acknowledged that not every predictor of successful comprehension needs to be activated at once during reading. Early conceptions of this phenomenon considered L2 reading to be broken down into coarse categories of skills: L1 literacy and L2 language proficiency, and deficits in one category could be compensated for with strengths in the other (Bernhardt, 2005). This view was expanded beyond the broad categories of L1 literacy and L2 language ability to include other potential compensatory strengths such as reading strategy knowledge and background knowledge (McNeil, 2011, 2012) in line with Stanovich's (1980) postulation that individuals will rely on multiple top-down and bottom-up resources as needed to achieve comprehension. Urquhart and Weir (2014) highlight goal-setting as an important aspect of reading ability, noting that modifying one's reading behavior based on the reading purpose is important. In other words, the type of reading task will influence the skills and behavior necessary to complete the task. This idea is expanded in the Reading as Problem Solving Model (RESOLV; Rouet et al., 2017) wherein a reader constructs a representation of a text with respect to the reading purpose and task at hand. Readers moderate the speed of reading and the level of attention to the text depending on whether the reader is skimming for gist (faster pace, global attention), scanning for details (faster pace, local attention), reading for informational purposes (slower pace, global attention), or having processing difficulty (slower pace, local attention) (Carver, 1997; Grabe, 2009). Understanding these factors and how this is elicited by reading tasks is important for designing effective measures of reading comprehension (Alderson, 2000; Borsboom, 2005).

However, it is difficult to observe reading behavior, let alone strategic reading. Part of why the previous debate about how L2 reading and whether it was more derived from L2 proficiency or L1 literacy came from this methodological difficulty in observing reading behavior. Reading abilities of either order have been difficult to measure directly, and as such, cognitive validity of reading tests could only be indirectly examined. That is until more sophisticated methods for tapping into cognitive processes of reading, such as eye-tracking became available (Conklin et al., 2018). Now, behavior related to both lower-order and higher-order reading abilities can be somewhat more directly observed.

Eye-tracking in second language acquisition

Observing reading processes and their contribution to successful comprehension has been a goal of Second Language Acquisition (SLA) research, but historically there have been few means by which to observe cognition in real time. Investigations into the processes which lead to successful comprehension have been usually been *post-hoc* in nature, but concurrent methods, such as eye-tracking, have become more commonplace (Godfroid, 2019). The utility of eye-tracking methods in investigating SLA rests on the assumption of the Eye-Mind Hypothesis (Just and Carpenter, 1980) stating that "eye movements are over orienting responses that signal the alignment of attention with the object at the point of gaze" (Godfroid, 2019, p. 23). Visual attention can give us insight into how readers allocate cognitive resources to text. Although eye-tracking in reading is often restricted to processes related to local word parsing, there has also been attention paid to how Eye-tracking data can inform us about high-order reading cognition. For example Yeari et al. (2017) utilized eye-tracking methods to find that readers pay more or less attention to peripheral information depending on their reading purpose. Dirix et al. (2020) found that having readers

engage with a text for informational purposes elicited shorter overall reading times and shorter fixations than when readers engaged with a text for studying purposes, and that these differences were increased for L2 text-reading. They additionally found that students could compensate for slower processing with more overall attention to the text. Huang et al. (2022) examining Chinese L2 English learners' reading of texts with unfamiliar words. They found that working memory and duration of first fixation affected how readers processed unfamiliar words. Comprehension performance was affected by the longer duration of first fixation on unfamiliar words, yet unfamiliar word fixation affected comprehension less for learners who demonstrated higher working memory capability. This result demonstrates that successful reading can involve compensation for one weakness in reading with another resource.

Less attention has been paid to real-time reading behavior during L2 reading comprehension assessment. Bax and Chan (2019) measured second language English readers' eye-movements during reading test completion, finding that more successful readers made shorter fixations on average and paid more attention to areas of text based on relevance. In studies by Prichard and Atkins (2016, 2019), L2 English readers were found to underutilize strategic reading when they had time pressure to complete a reading task. Readers who were able to consciously apply strategic reading to their task did better in their comprehension. Outside of these studies, little research has been conducted on L2 reading assessment, especially with the analysis of interactions between components of reading ability in mind, but it is clear that eye-tracking can provide an avenue to understanding reading behaviors in relation to comprehension ability for L2 learners (Conklin et al., 2018).

Research questions

The goal of this study was to investigate whether differences in real-time reading behavior, as measured using eye-tracking, uniquely impacts second language reading comprehension performance, and to investigate interactions between reading behavior and other individual differences. Specifically,

(1) To what extent do online reading behaviors predict variance in reading comprehension scores beyond that predicted by offline measures of individual cognitive and noncognitive differences (logic, memory, motivation, proficiency)?
(2) To what extent do linear models reveal compensatory effects within individual differences impacting comprehension outcomes?

Methods

The data for this study involved second language English readers completing three sequential reading comprehension tasks each while reading one of a pool of six texts. During reading task completion, an eye-tracker recorded reader behavior. Each of the aspects of data collection and analysis are described below.

Participants

The data for this study was collected from 102 international students (graduate and undergraduate, with ages ranging between 19 and 52) at a large university in the southeastern United States as part of a larger study on second language reading assessment. The students represented a wide range of language backgrounds, including Mandarin, Spanish, Korean, Telugu, Cantonese, Urdu, Vietnamese, and 21 other language groups. Participants had spent an average of 4.67 years in an English-speaking environment, with an average of 5.1 years of English classroom experience.

Texts

The reading procedure involved reading three texts from a pool of six texts. The six texts were all passages from high school science textbooks on the following topics: "biotechnology and DNA," "the compound microscope," "chemical properties of water," "the science of hunger," "the psychology of making choices," and "attitudes and roles." Texts ranged from 315 to 350 words, and each consisted of four paragraphs. The texts were selected based on their similarity in terms of lexical and syntactic complexity, as well as their intended reading level of US high school grade 10 (Flesch Kincaid reading level is reported in Supplementary Appendix A). Although there is an inherent advantage in comprehension for any examinees with background knowledge on each particular topic, the texts were selected from introductory writings on the topic and reviewed by a panel of three applied linguists for broad approachability.

Tasks

Three reading comprehension tasks were completed by participants during the eye-tracking procedure. Each task reflected an oft-used second language reading test format along the spectrum of selected-response to constructed-response. The tasks were a multiple-choice (MC) reading task (selected-response, discrete-point scoring), a cloze task (constructed response, discrete-point scoring), and a summary task (constructed response, human scored). The MC task for each text involved answering five questions: one main-idea question, two detail questions, and two inferencing questions. Each question had three answer choices. Questions were presented to the right of the text and participants could see the text and questions at the same time without scrolling or leaving the screen.

The cloze task involved reading the text, but with 15 words replaced by blanks. There was no word bank to fill in the blanks, and participants needed to use comprehension processes to reconstruct the text. Words were blanked using a rational deletion method (Kleijn, 2018) targeting a content word or coherence-maintaining word every 15 words rather than a random or systematic deletion method to ensure that the task focused on comprehension processes as much as possible. Cloze tasks were scored by human raters so that near synonyms could be accepted

as correct answers. Scoring was otherwise objectively rated based on an answer key.

The summary task asked readers to produce a 100-word summary, or a "brief account" (Seidlhofer, 1990), of the text for a hypothetical fellow student who did not read the text. The provision of a specific audience and task encouraged summarizers to focus on content transmission and not linguistic copying and recall. As with the MC task, the task pane in which examinees typed their summaries was presented to the right of the text so the examinees could navigate between text and task without scrolling or changing screens. Summaries were scored by human raters for level of detail, evidence of mental modeling, and adherence to the task. Each text is presented in Supplementary Appendix A.

Eye-tracking metrics

Readers' real-time reading behavior was recorded with an ASL EyeTrac 6 device. Participants were seated two feet from a computer screen as they completed the reading comprehension tasks, keeping their head in a stable position using a chin rest. Each participant was calibrated with a practice exercise to ensure accuracy of fixations to within 0.2 inches before recording began. Fixation location and duration data were gathered by the eye-tracking device, along with length of *saccades* (jumps between fixations). Fixations were considered to be any pause in eye-movement >100 ms (Manor and Gordon, 2003). Lines of text and paragraphs were designated using *post-hoc* areas of interest (AOIs). Further AOIs were marked for each task area.

Various metrics were derived from the raw data which are relevant for understanding text-level reading behavior. The derived metrics are "late" processing measures, which reflect integrating of larger portions of text. These contrast with "early" processing measures, primarily focused on individual words and phrases. The metrics calculated in this study are average saccade length, total numbers of text fixations per word in reading text area and task areas, average fixation durations on the text and in task areas, and average fixations per word per dwell in AOIs. Unique for the assessment context, the number of transitions between a fixation on text and a fixation on a task area was calculated. Metrics related to rereading were also gathered, but they were largely multicollinear with total fixations per word, indicating that text level reading in an assessment setting naturally involves a great deal of rereading. Eye-tracking metrics were further evaluated for normality and text topic effects. These analyses are not reported in detail and were merely performed to ensure the assumptions were met for subsequent analyses. The metrics utilized in analyses are presented in Table 1.

Although there was a time limit for the overall data collection procedure of 90 min, there was substantial variance in the amount of time taken to complete the individual reading tasks, so for each task, the eye-tracking metrics were checked for multicollinearity with reading time. The following metrics were found to be multicollinear ($r \geq 0.7$) with reading time and were excluded from further analysis: transitions in the cloze task ($r = 0.739$), text fixations per word in the cloze task ($r = 0.896$), task fixations per word in the cloze task ($r = 0.729$), text fixations per word in the MC task ($r = 0.762$), and task area fixations per word in the MC task ($r = 0.744$). No fixation metrics were multicollinear with reading time for the summary task.

Individual differences

Considering the large number of cognitive factors which impact comprehension aside from eye-movement behavior, data from individual differences were gathered to understand what moderating effects might occur on how attention impacts task performance in reading assessment.

Language proficiency

Academic reading ability in a second language depends heavily on general grammatical knowledge and vocabulary size. Due to the diverse background of the participants, no standardized measure of proficiency could be gathered *a priori* for all participants, so an 18-item gap-fill c-test was developed to target morpho-syntax and academic vocabulary. The test involved 18 sentences with a word which was left half blank. The test is based on the productive orthographic vocabulary size tests (Laufer and Nation, 1999) which have been found to strongly predict reading comprehension in a second language (Cheng and Matthews, 2018).

Reading speed

Reading fluency is an important lower-order literacy skill (Grabe, 2009; Gauvin and Hulstijn, 2010; Stoller et al., 2013), which has been found to be connected to reading behavior in monolingual data (Taylor and Perfetti, 2016). Reading fluency was here operationalized as reading speed in words per minute during a silent reading of a 12th grade-level academic text with 375 words about geology. The participants were asked comprehension questions afterward to ensure the participants read intentionally but the questions were not scored.

Reading motivation

Motivation is an important factor in understanding academic reading comprehension (Wigfield and Guthrie, 1997; Schaffner and Schiefele, 2013). A survey was developed to measure reading motivation and was administered before the reading trials. All items were discrete-point, using a 5-point Likert scale, and included 10 items. Five items measured intrinsic motivation to read, and five items measured extrinsic motivation to read. Intrinsic motivations include personal reasons such as enjoyment or personal enrichment, and extrinsic motivations include practical reasons such as career-usefulness of reading texts or social engagement through reading. These items were derived from previous surveys of motivation (Wigfield and Guthrie, 1997; Ryan and Deci, 2000). A confirmatory factor analysis was used to investigate the two-factor nature of the survey, resulting in a significant model ($\chi^2 = 58.23$, $p = 0.006$). However, only the intrinsic motivation questions reliably factored together in a unified construct, and so the intrinsic motivation metric was featured subsequent modeling of comprehension. The entire motivation survey is presented in Supplementary Appendix C.

TABLE 1 Description and operationalization of eye-tracking measures.

Measure	Purpose for measurement	Target area	Operationalization notes
Fixations on text per word	Global, careful reading	Entire text area	Average of all fixations made on the reading text in a given trial
Mean length of saccade	Global reading	Entire trial area	Average absolute distance between sequential fixation coordinates throughout a trial
Mean fixations per line dwell	Linear, local reading	Line areas of interest	Average count of fixations per dwell across dwells in line AOIs. Controlled for number of words in AOI
Mean fixations per paragraph dwell	Local, careful reading	Paragraph areas of interest	Average count of fixations per dwell across dwells in paragraph AOIs. Controlled for number of words in AOI
Mean duration of fixations on text	Careful reading	Entire text area	Average time (ms) of fixations in any text area of interest. Controlled for size of AOI
Mean duration of fixations on task	Careful reading, Task integration	Task areas of interest	Average time (ms) of fixations in any task area of interest. Controlled for size of AOI
Fixations on task per word	Task integration	Task areas of interest	Average of all fixations made on the task areas in a given trial. Size of the areas in the respective tasks is controlled for
Number of gaze transitions between text and task	Task integration, global reading	Text and task areas of interest	Raw count of saccades which moved from a text area of interest to a task area of interest

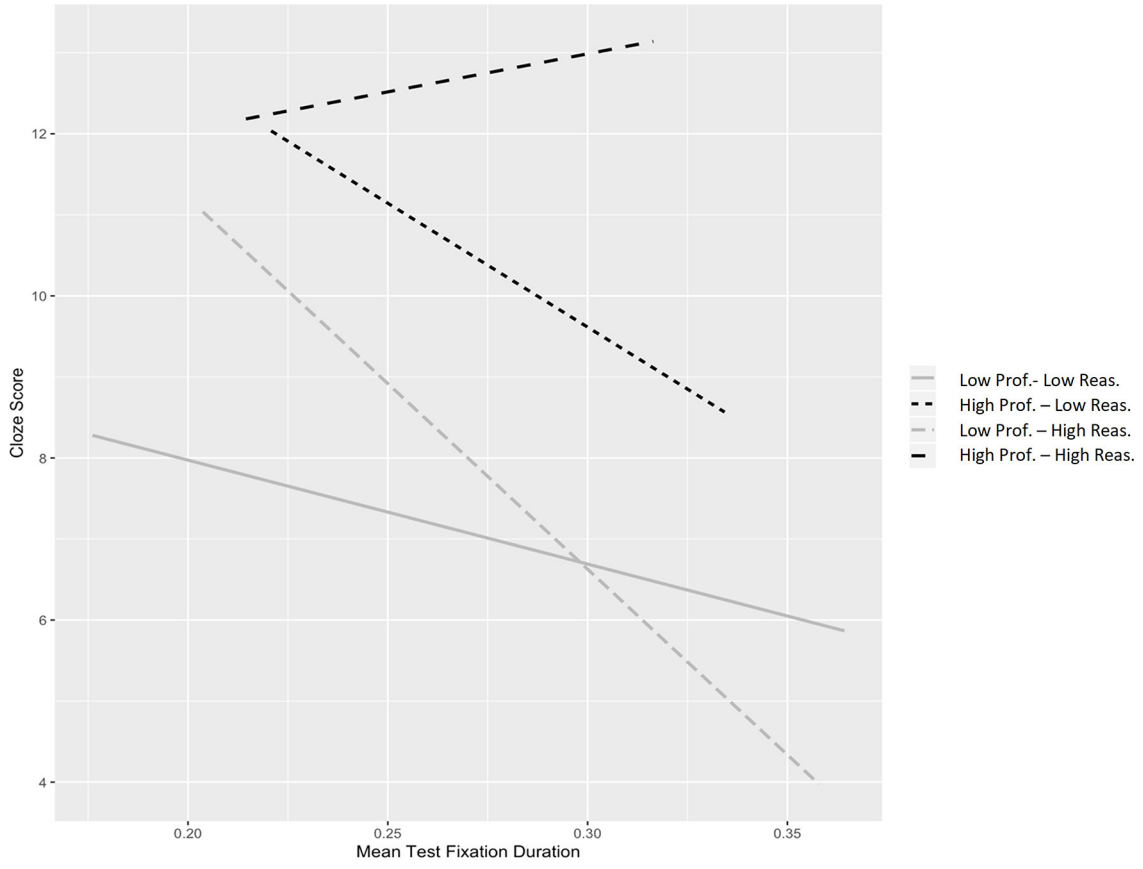


FIGURE 1
Cloze scores plotted against mean fixation duration on text, with groups for L2 proficiency and reasoning. Prof., proficiency; Reas., reasoning.

Reasoning

Logical reasoning, or inductive reasoning, has been predictive of reading comprehension ability in previous research (Klauer and Phye, 2008). This facet of reasoning specifically refers to the ability to extrapolate information from patterns. For this study, inductive reasoning was measured using a 10-item incomplete series test where participants saw a pattern of three shapes and selected the best of four options to complete the sequence.

TABLE 2 Linear regression model to predict cloze task scores.

Predictor	*B*	SE	*t*	*p*-value	r^2	Δr^2
Intercept	−0.021	0.072	−0.293	0.770		
L2 proficiency × reasoning × mean text fixation duration	0.151	0.070	2.150	0.034*	0.021	
L2 proficiency	0.663	0.074	8.959	<0.001*	0.442	0.421
Reasoning	0.278	0.078	3.551	0.001*	0.511	0.069
Mean text fixation duration	−0.222	0.072	−3.099	0.003*	0.559	0.048

B, standardized coefficients.
*Significant at $p < 0.05$.

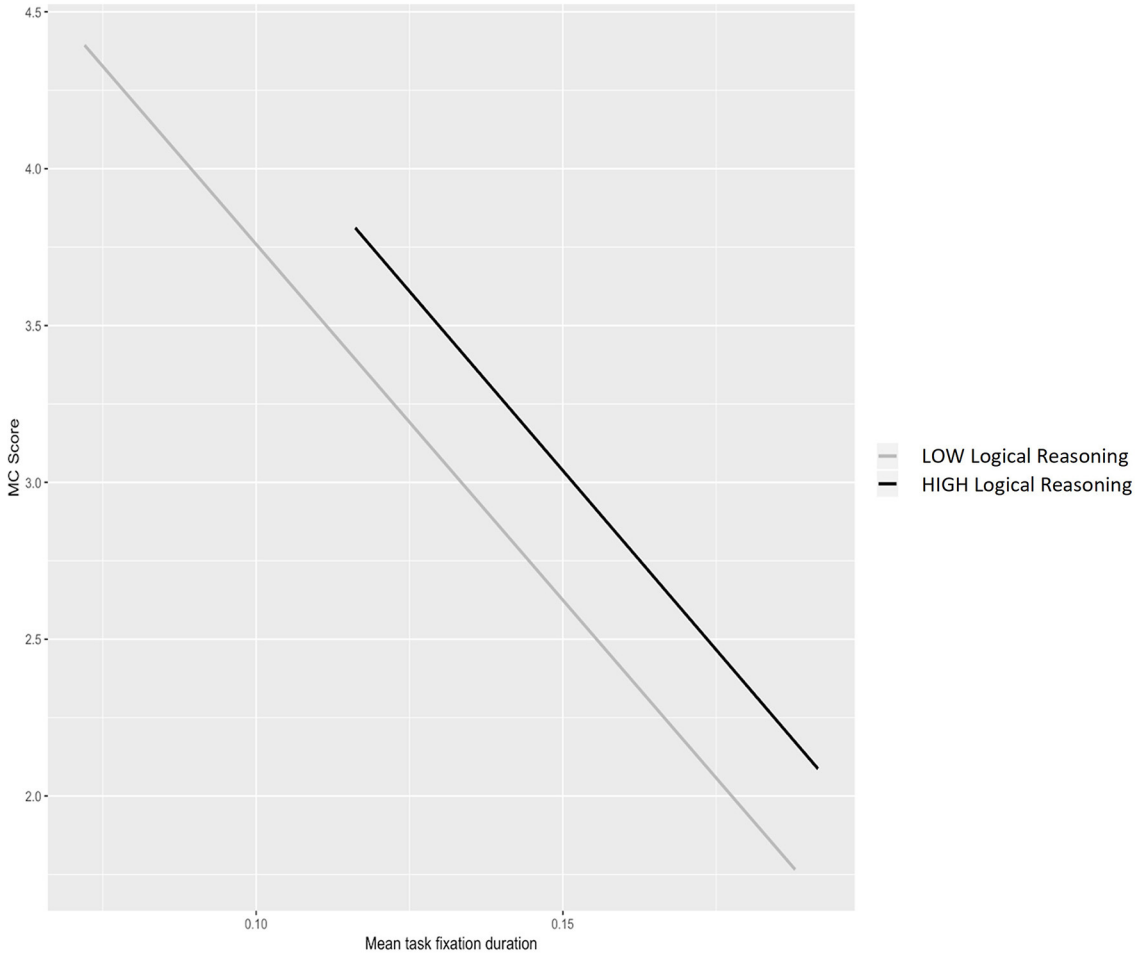


FIGURE 2
MC score plotted against mean task area fixation duration, with groupings for above-median and below-median reasoning.

Working memory

Working memory has been found to contribute to reading comprehension in monolingual readers (Cain et al., 2001; Calvo, 2005; Carretti et al., 2009) and multilingual readers (Alptekin and Erçetin, 2010; Lipka and Siegel, 2012; Erçetin and Alptekin, 2013; Joh and Plakans, 2017). Working memory was measured using a 2-back test, where participants were shown a series of simple images. Participants compared the image on screen to the image which they saw two images previously, deciding if they were the same within 1 s. They saw a total of 35 images, among which 15 2-back matches were randomly distributed in the sequence of pictures. Participants were scored by the percent of correct responses.

Scoring

Each participant's responses were scored in a task-appropriate manner. MC task responses were scored automatically by key, and a score of 0 to 5 was assigned to each test-taker. Trained raters scored the cloze tests with an answer key using an acceptable

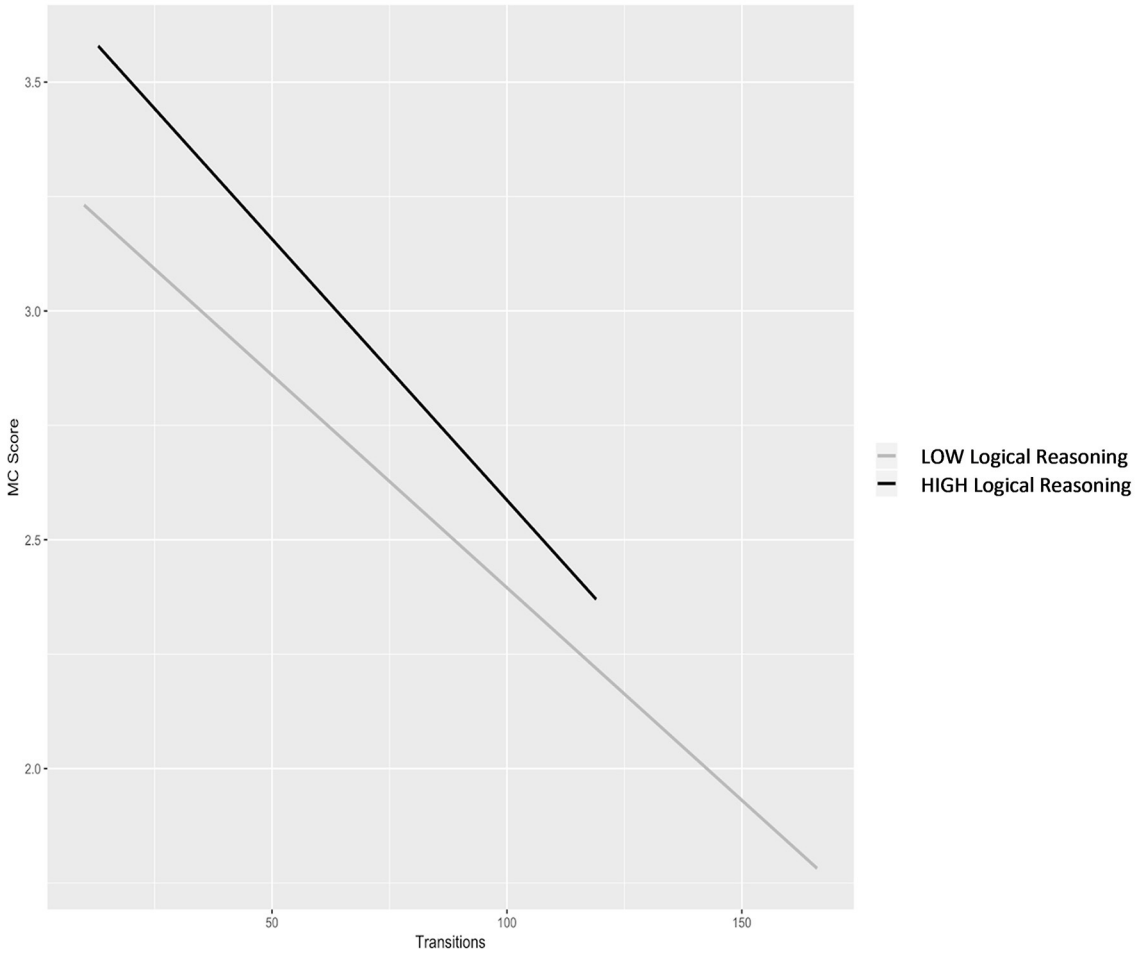


FIGURE 3
MC score plotted against number of transitions, with groupings for above-median and below-median reasoning.

TABLE 3 Linear regression model to predict MC task scores.

Predictor	*B*	SE	*t*	*p*-value	r^2	Δr^2
Intercept	<0.001	0.092	0.000	1.00		
Mean fixation duration (task)	−0.347	0.092	−3.766	<0.001*	0.135	
Transitions	−0.280	0.092	−3.036	0.003*	0.213	0.078

B, standardized coefficients.
*Significant at $p < 0.05$.

response scoring method. Each cloze blank had an intended response based on the source text, but scorers also accepted near-synonyms. Each correct response to a blank in the passage was given a point, for a score range of 0 to 15 points. Trained raters also scored the summary tasks. The raters consisted of a pool of seven applied linguists. Summaries were rated using an analytic rubric developed by the researcher (see Supplementary Appendix B for the full summary rating guidelines). This rubric was developed based on constructs in Taylor (2013) used for rating summaries. The constructs include content accuracy, level of modeling (distinguishing between main ideas and subordinate details), task completion, and language quality. Only accuracy, modeling, and task completion were considered as part of the comprehension score, with the language score being used to control for productive language ability and ensure raters did not factor linguistic aspects into their content scores. The language score component was only included on the rubric to mitigate the effect of raters' judgments of productive language quality in their assessment of reading comprehension.

Each summary was given a score out of 4 for each construct, and each summary was rated by at least two raters. If ratings from the two raters were misaligned in any category by more than one point, a third rater was called. Only 8.5% of ratings resulted in a third rater's adjudication, and no fourth ratings were necessary. The summary ratings were analyzed for reliability using Multi-faceted Rasch Analysis (Linacre and Wright, 2002; Linacre, 2023), Although the complete results of such an analytic measure are too voluminous to report here, importantly, the rubric constructs demonstrated independence with high separation reliability of 0.9, and the raters each exhibited acceptable fit, ranging from 0.72 to 1.12. This is within the acceptable range of model fit of 0.5 to 1.5 (Linacre, 2023), indicating good internal consistency among the raters.

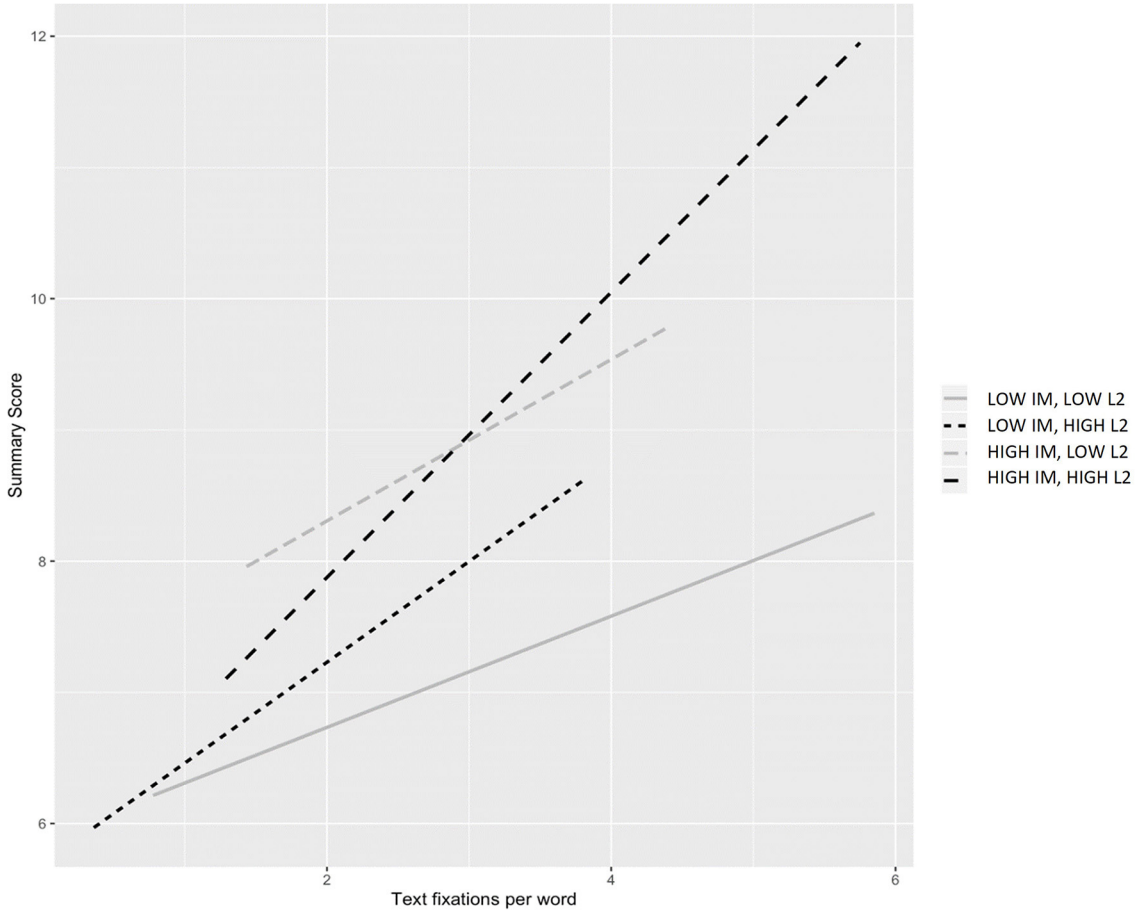


FIGURE 4
Summary score plotted against text fixations per word, with groupings for above-median and below-median motivation and L2 proficiency. IM, intrinsic motivation; prof., L2 proficiency.

The average of the closest two was used as the final score for each construct, and an additional Total Comprehension score was calculated as the sum of the accuracy, modeling, and task completion ratings. This total score was the score used as the dependent variable in summary modeling analyses.

Analyses

Three linear models were constructed to predict comprehension score in each task, using predictors of eye-tracking metrics along with individual differences which exhibited meaningful correlation with scores. A separate linear model was developed for each reading task. Correlations were calculated between each pair of metrics and with task scores. Eye-tracking and individual differences metrics which had significant and at least a weak correlation with score, were included in a linear regression model to predict score.

Results

This section will cover the results of the analyses described in the previous section on eye-movement and reading comprehension. Comprehension scores for each of the different reading tasks were predicted with unique models, the construction of which began with examination of correlations. Based on correlations, eye-tracking metrics with at least a weak significant correlation with scores were selected for linear regression modeling. Similarly, individual difference metrics at least weakly significantly correlated with score were included as well. Text topic was included as a control variable.

Predicting cloze scores

One eye-tracking metric was found to correlate with cloze scores: mean fixation duration on text ($r = -0.306$). The correlation was negative, implying faster eye-movement via lower fixation durations was related to higher performance. Two individual differences were found to significantly correlate with cloze scores: L2 proficiency ($r = 0.630$) and logical reasoning ($r = 0.212$). The metrics were not correlated with each other or with average fixation duration on text.

Before constructing the linear model, visual inspection of the three variables was conducted to ascertain the presence of interactions. Figure 1 shows cloze scores along the y-axis, with

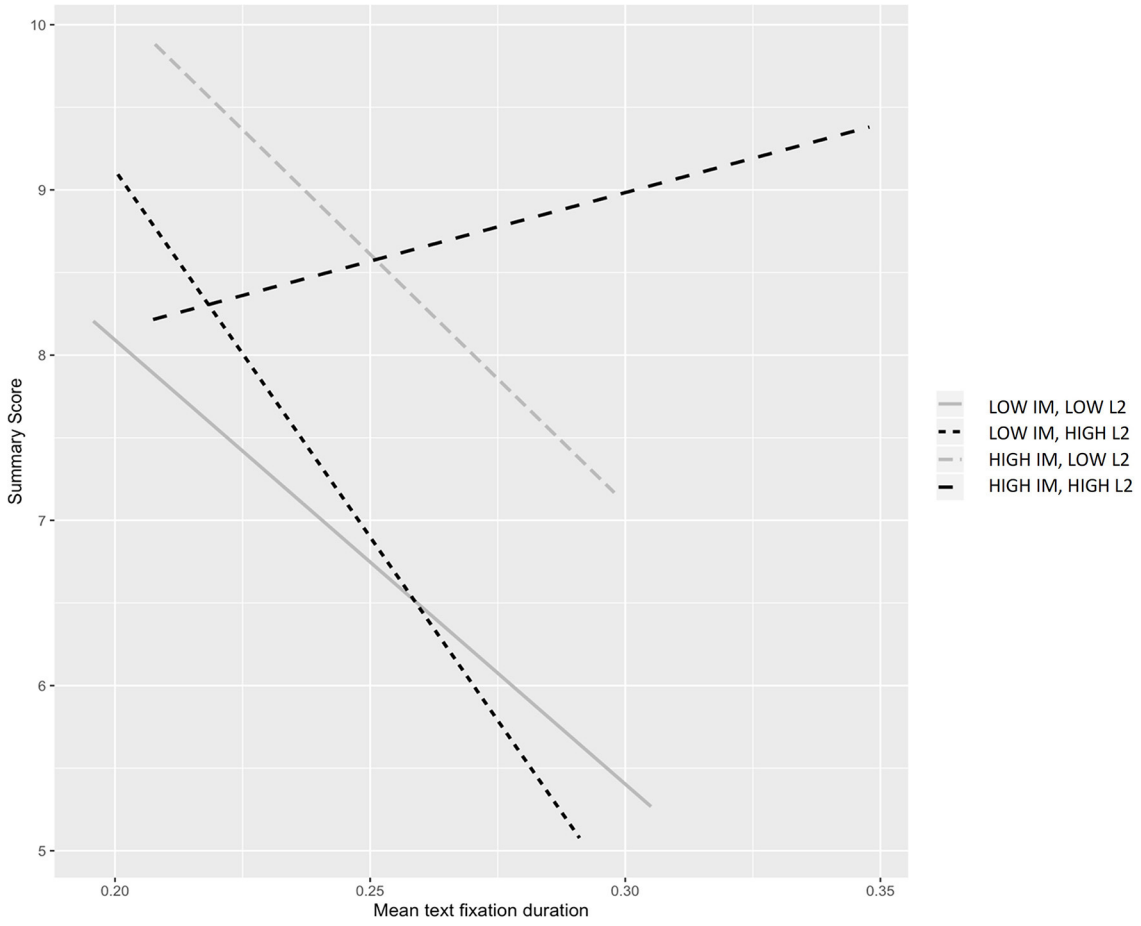


FIGURE 5
Summary score plotted against mean text fixation duration, with groupings for above-median and below-median motivation and L2 proficiency. IM, intrinsic motivation; prof., L2 proficiency.

mean fixation duration on text along the x-axis, and groupings for L2 proficiency level and reasoning level (each split into two groups around the median). The different slopes of the mean text fixation duration fit lines between proficiency levels and reasoning levels indicate a possible interaction effect. As such, these interactions were included in the linear modeling.

Variables were standardized for the linear model, and a linear model with three variables as well as on three-way interaction was constructed. The model was found to be significant, $F_{(4,94)} = 27.64$ ($p < 0.001$), and a description of the model is presented in Table 2. The model was found to have a large effect size, explaining 55.9% of variance in scores. Average fixation duration on text, as well as interactions with individual differences, was found to be uniquely account for variance in the model, though the effect size is very small. L2 proficiency and reasoning were positive predictors of score, and average fixation duration was a negative predictor, implying that shorter fixations related to higher scores. The interaction variable is more complex, but when interpreted alongside visual presentation of data in Figure 1, it can be seen that when both L2 proficiency and reasoning are above average, the negative impact of fixation duration reverses somewhat, i.e., readers ability to make fast fixations is less important when reasoning and L2 proficiency are high. This effect is small, but still indicates that these metrics may have a compensatory effect between them.

Predicting MC scores

Two eye-tracking metrics were found to correlate with MC scores: transitions between text and task areas ($r = -0.293$), and mean fixation duration on the question area ($r = -0.379$). Each of the correlations were negative, implying fewer transitions and shorter fixations on the question area were related to higher MC performance. Only a single individual difference metric was found to significantly correlate with MC scores, logical reasoning ($r = 0.221$). Reasoning was not significantly correlated with any eye-tracking metrics.

Before constructing the linear model, visual inspection of the two variables was conducted to ascertain the presence of interactions. Figure 2 through Figure 3 show MC scores along the y-axis, with eye-tracking metrics along the x-axis, and groupings for reasoning (split into two groups around the median). The participants were split into groups for above median or below median in reasoning to make the plots reader friendly, and this grouping is not used in further analysis. The similar slopes of the average fixation duration and transitions fit lines between reasoning levels indicates that higher reasoning scores trend with higher comprehension scores, and there is likely little to no interaction effect between the reasoning and eye-movement behavior.

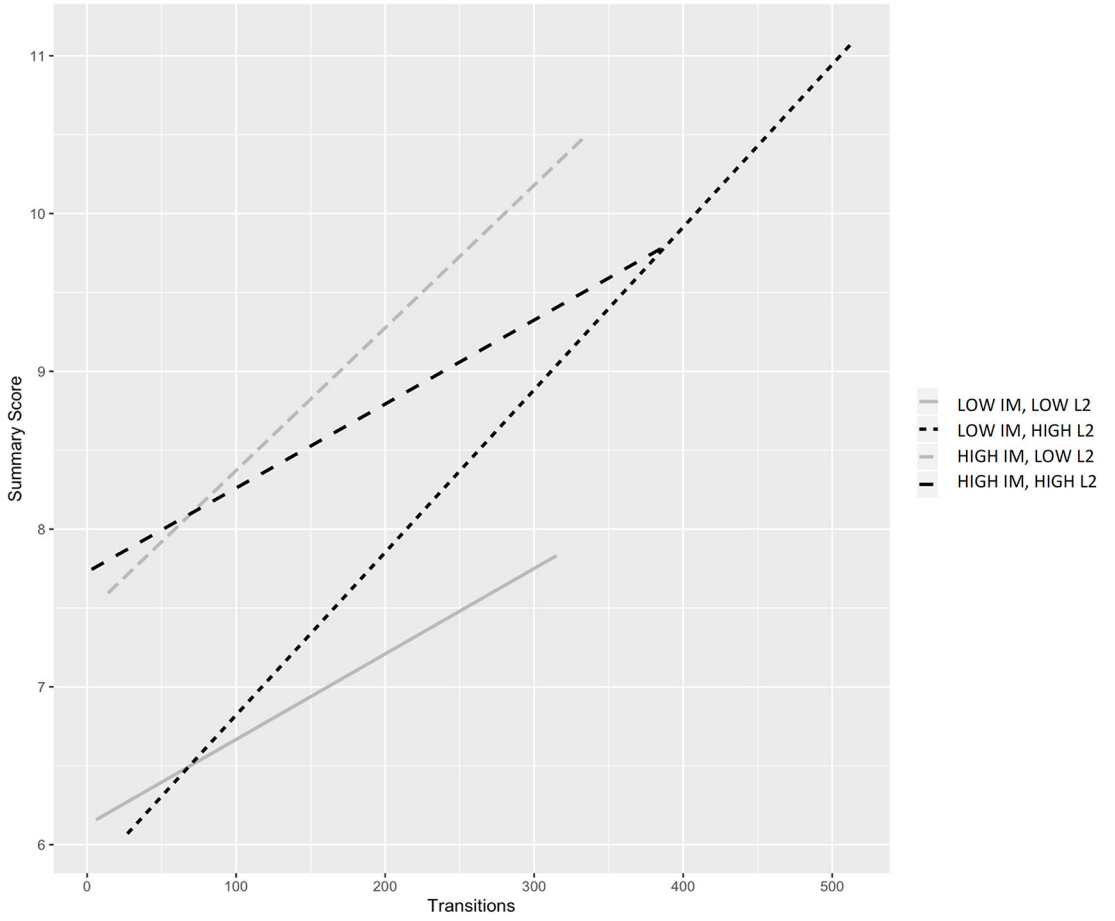


FIGURE 6
Summary score plotted against number of transitions, with groupings for above-median and below-median motivation and L2 proficiency. IM, intrinsic motivation; prof., L2 proficiency.

The linear regression model for MC score included as predictors reasoning, average duration of fixation on questions, and transitions. Of the included predictors, mean fixation duration on questions and transitions were significant predictors, but reasoning and any interaction variables were not. These effects were thus removed from the model. The final 2-predictor model was found to be significant, $F_{(2,96)} = 12.583$ ($p < 0.001$) and Table 3 contains a description of the model. The model had a moderate effect size in predicting score, with $r^2 = 0.213$. Mean fixation duration on questions was the most significant predictor, showing that making shorter fixations on the question area contributed to higher scores. Transitions was also a significant predictor, with fewer transitions being predictive of higher score.

Predicting summary scores

Three eye-tracking metrics were found to correlate with summary scores: transitions between text and task areas, this time positively correlated ($r = 0.302$), fixations per word in text area ($r = 0.364$), and mean fixation duration on the text ($r = -0.214$). As in the cloze data, mean duration of fixations on the text was negatively correlated with summary score. Fixations per word on text was significantly correlated with transitions ($r = 0.477$), but there were no multicollinear variables. Two individual difference metrics were found to significantly correlate with summary scores: L2 Proficiency ($r = 0.297$) and Intrinsic Motivation ($r = 0.345$).

Before constructing the linear model, visual inspection of the three variables was conducted to ascertain the presence of interactions. Figures 4–6 show Summary scores along the y-axis, with eye-tracking metrics along the x-axis, and groupings for individual differences split around the median. The participants were split into groups for above median or below median in reasoning to make the plots reader friendly, and this grouping is not used in further analysis. Each graph reveals interaction effects between the eye-tracking metrics and the individual differences, but the most can be found in the graph for mean fixation duration. Here, mean fixation duration normally has a negative correlation with summary score, yet at higher levels of both motivation and L2 proficiency, the relationship between fixation duration and summary score is positive. These interactions are further explored for significance in the linear regression model.

The linear regression model for Summary score included as predictors intrinsic motivation, L2 proficiency, fixations per word

TABLE 4 Linear regression model to predict summary task scores.

Predictor	*B*	SE	*t*	*p*-value	r^2	Δr^2
Intercept	0.004	0.083	0.043	0.966		
Intrinsic Mot. × L2 proficiency × mean fix. duration	0.253	0.087	2.902	0.005*	0.058	
L2Proficiency × mean fix. duration	0.211	0.078	2.707	0.008*	0.063	0.005
Intrinsic motivation	0.188	0.086	2.188	0.031*	0.149	0.086
L2 proficiency	0.177	0.087	2.044	0.044*	0.207	0.058
Fixation per word (Text)	0.385	0.086	4.498	<0.001*	0.331	0.124
Mean fixation duration (text)	−0.275	0.089	−3.098	0.003*	0.397	0.066

B, standardized coefficients.
*Significant at $p < 0.05$.

on text, mean fixation duration on text, and number of transitions. Number of transitions and the interactions with it were not found to be significant to the model and were removed. The final model was found to be significant, $F_{(6,92)} = 9.641$ ($p < 0.001$). Table 4 contains a description of the model. The effect size of the model was large, with about 39.7% of the variance explained for summary scores ($r^2 = 0.397$). The three-way interaction with L2 proficiency, motivation, and mean fixation duration was found to be significant and a positive predictor of summary scores, where mean fixation duration alone was a significant negative predictor. The stronger of the two predictors was mean fixation duration alone, indicating that the positive interaction does not mean readers with stronger proficiency and motivation necessarily benefit from longer fixations, but rather mitigate slower fixations with their other abilities. A positive pairwise interaction between L2 proficiency was also significant in the model, but not to the extent of the three-way interaction. This still further shows the strength of L2 proficiency to compensate for more rapid fixations.

In addition to mean fixation duration, three other main effects were found to be significant. Fixations per word on text was the most meaningful predictor, indicating higher numbers of fixations predicted higher summary scores with a moderate effect size. High motivation was a moderate positive predictor as well, and L2 proficiency had a main effect, but it was not as impactful on score as its interaction effects with text duration.

Discussion

The online reading behavior measured in this study was used to understand its impact on reading comprehension and interactions with individual differences across various reading assessment tasks. Each reading task elicited a different linear model to predict comprehension scores using individual differences and eye-tracking metrics. These are briefly summarized below.

Score on the cloze was related to L2 proficiency, reasoning, and efficiency of fixations. Shorter fixations on text areas was predictive of cloze score, with a small but meaningful effect size ($\Delta r^2 = 0.048$), though this was not as meaningful as the predictive effects of L2 Proficiency (to a large extent) and reasoning. The three way interaction between these variables indicated that at higher levels of proficiency and reasoning, the effect of fixation efficiency diminished as other skills could compensate.

The model predicting score on the MC task was much weaker, with two eye-movement measures related to processing the question area of the text being meaningful in the model. Having shorter fixation durations on the questions and fewer transitions between question and text predicted higher comprehension scores. Though there was a possible interaction between reasoning and number of fixations, with higher reasoning scores relating to lower fixations, neither these main effects nor this interaction was significant in the score model.

The model predicting summary task scores included multiple predictors, with motivation, proficiency, and fixations positively predicting summary scores with at least a weak effect size, and mean fixation duration negatively predicted scores. There was again an interaction, with longer fixation durations no longer having a negative impact on score at higher levels of proficiency and/or motivation. Readers with higher motivation appear to be able to compensate for the impact of slower processing ability on comprehension with more L2 linguistic resources.

To answer the first research question, *to what extent do online reading behaviors predict variance in reading comprehension scores beyond that predicted by other individual differences*, we can look at the appearance of eye-tracking main effects in the models of comprehension for each reading task. For each reading task model, a fixation duration metric was found to predict scores, with shorter average fixations predicting higher score. This is in line with previous research which showed that skilled readers make short, efficient fixations (Ashby et al., 2005; Bax, 2013; Krieber et al., 2016). The summary task was distinct from the cloze and MC tasks in that an eye-tracking metric positively predicted scores. For the summary task, a greater number of fixations on the reading text was predictive of higher summary scores with a medium effect size. It is possible that the summary task pushes readers to build a more detailed mental model of the text and is more cognitively demanding, so more fixations are necessary. This is attested in Bax (2013) who found eye-movement behavior related to higher-order processing in summary comprehension tasks.

In relation to the second research question, *to what extent do linear models reveal compensatory effects within individual differences impacting comprehension outcomes,* interactions were present in two models of reading comprehension. The results from this study align with previous research which asserts that

readers can compensate for certain weaknesses in reading ability by utilizing other related skills (Stanovich, 1980; McNeil, 2012). McNeil's (2012) framework made predictions about how readers at different levels would rely on strategic, literate, or linguistic resources. Although the current study did not seek to ascertain which aspect of skills would impact comprehension most at different levels of reading, we nonetheless established that L2 language ability, reading behavior, and strategic abilities have unique contribution to reading comprehension, and readers can compensate for weaknesses in one skill with strengths in another. The specific compensations related to reading efficiency, where efficiency was less critical for comprehension when readers had higher L2 proficiency and/or another skill (logical reasoning for the cloze task and motivation for the summary task). This deviates slightly from previous research which found interaction effects with eye-tracking metrics on reading comprehension. In Huang et al. (2022), working memory was found to be a significant predictor of comprehension, and was able to compensate for the effect of unfamiliar words which caused slower processing. However, the Huang et al. (2022) study was looking at smaller texts with shorter reading times, so the results of the current study extend our understanding of how measures of efficient processing materialize at different lengths of text. For longer texts and tasks allowing simultaneous access to text and task, working memory may not be the most predictive cognitive measure, and may not compensate for late-measure eye-tracking metrics as measured in this study.

Conclusion

This study has taken a novel look at how reading behavior, measured through eye-tracking, differed across reading tasks in terms of impact on task performance. Beyond furthering our understanding of the second-language reading process, there are implications for language teaching and testing as well. It is worth acknowledging as teachers that readers benefit from learning various aspects to reading, from refining language proficiency to practicing extensive reading for speed to engaging in reasoning and motivation-enhancing tasks. Since there is variance in how different abilities contribute to comprehension performance across tasks, it is also worth teaching developing readers goal-setting strategies to help them compensate for the demands set by their reading purpose. For example, reading for discrete information as in the cloze and MC tasks demands quick, efficient reading, but reading for global comprehension as in the summary task required more comprehensive attention to the text. Being able to moderate one's approach to reading in different tasks is critical.

These findings must be taken in light of the study's limitations. Previous research (Cook and Wei, 2019) has advised against drawing direct connections between eye-tracking metrics and underlying processes. This is especially true for the current study which utilized very coarse-graining eye-tracking metrics. Fixations per word and average fixation duration are both general measurements based on participants' entire trial of reading data. More attention to areas of interest and phrasal/word-level eye-tracking information could provide more to the picture of eye-movement behavior's contribution to comprehension. Further research is needed to better understand how finer shades of measurement of fixation duration impacts comprehension and relates to other individual differences. It is also necessary to state that while we observed the impact of reading efficiency in this study, we were not able to ascertain whether readers consciously engaged in faster or slower reading as part of an active reading strategy. More research is needed to connect eye-movement behaviors to conscious engagement in specific types of reading strategy activation.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Georgia State University Institutional Review Board [University Research Services & Administration (URSA)]. The patients/participants provided their written informed consent to participate in this study.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

Data collection for this research would not have been possible without support from the Georgia State University Adult Literacy Research Center Dissertation Support Grant.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2023.1176986/full#supplementary-material

References

Alderson, J. C. (2000). *Assessing Reading (Atlanta Library North 4 LB1050.46.A43 2000)*. Cambridge : Cambridge University Press. doi: 10.1017/CBO9780511732935

Alptekin, C., and Erçetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *J. Res. Read.* 33, 206–219. doi: 10.1111/j.1467-9817.2009.01412.x

Ashby, J., Rayner, K., and Clifton, C. (2005). Eye movements of highly skilled and average readers: differential effects of frequency and predictability. *Q. J. Exp. Psychol.* 58, 1065–1086. doi: 10.1080/02724980443000476

Bachman, L. F., and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: OUP Oxford.

Bax, S. (2013). The cognitive processing of candidates during reading tests: evidence from eye-tracking. *Lang. Test.* 30, 441–465. doi: 10.1177/0265532212473244

Bax, S., and Chan, S. (2019). Using eye-tracking research to investigate language test validity and design. *System* 83, 64–78. doi: 10.1016/j.system.2019.01.007

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annu. Rev. Appl. Linguist.* 25, 133–150. doi: 10.1017/S0267190505000073

Borsboom, D. (2005). Measuring the Mind: *Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511490026

Cain, K., Oakhill, J. V., Barnes, M. A., and Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Mem. Cognit.* 29, 850–859. doi: 10.3758/BF03196414

Calvo, M. G. (2005). Relative contribution of vocabulary knowledge and working memory span to elaborative inferences in reading. *Learn. Individ. Differ.* 15, 53–65. doi: 10.1016/j.lindif.2004.07.002

Carretti, B., Borella, E., Cornoldi, C., and De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: a meta-analysis. *Learn. Individ. Differ.* 19, 246–251. doi: 10.1016/j.lindif.2008.10.002

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Sci. Stud. Read.* 1, 3. doi: 10.1207/s1532799xssr0101_2

Cheng, J., and Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Lang. Test.* 35, 3–25. doi: 10.1177/0265532216676851

Conklin, K., Pellicer-Sanchez, A., and Carroll, G. (2018). *Eye-tracking: A Guide for Applied Linguistics Research*. Cambridge: Cambridge University Press. doi: 10.1017/9781108233279

Cook, A. E., and Wei, W. (2019). What can eye movements tell us about higher level comprehension? *Vision* 3, 45. doi: 10.3390/vision3030045

Dirix, N., Vander Beken, H., De Bruyne, E., Brysbaert, M., and Duyck, W. (2020). Reading text when studying in a second language: an eye-tracking study. *Read. Res. Q.* 55, 371–397. doi: 10.1002/rrq.277

Erçetin, G., and Alptekin, C. (2013). The explicit/implicit knowledge distinction and working memory: implications for second-language reading comprehension. *Appl. Psycholinguist.* 34, 727–753. doi: 10.1017/S0142716411000932

Gauvin, H. S., and Hulstijn, J. H. (2010). Exploring a new technique for comparing bilinguals' L1 and L2 reading speed. *Read. Foreign Lang.* 22, 84–103.

Godfroid, A. (2019). *Eye Tracking in Second Language Acquisition and Bilingualism: A Research Synthesis and Methodological Guide*. London: Routledge. doi: 10.4324/9781315775616

Grabe, W. (2009). Reading in a Second *Language: Moving from Theory to Practice*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139150484

Huang, L., Ouyang, J., and Jiang, J. (2022). The relationship of word processing with L2 reading comprehension and working memory: insights from eye-tracking. *Learn. Individ. Differ.* 95, 102143. doi: 10.1016/j.lindif.2022.102143

Joh, J., and Plakans, L. (2017). Working memory in L2 reading comprehension: the influence of prior knowledge. *System* 70, 107–120. doi: 10.1016/j.system.2017.07.007

Just, M. A., and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychol. Rev.* 87, 329–354. doi: 10.1037/0033-295X.87.4.329

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Klauer, K. J., and Phye, G. D. (2008). Inductive reasoning: a training approach. *Rev. Educ. Res.* 78, 85–123. doi: 10.3102/0034654307313402

Kleijn, S. (2018). Clozing in on *Readability. How Linguistic Features Affect and Predict Text Comprehension and On-line Processing*. Dutch: LOT, Netherlands Graduate School.

Koda, K. (1988). Cognitive process in second language reading: transfer of l1 reading skills and strategies. *Second Lang. Res.* 4, 133–156. doi: 10.1177/026765838800400203

Koda, K. (1990). The use of L1 reading strategies in L2 reading: effects of L1 orthographic structures on L2 phonological recoding strategies. *Stud. Second Lang. Acquis.* 12, 393–410. doi: 10.1017/S0272263100009499

Krieber, M., Bartl-Pokorny, K. D., Pokorny, F. B., Einspieler, C., Langmann, A., Körner, C., et al. (2016). The relation between reading skills and eye movement patterns in adolescent readers: evidence from a regular orthography. *PLoS ONE* 11, e0145934. doi: 10.1371/journal.pone.0145934

Laufer, B., and Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Lang. Test.* 16, 36–55. doi: 10.1191/026553299672614616

Linacre, J. M. (2023) *Facets Computer Program for Many-facet Rasch Measurement*, version 3.85.1. Beaverton, OR: Winsteps.Com.

Linacre, J. M., and Wright, B. D. (2002). Construction of measures from many-facet data. *J. Appl. Meas.* 3, 486–512.

Lipka, O., and Siegel, L. (2012). The development of reading comprehension skills in children learning English as a second language. *Read. Writ.* 25, 1873–1898. doi: 10.1007/s11145-011-9309-8

Manor, B. R., and Gordon, E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *J. Neurosci. Methods* 128, 85–93. doi: 10.1016/S0165-0270(03)00151-1

McNeil, L. (2011). Investigating the contributions of background knowledge and reading comprehension strategies to L2 reading comprehension: an exploratory study. *Read Writ.* 24, 883–902. doi: 10.1007/s11145-010-9230-6

McNeil, L. (2012). Extending the compensatory model of second language reading. *System* 40, 64–76. doi: 10.1016/j.system.2012.01.011

Perfetti, C. (2007). Reading ability: lexical quality to comprehension. *Sci. Stud. Read.* 11, 357–383. doi: 10.1080/10888430701530730

Prichard, C., and Atkins, A. (2016). Evaluating L2 readers' previewing strategies using eye tracking. *Read. Matrix* 16, 110.

Prichard, C., and Atkins, A. (2019). Selective attention of L2 learners in task-based reading online. *Read. Foreign Lang.* 31, 269–290.

Rouet, J.-F., Britt, M. A., and Durik, A. M. (2017). RESOLV: readers' representation of reading contexts and tasks. *Educ. Psychol.* 52, 200–215. doi: 10.1080/00461520.2017.1329015

Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020

Schaffner, E., and Schiefele, U. (2013). The prediction of reading comprehension by cognitive and motivational factors: does text accessibility during comprehension testing make a difference? *Learn. Individ. Differ.* 26(Supplement C), 42–54. doi: 10.1016/j.lindif.2013.04.003

Seidlhofer, B. (1990). Summary judgments: perspectives on reading and writing. *Read. Foreign Lang.* 6, 413–424.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Read. Res. Q.* 16, 32–71. doi: 10.2307/747348

Stoller, F. L., Anderson, N. J., Grabe, W., and Komiyama, R. (2013). Instructional enhancements to improve students' reading abilities. *English Teach. Forum* 51, 2–11.

Taylor, J. N., and Perfetti, C. A. (2016). Eye movements reveal readers' lexical quality and reading experience. *Read. Writ.* 29, 1069–1103. doi: 10.1007/s11145-015-9616-6

Taylor, L. (2013). Testing Reading *through Summary: Investigating Summary Completion Tasks for Assessing Reading Comprehension Ability*. Cambridge: Cambridge University Press.

Urquhart, A. H., and Weir, C. J. (2014). *Reading in a Second Language*: *Process, Product and Practice*. London: Routledge. doi: 10.4324/9781315841373

Wigfield, A., and Guthrie, J. T. (1997). Relations of children's motivation for reading to the amount and breadth or their reading. *J. Educ. Psychol.* 89, 420–432. doi: 10.1037/0022-0663.89.3.420

Yeari, M., Elentok, S., and Schiff, R. (2017). Online and offline inferential and textual processing of poor comprehenders: evidence from a probing method. *J. Exp. Child Psychol.* 155(Supplement C), 12–31. doi: 10.1016/j.jecp.2016.10.011

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 13 July 2023
DOI 10.3389/fpsyg.2023.1124227

OPEN ACCESS

EDITED BY
Lena Ann Jäger,
University of Potsdam, Germany

REVIEWED BY
Vsevolod Kapatsinski,
University of Oregon, United States
Filip Dechterenko,
Academy of Sciences of the Czech Republic (ASCR), Czechia
Alex Warstadt,
New York University, United States

*CORRESPONDENCE
Katrine Falcon Søby
kafs@kp.dk

RECEIVED 14 December 2022
ACCEPTED 22 May 2023
PUBLISHED 13 July 2023

CITATION
Søby KF, Ishkhanyan B and Kristensen LB (2023) Not all grammar errors are equally noticed: error detection of naturally occurring errors and implications for eye-tracking models of everyday texts.
Front. Psychol. 14:1124227.
doi: 10.3389/fpsyg.2023.1124227

Not all grammar errors are equally noticed: error detection of naturally occurring errors and implications for eye-tracking models of everyday texts

Katrine Falcon Søby*, Byurakn Ishkhanyan and Line Burholt Kristensen

Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark

Grammar errors are a natural part of everyday written communication. They are not a uniform group, but vary from morphological errors to ungrammatical word order and involve different types of word classes. In this study, we examine whether some types of naturally occurring errors attract more attention than others during reading, measured by detection rates. Data from 211 Danish high school students were included in the analysis. They each read texts containing different types of errors: syntactic errors (verb-third word order), morphological agreement errors (verb conjugations; gender mismatches in NPs) and orthographic errors. Participants were asked to underline all errors they detected while reading for comprehension. We examined whether there was a link between the type of errors that participants did not detect, the type of errors which they produce themselves (as measured in a subsequent grammar quiz), and the type of errors that are typical of high school students in general (based on error rates in a corpus). If an error is infrequent in production, it may cause a larger surprisal effect and be more attended to. For the three subtypes of grammar errors (V3 word order, verb errors, NP errors), corpus error rates predicted detection rates for most conditions. Yet, frequency was not the only possible explanation, as phonological similarity to the correct form is entangled with error frequency. Explicit grammatical awareness also played a role. The more correct answers participants had in the grammar tasks in the quiz, the more errors they detected. Finally, we found that the more annoyed with language errors participants reported to be, the more errors they detected. Our study did not measure eye movements, but the differences in error detection patterns point to shortcomings of existing eye-tracking models. Understanding the factors that govern attention and reaction to everyday grammar errors is crucial to developing robust eye-tracking processing models which can accommodate non-standard variation. Based on our results, we give our recommendations for current and future processing models.

KEYWORDS

grammar, error detection, proofreading, production, processing models

1. Introduction

Everyday texts, whether it is an email to a colleague or a high school essay, are rarely edited. Such texts often contain grammar errors like anomalous use of word order and lack of agreement between verb and subject (Lunsford and Lunsford, 2008). Attention to these errors is not uniform. In some cases, readers react to the error. In other cases, the error goes by unnoticed. This variation in the reader's attention and response to errors poses a challenge to existing models of eye movement control in reading, such as E-Z Reader (Reichle et al., 2003) and SWIFT (Engbert et al., 2005). Enhancing our understanding of the factors that govern attention and reaction to everyday grammar errors is necessary for developing robust models of eye movement control (Søby et al., 2023). We need models that take into account variation in the type of naturally occurring grammar anomalies that occur in non-standard language and variation in the reader's grammatical awareness and proficiency, as both these factors may modulate attention and eye movements.

Differential attention to language errors has been examined in previous studies using different methods. Proofreading studies show that attention is not equally distributed between different types of language errors (Hacker et al., 1994; Shafto, 2015). Typos like *toujousr* for *toujours* attract more attention than grammar errors, which again attract more attention than orthographic errors with phonological similarity to the correct form, e.g., *essentiellemment* for *essentiellement* (Larigauderie et al., 2020).

Change blindness studies also provide evidence for differential attention allocation. In this paradigm, a participant reads two almost identical sentences, one after another, and responds to whether the two sentences are identical or not. Only one word is changed from the first display of the sentence to the second. Change blindness studies show that readers attend more to changes in lexical elements (e.g., full verbs and demonstrative pronouns) than to changes in grammatical elements (e.g., auxiliaries and articles; Christensen et al., 2021) and that readers attend more to changes in focused words than in non-focused words (Sturt et al., 2004).

Across EEG and eye-tracking studies, the difference between syntactic and semantic/pragmatic anomalies is well-documented (Ainsworth-Darnell et al., 1998; Ni et al., 1998; Braze et al., 2002; Hahne and Friederici, 2002; Hagoort, 2003). Grammar errors, however, are usually treated as a homogenous group, although grammar errors involve various subtypes (word order errors, verb agreement errors, gender mismatch errors etc.) which are not necessarily noticed to the same degree or not necessarily processed in the same way. With the present study we ask, if sensitivity to different kinds of grammar errors differs too, and what the consequences are for existing models of eye movement control in reading.

Using an error detection paradigm, we study the differences in attention to different types of naturally occurring grammar errors in written Danish. Some error types involve attention to confusion of large elements (e.g., word order errors), while others involve smaller segments at the level of words, suffixes and letters. Some errors appear initially in a sentence. Other errors have a medial or final position. Some grammar errors have phonological similarity with the correct form, and others are distinct. Many of these factors co-vary in naturally occurring errors and cannot be completely disentangled. In our study, we focus on how error type, error frequency in written production and phonological similarity to the correct form affect readers' perception of and attention to grammar errors in Danish. For word order errors, we also consider the position of the misplaced word in the sentence.

Previous studies of writers' spelling accuracy show that exposure to incorrectly spelled words tends to negatively influence later spelling accuracy for those same words (Jacoby and Hollingshead, 1990). Building on these findings, we propose that previous exposure to specific types of incorrectly inflected or misplaced words may also affect attention to this specific type of grammar errors during reading. We also examine the relationship between the type of errors that young readers tend to overlook in texts, the type of errors these young readers produce themselves (when performing a grammar quiz), and the type of errors that are typical of their age group in general (based on corpus studies of naturally occurring texts). Some grammar errors in our study represent types of errors that frequently occur in Danish high school essays. Other errors are less typical of high school students, but characteristic of L2 learners of Danish. We investigate if these typical L2 grammar errors attract more attention than the grammar errors typical of high school students. Our expectation is that attention to a specific type of grammar error is not only a matter of the reader's explicit grammar awareness (as measured in the grammar quiz), but also of whether the specific type of error is common in everyday texts by native speakers. If a specific type is frequent among the peers of the reader, the reader may have more exposure to this type of error and a mental representation of the error. The reader may therefore find it less striking and be less likely to detect it than errors that are infrequent in texts written by peers.

2. Background

Our error detection study does not involve eye-tracking data, but in combination with insights from previous eye-tracking studies on processing of grammar errors, it can address shortcomings in current models of eye movement control during reading. In this section, we present previous eye-tracking studies on processing of grammar errors (section 2.1), and describe the role of grammar errors in existing models of eye movement control in reading (section 2.2). In section 2.3, we describe the error detection paradigm, and how this may contribute to research in attention during reading. We also present the error types chosen for this study. Finally, in section 2.4, we provide an overview of the main factors presumed to influence attention to errors.

2.1. Previous eye-tracking studies on processing of grammar errors

Previous eye-tracking studies of grammar errors differ with respect to language, error types, purpose of the study, and the included reading measurements. Therefore, the findings cannot be easily summarized.

First, the eye-tracking studies have been conducted in different languages (English, Hebrew and Norwegian), making it difficult to

Abbreviations: DEF, Definite; INF, Infinitive; N, Neuter; PRS, Present tense; U, Uter (common gender).

compare across studies. For example, it is difficult to compare Hebrew subject-verb gender agreement to Norwegian word order.

Second, the ungrammatical items are very different, ranging from word order errors such as *The white was cat big* (Huang and Staub, 2021), Norwegian *ASV word order instead of AVS (Søby et al., 2023) to various morphosyntactic agreement errors such as gender agreement (Deutsch and Bentin, 2001; Dank et al., 2015), subject-verb agreement (Pearlmutter et al., 1999; Lim and Christianson, 2015) or modals followed by a progressive form, e.g., *It seems that the cats will not usually eating the food we put on the porch* (Ni et al., 1998), and/or a past tense form (Braze et al., 2002).

Third, previous studies have had different reasons for including ungrammatical items. Their experimental contrasts differ and their results can be difficult to compare. Huang and Staub (2021) examined failure to notice transposition errors to enter a debate about serial vs. parallel processing. Other studies focus on the differences between pragmatic and syntactic processing (Ni et al., 1998; Braze et al., 2002), or the interrelation between semantic and syntactic factors during processing of agreement in Hebrew (Deutsch and Bentin, 2001). Other studies again have investigated the attraction phenomenon, i.e., when a word erroneously agrees with a local distractor noun instead of the head noun, e.g., *The key to the cabinets were rusty from many years of disuse* (Pearlmutter et al., 1999), both in English (Lim and Christianson, 2015) and for subject-predicate agreement in Hebrew (Dank et al., 2015).

Finally, the studies use different reading measurements. While some measure very early effects, such as first fixation duration (Deutsch and Bentin, 2001; Braze et al., 2002; Dank et al., 2015; Lim and Christianson, 2015; Huang and Staub, 2021; Søby et al., 2023); others do not (Ni et al., 1998; Pearlmutter et al., 1999).

Taking these reservations into account, it seems that the different types of grammar errors elicit similar responses in participants' eye movements across languages, with similar time courses. Most of the studies find more regressions out from the error, meaning that participants respond immediately. Most studies also find increased reading times, but the time course varies (see Søby et al., 2023). Very early effects are found on first fixation duration by Deutsch and Bentin (2001), Dank et al. (2015), Huang and Staub (2021), and partly by Søby et al. (2023). Other studies only find increased total durations on the critical region (Pearlmutter et al., 1999) or no reading time effects at all (Ni et al., 1998). Typically the effects of ungrammaticality quickly disappears, either in the critical or subsequent regions.

Only one of the previous eye-tracking studies has explicitly examined whether readers perceived the ungrammatical items as errors or not. Huang and Staub (2021) used readers' grammaticality judgments of each sentence to distinguish between detected and undetected errors. None of the studies have made direct comparisons between different types of grammar errors to examine whether participants elicit stronger or different reactions to some errors than others. Therefore, little is known about the factors that govern attention and reaction to different types of grammar errors. Furthermore, the ecological validity of grammar errors have not been the focus of previous studies. Errors such as transposed words are constructed for the purpose of the experiment, but infrequent in natural language, and therefore may not reflect reading processes for naturally occurring language. Understanding the factors that govern attention and reaction to different types of naturally occurring errors is a necessary prerequisite when developing robust eye-tracking models for reading everyday texts (Søby et al., 2023).

2.2. The role of grammar errors in existing models of eye movement control in reading

Attention to, and processing of, grammar errors have not been a focal point in existing models of eye movement control in reading. Existing models can be divided into two types. *Serial-attention* models share the assumptions that attention is allocated serially, and only to one word at a time, while *attention-gradient* models assume that attention is allocated as a gradient, i.e., to multiple words at a time (Warren, 2011, p. 919). The major models are the influential E-Z Reader (Reichle et al., 2003, 2009; Reichle, 2011), a serial-attention model, and SWIFT, an attention-gradient model (Engbert et al., 2005; Engbert and Kliegl, 2011). Serial-attention models are furthermore described as *cognitive control* models, because they assume that "lexical processing is the 'engine' that determines when the eyes will move from one word to the next during reading" (Reichle, 2011, p. 768), in contrast to models like SWIFT, in which cognition is assumed to play a reduced role for eye movements. For example, the signal to move the eyes forward in SWIFT is provided by an autonomous random timer.

Both E-Z Reader and SWIFT account for effects of lexical processing on eye movements, but a widely acknowledged shortcoming of both models is that they cannot account for effects of higher-level language processing on eye movements (Clifton and Staub, 2011; Warren, 2011). The issue has not been addressed in SWIFT, but for E-Z Reader, Reichle et al. (2009) added a post-lexical integration stage, which is assumed to reflect all of the postlexical processing that is required to integrate a word, *n*, into the higher-level representations which readers construct online. As exemplified by Reichle et al. (2009, p. 5f), this could be to link word *n* into a syntactic structure, to generate a context-appropriate semantic representation, and to incorporate its meaning into a discourse model. Reichle et al. (2009, p. 6) state that "the integration stage […] is a placeholder for a deeper theory of postlexical language processing during reading. Our goal in including this stage is therefore quite modest: to provide a tentative account of how […] postlexical variables might affect readers' eye movements."

In E-Z Reader ver. 10 (Reichle et al., 2009; Reichle, 2011), lexical processing of a word takes place in two stages. First, the early stage of lexical processing (or word identification), also known as L_1 or the *familiarity check*, takes place. This stage corresponds to the identification of the orthographic form of the word, assuming that "this is not full lexical access, as the phonological and semantic forms of the word are not yet fully activated" (Reichle et al., 2003, p. 452). When completed, i.e., when the feeling of familiarity concerning the word exceeds a threshold corresponding to the familiarity check, it triggers the initiation of the programming of a saccade to move the eyes to the next word (Reichle, 2011). The time required to finish the familiarity check depends on the frequency of a word and its cloze probability, defined as the proportion of subjects who are able to guess word *n*, when shown the rest of the sentence (Reichle et al., 2009:3). This predicts that frequent and/or predictable words are processed faster than infrequent and/or unpredictable words (Reichle, 2011).

We assume that the same reasoning applies to frequent and/or predictable errors, but the E-Z Reader model does not explicitly account for input with frequent vs. infrequent errors.

The later stage of lexical processing (L_2) involves the identification of the word's phonological and/or semantic forms, to enable additional linguistic processing (Reichle et al., 2003). This stage corresponds to what is typically referred to as *lexical access,* and with the completion of lexical access, attention shifts to the next word, which can now be processed. Simultaneously, post-lexical processing (i.e., integration) starts on the identified word. This post-lexical processing corresponds to the minimal amount of processing necessary to continue to move attention (and the eyes) forward through the text (Reichle, 2011, p. 776). In most cases, integration is completed without difficulty, meaning that post-lexical processing only has minimal effect on readers' eye movements. Reichle et al. (2009, p. 6) assume that complete incremental post-lexical processing is not always necessary and does not always occur, which is broadly consistent with the "good enough" view of language processing (Ferreira and Patson, 2007). However, integration difficulty may occur. When integration fails, it causes the eyes and attention to pause and/or move backwards (Reichle, 2011). Integration failures happen by default when word $n+1$ is identified before word n is integrated. Rapid integration failure can happen due to severe semantic or syntactic violations (Reichle et al., 2009). If the integration of n fails rapidly, the forward saccade to $n+1$ is canceled, which results in a pause (increasing first fixation duration and gaze duration) and/or a refixation (increasing gaze duration) or an interword regression (Reichle et al., 2009). If the integration failure of n takes place after the eyes have moved to $n+1$, i.e., fails more slowly, a regressive eye movement is made (Clifton and Staub, 2011, p. 904). Thus, the model predicts that problems with integration can have very rapid effects, influencing first-fixation duration on the word that is being integrated. This, however, only happens when the integration failure occurs before the labile stage of saccadic programming (i.e., the stage which can be canceled) to move the eyes forward in the text has completed (Reichle et al., 2009).

The assumption that contextual information (besides cloze probability) only affects postlexical integration is challenged by studies of parafoveal processing, i.e., processing of upcoming words that have been attended, but not yet fixated (Warren, 2011). For example, Veldre and Andrews (2018) used the gaze-contingent boundary paradigm to assess whether parafoveal processing of a word contributes to its subsequent identification. In this paradigm, a target word in a sentence is replaced with another word, until the reader's eyes cross an invisible boundary (e.g., before the space to the left of the target word), after which the word is changed back to the target word. Veldre and Andrews (2018) conducted two experiments, in which they compared contextually plausible previews (which either contained a morphosyntactic agreement violation or not) to implausible previews (either containing a syntactic word class violation or not). The plausible previews were not predictable from the sentence context, as measured in a cloze task. Veldre and Andrews (2018) found that the contextual plausibility and grammatical correctness of an upcoming word can affect processing, early enough to affect skipping of that word. According to the authors, the plausibility effects on skipping rates are unlikely to be reconciled with E-Z Reader's current post-lexical integration mechanisms.[1]

Furthermore, the E-Z Reader model does not address what happens when readers encounter other types of misspellings or grammar errors, besides severe syntactic violations. If the early familiarity check identifies the orthographic form of the word, it should be able to respond to orthographic errors (e.g., *posibility*), but not anomalous use of existing morphological forms (e.g., *eats* for *eat*). The model does not answer the question of why some types of errors are detected while others are not, nor the question of why readers do not always notice the same error.

Finally, Warren (2011) argues that the E-Z Reader model will be incomplete without allowing some role for even higher-level influences, based on research on semantic anomalies. Readers sometimes fail to notice semantic anomalies, suggesting that processing is sometimes shallow (Ferreira et al., 2002). "If different readers, reading for different purposes, perform post-lexical processing more or less quickly or completely […], the precise combination of reader, purpose and motivation will affect the patterns of eye movements to semantic violations" (Warren, 2011, p. 922). In our study, we examine how error detection differs between readers with differences in grammatical awareness and proficiency.

2.3. The error detection paradigm

Both the eye-tracking and error detection paradigms can be used to measure attention during reading. Here we assume that eye-tracking provides a more sensitive measure than error detection. Yet, the exact correlations between the two measures is not well-explored. It may be the case that the error detection paradigm treats two events as the same, while they involve different eye movements. Although we assume that error detection is more rough, there are several advantages to using this paradigm for our purpose. In the previous eye-tracking studies of ungrammaticality, sentences were presented individually. With error detection, we can introduce participants to long, consecutive texts, simulating natural text reading. Furthermore, we can include many different types of grammar errors, unlike previous eye-tracking studies which have included relatively few error types (e.g., pragmatic vs. syntactic). Having many different types of errors in different conditions would result in a long and tiresome eye-tracking experiment. Finally, using error detection, we can get participants' feedback on where errors occur, in a fast way, not having to ask after every trial. Although, error detection can only provide a rough measurement for attention during reading, it can provide insights into which types of errors are more noticed than others, and which other factors than error type is likely to play a role. The results are therefore relevant to future eye-tracking studies and processing models. If differences are found using error detection, they are also likely to be found using a presumably more sensitive measure such as eye-tracking.

1 Veldre and Andrews (2018) also argue that the results cannot be reconciled with the alternative *forced fixation account* of preview effects, proposed by Schotter et al. (2014b).

In our error detection study in Danish, we included one type of syntactic error (*ASV for AVS, see below) and two types of morphological errors (confusion of infinitive and present tense, and gender mismatches between articles or adjectives in NPs), as well as various common orthographic errors. These errors were chosen because they represent a broad range of error types, and they are all attested in natural L1 and/or L2 production, however with different frequencies. For example, ungrammatical verb-third word order (*ASV) instead of grammatical verb-second word order (AVS) is common in L2 Danish (Søby and Kristensen, 2019; Søby and Kristensen, to appear), but rare in L1 Danish, apart from multiethnic urban vernaculars (Quist, 2008). The three types of grammar errors naturally occur in different conditions, varying with respect to error frequency (measured as error rates in L1 production), and/or phonological similarity to the correct form, or placement in the sentence. Since the stimuli is based on naturally occurring errors, error frequency and phonological similarity tend to co-vary.

2.4. Attention to errors during reading

Many potential contributing factors besides error type might influence whether a reader reacts to an error. In this section, we elaborate on why some of the factors we are examining in our study are relevant to include, namely error frequency, phonological similarity to the correct form, and, for word order, placement in the sentence. Finally, we elaborate on the potential role of participants' own production of errors, and individual differences in error perception.

Previous letter detection studies and change-blindness studies review a wide a range of factors which can influence attention during reading (e.g., Smith and Groat, 1979; Sturt et al., 2004; Vinther et al., 2015; Christensen et al., 2021). For example, Smith and Groat (1979) found that the position on the line and in the sentence influenced detection of the letter *e*, so that the outer positions were more prominent than the middle. Using V3 errors with a length manipulation, we examine whether position effects within the sentence are also found for grammar errors.

The main focus of our study is on the role of error frequency. We hypothesize that error frequency, which is tied to the predictability of the error, predicts perception patterns. According to prediction-based approaches to sentence processing, unexpected input attracts attention (Kamide, 2008; Levy, 2008; Christiansen and Chater, 2016). If a reader sees input with common errors, the model will be updated according to the input, meaning that frequent errors should be predicted by the model, and thus should attract less attention than infrequent errors.

Besides error frequency, we expect that phonological similarity to the correct form negatively influences detection rates for grammar errors, in line with Larigauderie et al. (2020) who compared spelling errors which were either phonologically similar to or distinct from the correct form. One example from our stimuli is confusion of homophone verb pairs, such as present tense *kører* and infinitive *køre*, both pronounced [ˈkʰøːɐ]. We expect that confusion of heterophone verb pairs such as *rejser* [ˈʁɑjˀsɐ] and *rejse* [ˈʁɑjsə] will have higher detection rates. When the correct form is homophone to the error, the error is not grammatical in that context, but it is phonologically correct, and may therefore not disturb reading. For such silent errors readers may use all available cues whether they are phonological or orthographic (*cf.* Carassco-Ortiz and Frenck-Mestre, 2014). The E-Z Reader model does not account for homophony effects, but it may predict that the phonological form is more easily identified for homophone compared to heterophone errors in the later stage of lexical processing (L_2). The error frequency and phonological similarity to the correct form tend to co-vary, because L1 speakers of Danish produce more errors when for instance present tense and infinitive forms are homophone. Thus, effects of phonological similarity and frequency are often difficult to disentangle.

On top of that, individual differences are likely to influence error detection. If a type of error is frequent in a person's production, e.g., omitting the-r on verbs in present tense: **han køre* 'he drive.INF,' the rules for verbal inflection may not be fully mastered. It therefore seems likely that this person will overlook this type of error in general. Individual differences in the perception of what constitutes an error in a specific situation could also be a factor: How correct or incorrect on a continuum is an error to a specific reader? How do individual readers differentiate between unusual language and outright errors? And is the perception affected by the context in which it is read, e.g., experimental vs. natural? Our study is not equipped to answer these questions. Studies show that tolerance for various errors can be modulated by participants' perception of the speaker, so that the tolerance and willingness to repair is higher when the speaker is perceived as non-native (Konieczny et al., 1994; Hanulíková et al., 2012; Gibson et al., 2017).

In the public debate and prescriptive literature, some errors are pointed out as typical or basic errors, while other errors are much less debated or accentuated. Publically debated errors may be more prominent to readers (Blom and Ejstrup, 2019b). In Denmark, missing present tense-r is often accentuated in normative discourse. Blom and Ejstrup (2019b) found that readers' intolerance for errors are modulated by the type of error. Their participants were more annoyed with typical and basic grammar/spelling errors than with atypical and complicated errors. The missing present tense *-r* was the most annoying error. The authors also found a correlation between participants' irritation (with a specific item) and the number of errors detected, so that the more errors participants detected in general, the more irritated they were with that item.

2.5. The current study

The current study examines native speakers' attendance to different types of syntactic, morphological and orthographic errors (found in L1 and/or L2 Danish) during reading. We asked Danish high school students to read and comprehend two texts, while underlining all errors they noticed. We also tested their basic grammar skills, using a grammar quiz. The study included one type of syntactic error (V3 word order) and two types of morphological errors (confusion of infinitive and present tense, and gender mismatches between articles or adjectives in NPs), as well as various common orthographic errors. V3 errors are the least frequent, and orthographic errors the most common. In a corpus of 71 high school essays, we found 10 V3 errors, 16 gender mismatches in indefinite articles, 51 gender mismatches in adjectives, 178 confusions of infinitive and present tense, and 1,099 orthographic errors.

The study is designed as a four-in-one study. Each error type (V3, verb, NP, orthographic) constitutes its own subexperiment and appears in different conditions, controlled for a number of variables. We cannot directly compare attention to the four types, as there are too many confound variables, such as their position in the sentences and in the text. Thus, we only indirectly compare the detection rates for the three overall error categories (syntactic, morphological, orthographic) using descriptive statistics.

We examine the relationship between the type of errors that young readers tend to overlook in texts, the type of errors these young readers produce themselves (in the grammar quiz), and the type of errors that are typical of their age group in general (based on corpus studies of high school essays). Our expectation is that attention to a specific type of grammar error is not only a matter of the reader's explicit grammar awareness (as measured in the grammar quiz), but also of whether the specific type of error is common in everyday texts by native speakers. If a specific type is frequent among the peers of the reader, the reader may have more exposure to this type of anomaly and a mental representation of the error, i.e., common errors should be predicted to occur in input, based on prediction theory (Kamide, 2008; Christiansen and Chater, 2016). The reader may therefore find it less striking and be less likely to detect it than errors that are infrequent in texts written by peers, e.g., those found in L2 Danish. This means that for the overall categories of errors (syntactic, morphological and orthographic), we expect that the syntactic errors (V3 errors) have higher detection rates than morphological and orthographic errors, because V3 errors are rare in L1 writing (and are visually large). We also expect readers to overlook orthographic errors the most, because orthographic errors are highly frequent in the L1 writing.

Finally, for the two morphological subtypes of grammar errors (confusion of infinitive and present tense, and gender mismatches between articles or adjectives in NPs), we examine how error frequency and phonological similarity to the correct form may affect attention to errors. For the word order errors, we examine position effects within the sentence. The specific conditions and hypotheses for the three subtypes of grammar errors are presented in the results section where they are treated as three subexperiments. The fourth subexperiment on different types of orthographic errors is primarily included to create variation in the stimuli.

3. Methods

3.1. Participants

The participants were recruited from three different Danish upper secondary education programs (STX, HTX, and HHX).[2] Data were collected in August 2019 at six schools located in and around Copenhagen and Roskilde. Two hundred and forty students from 10 classes participated. We excluded participants with dyslexia (18), with late acquisition of Danish (>6 years, Hyltenstam and Abrahamsson, 2003) (2), or participants who misunderstood or did not finish the reading task (9). This left 211 participants in the analysis (98 women, 113 men), 17–20 years of age (M = 18.31 years; SD = 0.67 years). The majority were part of the STX Program (130), followed by HHX (43), and HTX (38). All participants (or their parents) gave informed written consent prior to the experiment. The study was approved by local research ethics committee at University of Copenhagen, and followed GDPR.

3.2. Experimental tasks and materials

The experimental tasks consisted of a reading task (section 3.2.1) which was followed by a grammar quiz and a questionnaire (section 3.2.2). All test materials are found in Supplementary material (section 3).

3.2.1. Reading task

The reading task consisted of two texts, A (689–692 words) and B (831–832 words). Every participant read both texts. There were four versions of the reading task material to ensure that each participant only saw the same item in one condition. That is, when reading the same sentence in the text, participants reading version 1 were presented with the verb error in one condition, participants reading version 2 were presented with it in another condition, etc. Each participant was presented with a total of 100 errors in text A and B together. Table 1 shows the distribution on subtypes. To avoid priming effects, target items did not occur elsewhere in the texts.

A further description of the stimuli is presented in the sections on each subexperiment. We varied the order of text A and B, so that half of the participants read A before B, and the other half read B before A. Thus, there were eight versions of the reading task in print.

3.2.2. Questionnaire and grammar quiz

The questionnaire addressed the participants' language and dialectal background as well as their attitude to language errors. The purpose of the grammar quiz was to ensure that the participants had the basic grammatical prerequisites to notice errors in the reading task. The grammar quiz included tests on all four types of errors, i.e., verb-second word order after sentence-initial adverbials, verb conjugations in infinitive and present tense, conjugation of adjectives, gender of indefinite articles, and spelling of the four types of target words. Most of the tasks were forced-choice between two options.

3.3. Procedure

The participants were informed that the study was about speed-reading and what readers notice when skimming a text. In the reading task, their task was to underline language errors. Participants had max. 7 min to read each text (A and B). Participants were instructed to skim as fast as possible and finish reading the whole text so they could answer the comprehension questions. Whenever they noticed a language error, they should underline it, but they should avoid going back in the text. Language errors were defined as different types of spelling and grammar errors, but not punctuation. They were instructed to underline the whole word containing the error, or multiple words if they were in the

2 The three education programs (STX, HTX, and HHX) all prepare for higher education, but have different profiles. STX is a general examination program, HTX is a technical examination program with a STEM profile and HHX is a commercial examination program with a business profile (Ministry of Higher Education and Science, 2022).

10.3389/fpsyg.2023.1124227

TABLE 1 Error types, conditions and number of target items in the reading task (text A+B).

Error types		Items
V3 errors (2 conditions, 8 items per condition)		**16**[a]
1) After short adverbial:	*og kl. 14 han ankommer til Berlin* and o'clock 2 he arrive.PRS in Berlin 'and at 2 o'clock, he arrives in Berlin'	8
2) After long adverbial:	*og først ud på eftermiddagen han ankommer til Berlin* and first out on afternoon.DEF he arrive.PRS in Berlin 'and first in the afternoon, he arrives in Berlin'	8
Verb errors (4 conditions, 8 items per condition)		**32**
1) Homophone; Present tense for infinitive:	*han vil kører* [ˈkʰøːɐ] he will drive.PRS 'he'll drive'	8
2) Homophone; Infinitive for present tense:	*han køre* [ˈkʰøːɐ] he drive.INF 'he drives'	8
3) Heterophone; Present tense for infinitive:	*han vil rejser* [ˈʁɑjˀsɐ] he will travel.PRS 'he'll travel'	8
4) Heterophone; Infinitive for present tense:	*han rejse* [ˈʁɑjsə] he travel.INF 'he travels'	8
NP errors (4 conditions, 8 items per condition)		**32**
1) Mismatch ADJ + N; Uter for neuter:	*et dejlig kæledyr* ART.N lovely-U pet.N 'a lovely pet'	8
2) Mismatch ADJ + N; Neuter for uter:	*en dejlig-t undulat* ART.U lovely-N budgie.U 'a lovely budgie'	8
3) Mismatch ART + N; Uter for neuter:	*en dejlig-t kæledyr* ART.U lovely-N pet.N 'a lovely pet'	8
4) Mismatch ART + N; Neuter for uter:	*et dejlig undulat* ART.N lovely-U budgie.U 'a lovely budgie'	8
Misspellings (4 types — 5 of each type)		**20**[b]
1) Missing double consonant, e.g., *startskudet* for *startskuddet* 'the starting signal'		5
2) Split compounds, e.g., *by vandring* for *byvandring* 'city walk'		5
3) Missing silent letter, e.g., *siste* [ˈsisd̥ə]/[ˈsisd̥] for *sidste* [ˈsisd̥ə]/[ˈsisd̥] 'last'		5
4) Reduction of syllable, e.g., *virklig* [ˈʋiɐ̯ɡ̊li] for *virkelig* [ˈʋiɐ̯ɡ̊li] 'really'		5
Total		**100**

[a]The V3 errors in version 1 + 2 were identical. The V3 errors in version 3 + 4 were also identical.
[b]The 20 spelling errors were identical in all four versions of the reading task.

wrong order. Underlinings could be canceled with a vertical line. Use of dictionaries and online tools were not allowed.

The researcher registered the starting time and gave statuses on remaining time. When the students finished reading the text, they wrote the finishing time and put the text away (if they did not finish, they marked how far in the text they got). The same procedure was repeated for the second text. Finally, the students completed the comprehension questions for both texts, the questionnaire and the grammar quiz. The whole session lasted around 45 min.

4. Analysis

The error detection data were analyzed with general linear mixed effects models for binomial data in RStudio (R Core Team, 2022,

version 2022.07.1), using the lme4 package (Bates et al., 2015, ver. 1.1.30). *p*-values were obtained using the lmerTest package (Kuznetsova et al., 2017, ver. 3.1.3). The dependent variable for all models was detection, i.e., whether the error was detected (=1) or not (=0). We did not penalize false hits. The conditions for each of the four error types were included in the models as fixed effects (p is the probability of correctly detecting an error):

1. Model for V3 errors: $\log(p/1\text{-}p)^3$ = Adverbial length [short vs. long] + Total grammar score + (1|Participant) + (1|Item) + Residuals
2. Model for Verb errors: log(p/1-p) = Type [infinitive for present tense vs. present tense for infinitive]*Homophony [homophone vs. heterophone pairs] + Total grammar score + (1|Participant) + (1|Item) + Residuals
3. Model for NP errors: log(p/1-p) = Type [agreement with article vs. adjective]*Gender [uter for neuter vs. neuter for uter] + Total grammar score + (1|Participant) + (1|Item) + Residuals
4. Model for orthographic errors: log(p/1-p) = Type [four different] + Spelling score + (1|Participant) + (1|Item) + Residuals

All models included random intercepts for participant and item. All models also included the scores from the grammar quiz. Participants made few wrong answers in the grammar tasks, so we summarized the results from the individual grammar-related tasks and included a total grammar score as a fixed effect in the models for detection of the three types of grammar errors. The model for orthographic errors included the score from the spelling task in the quiz as a fixed effect.

The models for the four error types did not include random slopes, presentation order (i.e., placement in the text) or irritation scores, as the models failed to converge when they were included. Only one subtype, NP errors, showed an uninterpretable effect of presentation order.

The output of the regression model was in logodds space. To increase interpretability, they were converted back to probabilities and plotted. Thus, the plots for the morphological errors show the models' predicted probabilities of detecting the target.

Finally, we made a general model, collapsing all error subtypes, with accuracy in percentage as the dependent variable, only including irritation scores as a fixed effect (see normal Q-Q plot in Supplementary Figure 3):

5. Model for all errors: accuracy (%) = Irritation score + Residuals.

5. General results

The participants detected 54% of all errors in the two texts (Table 2). As expected, the highest detection rate was found for syntactic errors (71% of all items were detected), followed by the two types of morphological errors (55% detected for NP errors; 59% for verb errors), and the lowest rate was found for orthographic errors (33%). The study is not designed to directly compare these overall categories (syntactic, morphological and orthographic), as there are a number of confounds, such as their position in the sentences and in the text. We therefore do not conduct any statistical tests between them. More detailed results are presented in the sections on each of the four error types (subexperiments).

5.1. Individual variation

As seen in Figure 1, there was individual variation among the participants, with respect to the number of words they underlined, and the share of correct (hits) vs. incorrect underlinings (false alarms). Out of 321,145 words, participants underlined 18,041 words (M = 85,50 words, SD = 31,38 words, range = 9–227 words). Of these only 2,565 words were not part of a target, i.e., false alarms (M = 12,16 words, SD = 13,59 words, range = 0–108). In total, 11,490 targets were underlined, i.e., hits (M = 54,45 words, SD = 21,32 words, range = 1–92 words). Notice that a target can consist of several words (targets are defined in the sections on the subexperiments).

In principle, participants could underline all words in the text and thus detect all errors, resulting in the highest possible score. This, however, was not an issue in general as participants only underlined

3 If p is the probability of detecting an error, 1 – p is the probability of not detecting an error. p/1 – p is the odds of detecting an error and log(p/1 – p) is the logarithm of the odds (logodds).

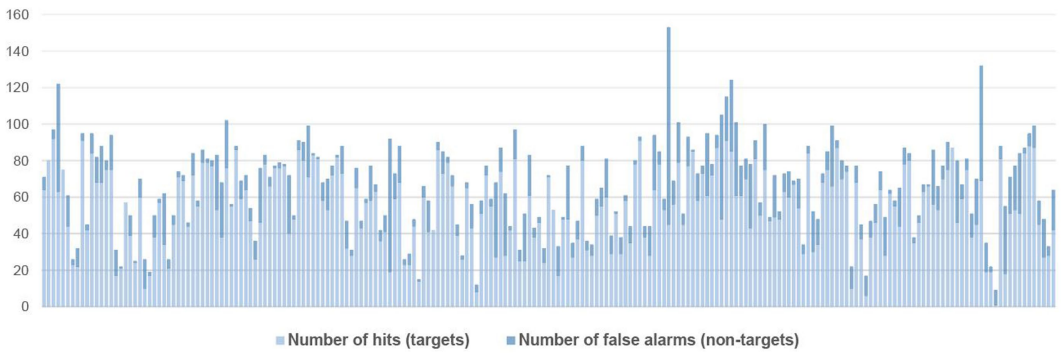


FIGURE 1
Number of underlinings (hits and false alarms) per participant.

10.3389/fpsyg.2023.1124227

0.8% non-target words in the texts (2,565 out of 321,145 words). Figure 1 shows that most participants were relatively exact in their underlinings, apart from 10 participants who had more false alarms than hits.

In the grammar quiz, participants generally made few errors (see sections on subexperiments). In the three grammar tasks (word order, NP agreement and verb conjugations), the highest possible score was 17, one point for each correct answer. Participants' scores had an average of 16.76 (SD = 0.67, range: 11–17). The Supplementary material (section 1.2) include a plot of the total quiz scores (grammar and spelling tasks) and the number of detected errors per participants.

The general model of all error types (5) included the participants' irritation scores (*cf.* Supplementary Table 13). We found a small effect of irritation ($\hat{\beta} = 1.82$, SE = 0.40, $t = 4.57$, $p < 0.001$), so that the more annoyed participants state to be with language errors, the more errors they detected in the reading task (see plots in Supplementary material, section 1.2).

6. Subexperiments

In the following sections, we present the hypotheses, stimuli and results for each of the four subtypes of errors. Sections 6.1–6.3 describe the three subexperiments on grammar errors. Section 6.4 describes the subexperiment on orthographic errors. The Supplementary material show all stimuli (section 2) and model results for the orthographic errors (section 1.1).

For the grammar errors, we start each section with information on error frequencies in L1 production. The error frequencies are based on a corpus of 71 high school essays from a final exam (127,957 words; 71 participants). For the morphological errors, we calculated the error rate by dividing the number of incorrect tokens with the number of correct and incorrect tokens. As an example, when a reader sees a verb in present tense, the error rate reflects how often the verb is incorrect. For the orthographic errors, the error rate is calculated by dividing the number of errors with the number of words in the corpus. For the syntactic errors, we report the absolute number of errors. Since there was a limited number of tokens for certain types of errors, we only use descriptive statistics (not inferential statistics) when accessing differences in error frequency.

TABLE 2 Number of errors in texts and share of detected errors.

Category type	Errors in texts (*N*)	Detected targets (*N*)	Share of detected targets (%)
Syntax			
V3	3,376	2,398	71.03%
Morphology			
Verb errors	6,752	3,992	59.12%
NP errors	6,752	3,719	55.08%
Orthography			
Misspellings	4,220	1,381	32.73%
Total	**21,100**	**11,490**	**54.45%**

6.1. V3 errors

A common word order error in L2 Danish is placing the verb in third position (V3), instead of second (V2; Søby and Kristensen, to appear). In (1a), the adverbial *nu* 'now' is placed in first position, followed by the subject *jeg* 'I' in second position, and the verb *bor* 'live' in third position. In the corrected version of the sentence in (1b), the verb is correctly placed in second position (the mandatory position for finite verbs in Danish main clauses).

(1) a. [original] **Men nu jeg bor i Denmark*
'but now I live in Denmark'

b. [corrected] *Men nu bor jeg i Danmark*
'but now live I in Denmark'

In the L1 corpus of high school essays, we only found 10 V3 errors. V3 errors are generally not considered typical L1 errors, but may occur in informal texts written by speakers of multiethnic urban vernaculars (Quist, 2008).

We expected these errors to be highly noticed by native speakers for two reasons. First, they are rare in L1 production. Second, large elements, i.e., entire words, are misplaced. In the experiment, the V3 errors were either presented after a short sentence-initial adverbial (1–2 words, consisting of 5–12 characters including spaces) or a long adverbial (4–6 words, 26–39 characters). In L2 Danish, V3 word order most frequently occurs after adverbials, both short and long (Søby and Kristensen, to appear). Examples of the stimuli are shown in Table 3. Previous letter detection studies have found position effects, so that elements in the start or end of a sentence tend to be more prominent than

TABLE 3 Conditions, number of V3 errors in texts and share of detected errors.

Conditions	Errors in texts (*N*)	Detected targets (*N*)	Share of detected targets (%)
Short A *og kl. 14 han ankommer til Berlin* and o'clock 2 he arrive.PRS in Berlin 'and at 2 o'clock, he arrives in Berlin'	1,688	1,200	**71.09%**
Long A *og først ud på eftermiddagen han ankommer til Berlin* and first out on afternoon.DEF he arrive.PRS in Berlin 'and first in the afternoon, he arrives in Berlin'	1,688	1,198	**70.97%**

TABLE 4 Model (1) estimates for V3 errors.

Random effects	Variance	Std. dev.		
Participant (intercept)	1.7076	1.3068		
Item (intercept)	0.4177	0.6463		
Fixed effects	**Estimate**	**Std. error**	***z*-value**	***p*-value**
(Intercept)	−10.91748	2.70189	−4.041	5.33e-05***
Length	−0.03394	0.08865	−0.383	0.702
Total grammar score (quiz)	0.72652	0.16094	4.514	6.36e-06***

Dependent variable: detection (1 = error detected, 0 = error not detected). Significance code: ***$p < 0.001$.

TABLE 5 Error rates in L1 texts, confusion of present tense and infinitive (*N*=194).

Type and error rates	Homophone	Heterophone
	e.g., *køre(r)* [ˈkʰøːɐ]	e.g., *rejse* [ˈʁɑjsə], *rejser* [ˈʁɑjˀsɐ]
Target form: present tense 1% errors (12,764 correct present tense verbs[1])	25% (*N* = 96)	0.30% (*N* = 35)
Target form: infinitive 1% errors (4,689 correct infinitives[1])	8.60% (*N* = 37)	1.10% (*N* = 10)

[1]Found using an automatic POS tagger [Centre for Language Technology, University of Copenhagen (CST), 2022], manually tagged for homophony.

in the middle (Smith and Groat, 1979). We therefore examined whether participants would detect more V3 errors after a short adverbial than a long adverbial.

The target verbs were all in present or perfect tense, and subjects were either pronouns, proper names or nouns in the definite form, with varying lengths. The texts also included 16 similar correct constructions with AVS, i.e., V2 word order (8 after short adverbials; 8 after long). All stimuli can be seen in Supplementary material (section 2).

The V3 errors were considered detected when either the adverbial, subject or verb was underlined by a participant, since the order of subject and verb would be correct if the adverbial was placed elsewhere. In Table 3, the number and share of detected targets are seen. There were no effects of adverbial length ($\hat{\beta}$ = −0.03, SE = 0.09, $z = -0.38$, $p = 0.70$), but there was an effect of total grammar score ($\hat{\beta}$ = 0.73, SE = 0.16, $z = 4.51$, $p < 0.001$; *cf.* Table 4). The higher grammar score in the quiz, the more V3 errors were detected. In the grammar quiz, participants had to place words in the correct order after conjunctions and adverbials. Out of 633 answers, only 3 were wrong (0.5%), confirming that V3 is not a typical L1 error.

6.2. Verb errors

Confusion of finite and infinite verb forms is the most frequent morphological error in the L1 corpus. More specifically, there are 181 cases of confusion of infinitive and present tense in the L1 corpus. When examining these, the error frequency seems influenced by phonological similarity (Table 5). L1 speakers produce more errors when the two verb forms are homophone (e.g., infinitive *køre* [ˈkʰøːɐ] and present tense *kører* [ˈkʰøːɐ]) than when the verb forms are heterophone (e.g., infinitive *rejse* [ˈʁɑjsə] and present tense *rejser* [ˈʁɑjˀsɐ]). This is both the case when examining the total number of errors and the error rates. For example, the error rate for using infinitive for present tense (homophone verb pairs) is 25%, i.e., out of all correct verbs in present tense (with the same pronunciation in infinitive) plus the cases where infinitive is used for a homophone present tense form, 25% are erroneous. L1 speakers also produce more errors of the type infinitive for present tense (132) than present tense for infinitive (49), i.e., they leave out an *-r* in writing. However, the error rates for the two types of confusion are both 1%, because there are more verbs in present tense in the corpus.

Based on error rates (which are entangled with phonological similarity), we expected that participants would detect more errors in the heterophone than homophone conditions. We did not expect differences between the two types of target forms (whether the target was infinitive or present tense), as there was no difference in error rates. Finally, the error rates in Table 5 also show a larger difference between the homophone and heterophone conditions when the target is present tense, compared to when the target form is infinitive. This predicts an interaction between homophony and type.

Table 6 shows the four experimental conditions for the verb errors. We used a 2 (heterophone vs. homophone) × 2 (target infinitive vs. present tense) design. Notice, that there is a visual difference between the two types of errors, because in one condition (present tense for infinitive), an extra *-r* is added, while an *-r* is missing in the other condition (infinitive for present tense). The heterophone vs. homophone verb pairs were controlled for length (number of letters in infinitive) and frequency. *T*-tests (correlated samples) showed no significant differences in length or frequency [Det Danske Sprog- og Litteraturselskab (DSL), 2022] for the homophone vs. heterophone verbs. The texts also included a minimum of 32 correct verbs (other lexemes), 8 in each condition. All stimuli can be seen in the Supplementary material (section 2).

Table 6 also shows the number and share of detected targets. In the condition present tense for infinitive, a target is considered detected if either the modal and/or the main verb is underlined.

As expected (based on error rates and phonological similarity), we found an effect of homophony ($\hat{\beta}$ = −1.21, SE = 0.09, $z = -13.38$, $p < 0.001$), so that participants detected more errors in heterophone than homophone pairs. Counter to the expectation based on error rates, we found an effect of type, so that more errors of the type infinitive for present tense were found, than for present tense for infinitive ($\hat{\beta}$ = −0.20, SE = 0.09, $z = -2.23$, $p < 0.05$). There was no interaction, contrary to the predictions based on error rates (*cf.* Table 7).

Figure 2 shows the model's predicted probability of responding correctly (i.e., detecting the error) in the different conditions. The probability of a correct answer (a detected error) is much higher in the heterophone than homophone conditions. Although, the effect of type

TABLE 6 Conditions, number of verb errors in texts and share of detected errors.

Conditions	Errors in texts (*N*)	Detected targets (*N*)	Share of detected targets (%)
HETEROPHONE PAIRS	**3,376**	**2,306**	**68.31%**
INFINITIVE FOR PRESENT TENSE: *han rejse* ['ʁajsə] he travel.INF	1,688	1,178	69.79%
PRESENT TENSE FOR INFINITIVE: *han vil rejser* ['ʁajˀsɐ] he will travel.PRS	1,688	1,128	66.82%
HOMOPHONE PAIRS	**3,376**	**1,686**	**49.94%**
INFINITIVE FOR PRESENT TENSE: *han køre* ['kʰøːɐ] he drive.INF	1,688	867	51.36%
PRESENT TENSE FOR INFINITIVE: *han vil kører* ['kʰøːɐ] he will drive.PRS	1,688	819	48.52%
Total	**6,752**	**3,992**	**59.12%**

TABLE 7 Model (2) estimates for verb errors.

Random effects	Variance	Std. dev.		
Participant (intercept)	2.7768	1.6664		
Item (intercept)	0.2204	0.4695		
Fixed effects	**Estimate**	**Std. error**	***z*-value**	***p*-value**
(Intercept)	−10.75433	2.99431	−3.592	0.000329***
Homophony	−1.20594	0.09014	−13.378	<2e-16***
Type	−0.20145	0.09025	−2.232	0.025614*
Homophony*type (Interaction)	0.01925	0.12440	0.155	0.877033
Total grammar score (quiz)	0.71993	0.17850	4.033	5.5e-05***

Dependent variable: detection (1 = error detected, 0 = error not detected). Significance codes: ***$p < 0.001$, *$p < 0.05$.

was significant, the plot shows that it is small. Also, according to the predictions based on error rates, the column with *han køre* should have been the smallest.

Finally, we found an effect of total grammar score ($\hat{\beta} = 0.72$, SE = 0.18, $z = 4.03$, $p < 0.001$), so that the higher total grammar score in the quiz, the more verb errors were detected. The grammar quiz contained 8 sentences where participants made a forced choice between infinitive or present tense for a missing verb. Out of 1.688 answers, there were only 25 errors (1.5%), made by 16 students. Twenty-two of 25 errors were in homophone verb pairs, supporting the role of phonological similarity on error production.

6.3. NP errors

In Danish, nouns are either uter (most common) or neuter gender. There are two indefinite articles, *en* (uter) and *et* (neuter) 'a.' Adjectives are inflected for gender, definiteness, and number. Typically, the suffix *-t* 'neuter,' *-e* 'definite,' or *-e* 'plural,' can be added to the uninflected basic form, corresponding to singular, indefinite, uter gender (Becker-Christensen, 2010). The most common adjective error in the L1 corpus is to leave out a suffix (*-t* or *-e*). Table 8 shows error rates for gender mismatches in adjectives and indefinite articles. Confusing the two indefinite articles is less common than missing gender agreement in adjectives, as seen in the error rates. Using uter for neuter is slightly more common than using neuter for uter.

Based on the error rates, we expected higher detection rates for mismatching articles than for mismatching adjectives, and higher detection rates for neuter for uter more than uter for neuter. The error rates in Table 8 show a slightly larger gender difference for adjectives than for articles, and we therefore predicted an interaction between word class and gender.

The four experimental conditions for the NP errors are seen in Table 9 (2 × 2 design). In continuous speech, there is phonological similarity between the correct and incorrect form in the condition mismatch with adjective, uter for neuter (where the suffix is missing). Notice, that there are also visual differences between the two word class conditions: when manipulating the adjectives, an element (*-t*) is either added or left out. When manipulating the articles, a *t* or an *n* is replaced with each other.

The neuter and uter nouns were controlled for length and frequency. The target items did not have the same syntactic function (e.g., object, subject complement or part of an adverbial) and thus were not in the same position in the sentences. The text also contained a minimum of 32 control items (16 uter NPs; 16 neuter NPs), which were inflected adjectives not already used as targets.

Table 9 shows the number and share of detected targets. Targets were considered detected if min. one of the three words in the NP was underlined.

As predicted based on error rates, we found an effect of word class ($\hat{\beta} = 0.90$, SE = 0.08, $z = 11.30$, $p < 0.001$), so that mismatches with articles were detected more than mismatches with adjectives. As expected based on error rates, we found an effect of gender ($\hat{\beta} = 0.72$, SE = 0.08, $z = 9.08$, $p < 0.001$), so that participants detected more neuter for uter than uter for neuter in general (*cf.* Table 10). We also found the expected interaction ($\hat{\beta} = -0.70$, SE = 0.11, $z = -6.23$, $p < 0.001$), which can be seen in Figure 3. It shows the model's predicted probability of responding correctly (detecting the error) in the different conditions. For the articles, the effect of gender is less pronounced than for the adjectives. The lowest detection rates were found for *et dejlig kæledyr* (mismatch with adjective; uter for neuter), as expected. However, the interaction might also be explained by the phonological similarity to the correct form in this condition, or visual differences between conditions. Perhaps, it is harder to spot a missing *-t* than an extra *-t* or to spot a *t* which is replaced with an *n*. Finally, we found an effect of total grammar score ($\hat{\beta} = 0.42$, SE = 0.12, $z = 3.38$, $p < 0.001$), so that the higher total grammar score in the quiz, the more NP errors were detected. In the grammar quiz, participants were given an adjective and asked to insert it before both an uter and a neuter noun. The article task was forced choice, and participants had to choose between uter or

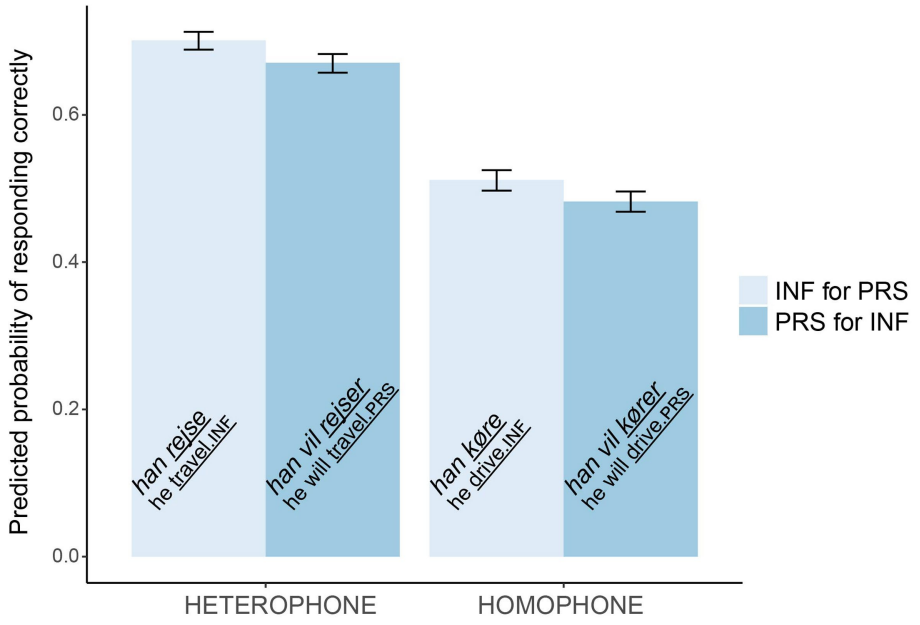


FIGURE 2
The model's predicted probabilities of detecting verb errors. Error bars show SDs.

TABLE 8 Error rates in L1 texts, gender mismatch between indefinite articles or adjectives with noun.

	N errors	N correct	Error rate (%)
Indefinite articles	**16**	**3,132**	**0.51%**
Uter for neuter (*en* for *et*)	6	984[1]	0.61%
Neuter for uter (*et* for *en*)	10	2,178	0.46%
Adjectives	**51**	**2798[2]**	**1.79%**
Uter for neuter (Ø for *-t*)	29	1,368	2.08%
Neuter for uter (*-t* for Ø)	22	1,430	1.49%

[1]Number of correct occurrences of et 'a' (neuter), found with a POS tagger [Centre for Language Technology, University of Copenhagen (CST), 2022].
[2]The number of correct adjectives with a correct -Ø or *-t* suffix. Found with a POS tagger [Centre for Language Technology, University of Copenhagen (CST), 2022]. Manually, the following were removed: adjectives with no/optional gender conjugations (ending with *-sk*, *-vis*), indeclinable adjectives (e.g., *ekstra* 'extra'), and adjectives ending with a *-t* (e.g., *stolt* 'proud').

neuter indefinite articles for four nouns. There were only 6 errors for the 844 articles (0.7%) and no errors for the 422 adjectives.

6.4. Orthographic errors

In general, we expected common types of misspellings to be noticed less than syntactic and morphological errors. In the high school corpus, orthographic errors are the most common type of error (0.86% of all words are misspelled). The 20 target items were created based on four types of misspellings which others have found to be common in L1 writing (e.g., Blom et al., 2017). Examples can be seen in Table 11. Table 11 also shows the number and shares of detected errors. Most of the errors are phonologically similar to the correct form. Some are entirely homophone (e.g., the error *virklig*), while other errors could be prosodically different, e.g., with respect to vowel length or stress.

The only significant effect of type was that reduced syllables were detected more often than missing double consonants, which were noticed the least ($\hat{\beta} = 1.40$, SE = 0.55, $z = 2.56$, $p < 0.05$). Finally, there was a significant effect of the score in the spelling task in the quiz, so that the more correct answers participants had in the spelling task, the more orthographic errors participants found in the reading task ($\hat{\beta} = 0.50$, SE = 0.08, $z = 6.52$, $p < 0.001$). In the spelling task, participants had to determine whether 8 words were spelled correctly. If not, they should write the correct form. There were 196 errors out of 1.688 answers (12% errors), made by 115 participants (1–5 errors per participant).

7. Discussion

Section 7.1 is a summary and discussion of the general findings of the study. In section 7.2, we discuss the relation between error detection rates and two seemingly dominant (and co-varying) factors in our study: the frequency of the error and its phonological similarity to the correct form. Section 7.3 discusses challenges for current and future models of eye movement control in reading and presents our recommendations based on the study.

7.1. General findings and effects of explicit grammar awareness

The present study examined the relationship between the type of errors young readers tend to overlook in texts, the type of errors these young readers produce themselves in the grammar quiz, and the type of errors that are typical of their age group in general (based on corpus error rates). When examining attention to naturally occurring grammar anomalies, some factors co-vary. Still, to use ecological stimuli is necessary if future models of language processing are to be able to accommodate naturally occurring, non-standard grammar.

10.3389/fpsyg.2023.1124227

TABLE 9 Number of NP errors in texts and share of detected errors.

Conditions	Errors in texts (*N*)	Detected targets (*N*)	Share of detected targets (%)
MISMATCH ART + N	**3,376**	**2034**	**60.25%**
Neuter for uter: *et dejlig undulat* ART.N lovely-U budgie.U 'a lovely budgie'	1,688	1,021	60.49%
Uter for neuter: *en dejlig-t kæledyr* ART.U lovely-N pet.N 'a lovely pet'	1,688	1,013	60.01%
MISMATCH ADJ + N	**3,376**	**1,685**	**49.91%**
Neuter for uter: *en dejlig-t undulat* ART.U lovely-N budgie.U 'a lovely budgie'	1,688	959	56.81%
Uter for neuter: *et dejlig kæledyr* ART.N lovely-U pet.N 'a lovely pet'	1,688	726	43.01%
Total	**6,752**	**3,719**	**55.08%**

TABLE 10 Model (3) estimates for NP errors.

Random effects	Variance	Std. dev.		
Participant (intercept)	1.260	1.1226		
Item (intercept)	0.192	0.4381		
Fixed effects	**Estimate**	**Std. error**	***z*-value**	***p*-value**
(Intercept)	−7.37728	2.07508	−3.555	0.000378***
Word class	0.90480	0.08013	11.292	<2e-16***
Gender	0.72108	0.07945	9.076	<2e-16***
Word class*Gender (interaction)	−0.70256	0.11276	−6.231	4.64e-10***
Total grammar score (quiz)	0.41706	0.12357	3.375	0.000738***

Dependent variable: detection (1 = error detected, 0 = error not detected). Significance code: ***$p < 0.001$.

In our study, grammar errors seem to attract more attention than orthographic errors. This finding is in line with Larigauderie et al. (2020) who studied attention to grammatical and orthographic errors in French. Their grammar errors were comparable to ours, as they related to number and gender agreement and misuse of the past participle form in French. Their orthographic errors (like most of ours) did not affect the phonology of the word. Previous proofreading studies of English (Hacker et al., 1994; Shafto, 2015), however, found the opposite pattern, as orthographic errors attracted more attention than grammar errors in their studies. It is likely that this discrepancy stems from differences in what is understood by a grammar error vs. an orthographical error. In Shafto (2015), the grammar errors were heterogeneous ranging from errors in verb agreement and number agreement to punctuation and capitalization errors, thus grouping types of errors which are quite distinct. The orthographic errors also included typos such as letter switches which resulted in an incorrect phonological form, and which are therefore also qualitatively different from the orthographic errors in our study. Larigauderie et al. (2020) found that typos were the most frequently detected type of error. In Hacker et al. (1994), the error categories were not clearly defined. Their grammar errors included errors in verb agreement as well as confusion of word classes (e.g., *affects* for *effects*). Altogether, these differences in the definitions of grammar vs. orthography may explain the seemingly contradictory results.

Error detection is not entirely explained by explicit grammar awareness. In the grammar quiz, the general performance was almost at ceiling with error rates ranging from 0.5% to 1.5% per task. Yet, all readers overlooked errors in the proofreading study.

Although there were generally few errors in the responses to the grammar quiz, the participants' total score in the grammar quiz did explain some of the variance in the detection rates. For the three types of grammar errors (V3 word order, verb errors, NP errors), we found an effect of the total grammar score, so that the more correct answers participants had in the three grammar tasks in the quiz, the more errors they detected. Similarly, the more correct answers participants had in the spelling task, the more orthographic errors they detected. Finally, we found that the more annoyed with language errors participants reported to be, the more errors they detected.

Unlike most previous psycholinguistic studies which either group many different types of grammar errors into one experimental condition (Hacker et al., 1994; Shafto, 2015) or only investigate one specific type as representative of all grammar errors (often using the cover term *syntactic violations*), our study distinguishes between different types of grammar errors. The descriptive statistics showed differences in detection rates between syntactic and morphological errors in our study, which seems to suggest that not all grammar errors are treated alike. Future eye-tracking studies may determine if this pattern is not just due to quantitative differences (degree of attention), but also due to qualitative differences (differences in how they are processed and attended to).

10.3389/fpsyg.2023.1124227

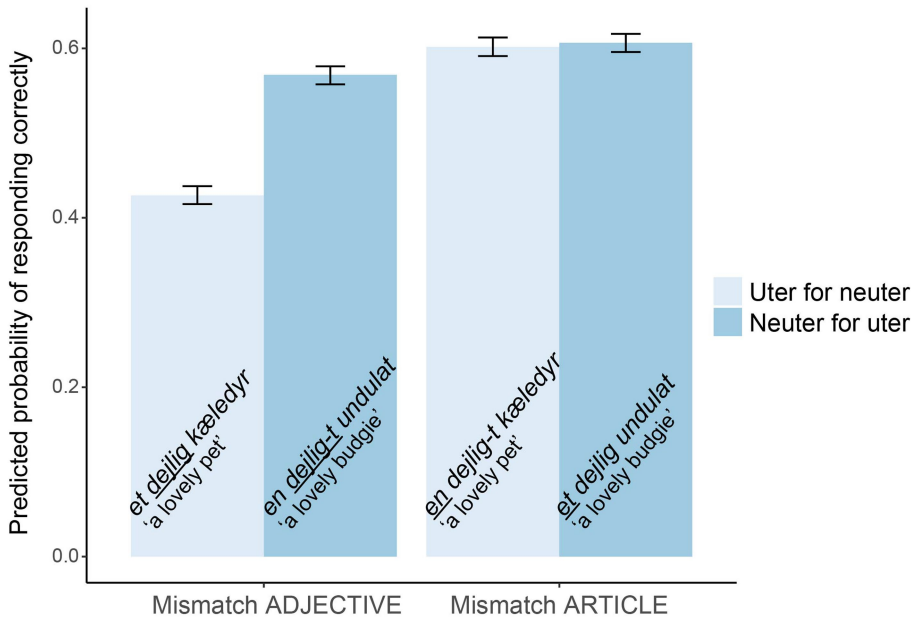


FIGURE 3
The model's predicted probabilities of detecting NP errors. Error bars show SDs.

TABLE 11 Types of orthographic errors, number of errors in texts and share of detected errors.

Types of orthographic errors (four types — five of each type)	Errors in texts (*N*)	Detected targets (*N*)	Share of detected targets (%)
Missing double consonant, e.g., *startskudet* for *startskuddet* 'the starting signal'	1,055	224	21.23%
Split compounds, e.g., *by vandring* for *byvandring* 'city walk'	1,055	342	32.42%
Missing silent letter, e.g., *siste* ['sisd̥ə]/['sisd̥] for *sidste* ['sisdə]/['sisd̥] 'last'	1,055	359	34.03%
Reduction of syllable, e.g., *virklig* ['vɪɐ̯ɣ̊li] for *virkelig* ['vɪɐ̯ɣ̊li] 'really'	1,055	456	43.22%
Total	**4,220**	**1,381**	**32.73%**

7.2. The relation between what students typically produce and what they notice

Models of natural reading processing must deal with naturally occurring errors. Yet, a complication of using naturally occurring errors is that several factors co-vary between conditions. In the following sections, we discuss two main potential contributing factors when it comes to readers' perception of and attention to grammar errors in Danish: the frequency of the error (section 7.2.1) and the phonological similarity between the error and the correct form (section 7.2.2).

7.2.1. Error frequency

Our study suggests that the frequency of grammar errors is a relevant factor to include in future models of eye movements during reading. Attention to a specific type of grammar error is not only a matter of the reader's explicit grammar awareness (as measured in the grammar quiz). If a specific type is frequent among the peers of the reader, the reader may have more exposure to this type of error and a mental representation of it. The reader may therefore find it less striking and be less likely to detect it compared to errors that are infrequent in texts written by peers. According to the descriptive statistics in our study, the error detection rates for the three overall error categories (syntactic > morphological > orthographic) were inversely proportional with the error rates in L1 writing. Syntactic errors have the lowest error rates in L1 writing and the highest detection rates. Orthographic errors have the highest error rates and the lowest detection rates. Within the three grammar subexperiments, we also found that error types with relatively high error rates (errors in homophone verb pairs, mismatching adjectives in NPs, overuse of uter in NPs) had lower detection rates than errors with lower error rates (errors in heterophone verb pairs, mismatching articles in NPs, overuse of neuter in NPs).

Yet, frequency is not the only possible explanation to these results. The higher share of detected syntactic errors could be influenced by differences in size (manipulating word order vs. letters). The homophony effect for verb errors is closely tied to the phonological similarity to the correct form (section 7.2.2.). In the subexperiment on NPs, phonological similarity to the correct form may also explain the interaction between word class and gender (section 7.2.2). Furthermore, frequency and word class co-varied. Also, the effect of word class could be influenced by differences in the placement of the error within the NP. It may be that phrase-initial errors (such as the article errors) attract more attention than errors placed in the middle of a phrase (such as the adjective errors). Thus, future studies are needed, in which effects of position in the phrase and frequency can be distinguished — and if possible, in which effects of frequency can be distinguished from phonological similarity to the correct form.

These reservations aside, it seems likely that frequency plays an important part in error detection, and that the role of frequency is worth studying in future studies with more controlled and less confounded stimuli. Frequency is, as mentioned in the introduction, tied to predictability. According to prediction-based approaches to sentence processing, unexpected input attracts attention (Kamide, 2008; Levy, 2008; Christiansen and Chater, 2016). If a reader sees input with frequent errors, the model will be updated according to the input, meaning that frequent errors should be predicted by the model, and thus should attract less attention than infrequent errors. The error rates in our study were based on texts written by high school students. We do not assume that high school students read each other's essays, but the errors they produce in school essays are likely to occur in their writing in general, including informal text directed at their peers. Furthermore, we assume that the error production patterns found in high school texts to a large extent reflect the error types found in the media.

Frequency does not explain all findings and it seems to be interacting with other factors in our study. Not all predictions based on error rates were confirmed: we did not expect an effect of type for the verb errors, but found higher detection rates for infinitive for present tense than vice versa. In the public debate and prescriptive literature, missing present tense *-r* is often accentuated as a typical or basic error (Blom and Ejstrup, 2019b), and in the study by Blom and Ejstrup (2019a), participants rated the missing present tense *-r* as the most annoying error of all included errors. This special status of the missing *-r* in present tense might explain why this error type was noticed more than the superfluous *-r* on infinitives, although the frequency in production (as measured by error rates) does not differ between the two. If looking at occurrences per 1,000 words, omitting the -r is, in fact, more frequent in written texts. Counter to our expectations, we did not find an interaction between homophony and type. The surprising result might also be explained by the great prescriptive focus on the most frequent error type (homophone; infinitive for present tense).

In our study, frequency measures were based on error rates in a small corpus of naturally occurring L1 texts. For erroneous use of gender in articles, the error rates were based on only 16 article errors, and the distribution between uter and neuter gender in errors may well be different in a larger corpus. Future studies with a larger corpus may use inferential statistics for a more adequate calculation and assessment of differences in error rates. They may also consider the pros and cons of using error rates vs. raw frequency (errors per 1,000 running words) as the basic measure. In most cases, these measures lead to the same predictions, but in one case, type for verb errors, our frequency-based predictions would have been different if we had based them on occurrences per 1,000 running words, instead of error rates. Homophony set aside, there are more errors per 1,000 words where the target form is present tense (1.02) than when it is infinitive (0.37). Thus, infinitive for present tense should be least noticed. This was, however, not the case, and this frequency measurement therefore does not seem better at predicting error detection than error rates.

To conclude, frequency (measured by error rates) in most cases predicted detection rates of different types of errors. Due to the confounded nature of the highly ecological error types in the stimuli, we cannot determine the exact nature of the interplay with other contributing factors.

7.2.2. Phonological similarity to the correct form

In naturally occurring language we often find errors that intersect grammar and phonology. Since we aimed to study error detection of naturally occurring grammar errors, our stimuli included such intersectional errors. We contrasted grammar errors where the confused forms were phonologically identical (homophone) with errors where the two forms were clearly distinct in pronunciation (heterophone). Our study showed significantly lower detection rates for verb errors in the homophone condition compared to the heterophone condition. These results suggest that phonology interferes with grammatical processing during error detection. Yet, the difference between homophone and heterophone forms may also be due to differences in frequency, as error rates in L1 writing are higher when the present tense and infinitive are homophone. In the verb error subexperiment, we therefore cannot disentangle the effect of phonological interference from that of frequency. Still, we find it plausible that phonological interference constitutes a separate effect when taking into account the findings from the subexperiment on NP errors. For NP errors, detection rates were low when the adjective was inflected in uter instead of the correct neuter form (e.g., *dejlig* instead of *dejligt*). This error with a missing *-t* is not only visually similar to the correct form (*cf.* section 7.3), but also phonologically similar. In distinct speech the final [d̥] in *dejligt* may be pronounced, but in running speech there is usually no audible difference. This similarity between forms may explain why we found an interaction between gender and word class. Frequency differences in error rates may also account for this effect. Yet, the differences in frequency are small. It therefore seems more likely that phonological similarity plays a key role in explaining the low detection rates for uter for neuter in adjectives.

Errors that intersect the boundary between grammar and phonology are not unique to Danish. "Silent suffix" errors with confusion of homophone verb forms are also frequent in other languages. In Dutch the 1st person verb *word* and the 3rd person verb *wordt* have the same pronunciation and are commonly confused (Sandra et al., 2004). In French, there is no audible difference between the verb forms *mange*, *manges* and *mangent*, and ERP studies show that responses to confusion of such homophone verb forms differs from responses to confusion of heterophone verb forms like *mange* vs. *mangez* (Carassco-Ortiz and Frenck-Mestre,

2014). This finding is in line with Larigauderie et al. (2020) who found that typographical errors (i.e., incorrect successions of letters resulting in incorrect phonology) are more frequently detected than orthographic errors which did not affect the phonology of the word. Potential interference from phonology is not limited to confusion of verb forms. The confusion of English *its* and *it's* is a prime example. Although our study cannot disentangle effects of phonological similarity from error frequency, we recommend that future eye-tracking models of reading and sentence processing models in general consider the possible role of phonological resemblance of errors to correct forms.

7.3. Challenges for current and future models of eye movement control in reading

Presumably, the error detection measure is less sensitive than eye-tracking. Although the degree of correlation between the two measures is uncertain, we assume that the overall results could be replicated using eye-tracking, which is a natural next step. More fine-grained differences may also be detected using eye-tracking, e.g., it may be that eye movements are affected, though errors are not underlined by the participant. This was, however, not found in the eye-tracking study by Huang and Staub (2021). Disruption in eye movement measures caused by transposition errors were only found in those sentences participants judged to be ungrammatical. The majority of previous eye-tracking studies of ungrammaticality did not ask participants whether they noticed and perceived the individual errors as ungrammatical or not. Using the error detection paradigm, we collected this information without interrupting participants' reading excessively and found that attention to different types of naturally occurring errors is not uniform. This variation in the reader's attention and response to errors poses a challenge to the major present models of eye movement control in reading (Reichle et al., 2003; Engbert et al., 2005). The E-Z Reader model (Reichle et al., 2009) addresses reactions to severe syntactic violations, but does not address what happens when readers encounter misspellings or other types of grammar errors. Results from previous eye-tracking studies of ungrammaticality indicate that different types of grammar errors (e.g., V3 and morphological agreement errors) elicit similar responses in participants' eye movements across languages, with similar time courses (*cf.* section 2.1) — including the very early effects, which E-Z Reader explicitly predicts for syntactic violations. If attention to different types of errors should be integrated in the E-Z Reader model, a first step could be to integrate detection of orthographic errors as part of the early familiarity check, and to account for both morphological and syntactic errors.

The E-Z Reader model does not explain why some errors are detected while others go by unnoticed, and why different readers do not always notice the same error. Also, as Warren (2011) points out, the model does not consider the precise combination of reader and the purpose or motivation for the reading. Our study both shows an effect of participants' explicit grammar awareness and general irritation with errors on detection rates.

In our study, we have demonstrated the complexity of measuring error frequency and determining when there is phonological similarity. It is therefore challenging to integrate these factors in models of eye movement control during reading. Still, the two factors are entangled, and even a rough measure of error frequency would improve current and future models when dealing with reading of everyday texts.

Previous letter detection experiments (Smith and Groat, 1979) have found position effects, e.g., that elements in the start or end of a line or within a sentence tend to be more prominent than elements in the middle. Our study on V3 errors manipulated the length of the sentence-initial adverbial, but we found no effects of the placement in the sentence (close to the start vs. further toward the middle). This lack of an effect of position was confirmed in an eye-tracking study where Norwegian readers read similar types of V3 with long and short adverbials (Søby et al., 2023). Smith and Groat (1979) did not consider different sentence structures in their analysis, only numerical order of the words, and the position effects varied between items. Further studies are needed to test the potential role of error position within the sentence.

For the verb and NP errors, there were visual differences between elements that were deleted, added and replaced with other elements. The NP data suggest that replacing two elements with another (i.e., *-t* and *-n* in indefinite articles) is noticed more than when an element is added or missing (*-t* in adjectives). However, for verb errors, a missing *-r* was more noticed than an extra *-r*. It therefore seems that other factors than visual differences are more important, e.g., word class or error frequency.

In this study, we have examined outright errors which both deviated from the norms defined by the Danish Language Council and from most participants' own answers in the grammar quiz. Language norms, however, are subject to language change and sociolinguistic variation. Natural texts therefore both contain outright errors and language anomalies in the gray zone between language errors and language variation. For instance, the inflection of Danish modal verbs seem to be subject to language change. In written production most high school students do not inflect the Danish modal verb *måtte* according to the norms defined by the Danish language council (Kristensen et al., 2023). These anomalies should also be considered in future studies.

Our study only included one type of task, i.e., proofreading while reading for comprehension. Using eye-tracking, Schotter et al. (2014a) found that the task (proofreading for letter transpositions vs. reading for comprehension) affected processing patterns. The patterns when reading for comprehension may therefore differ from what we find in our study. Still, based on our study, we recommend that future models take the following factors into account, as they may all modulate attention and eye movements:

1. Variation in the type of naturally occurring grammar errors that occur in non-standard language (e.g., syntactic errors compared to morphological errors, and different subtypes within these categories).
2. Variation in error frequencies as a general predictor, and importantly, when present: phonological similarity with the correct form (which tends to be entangled with error frequency).
3. Variation in the reader's grammatical awareness and proficiency.

Data availability statement

The dataset for this study and code for analyses can be found in an online repository: http://github.com/ResearchXX/ErrorDetection.

Ethics statement

The study involving human participants was reviewed and approved by the Faculty of Humanities' Research Ethics Committee, University of Copenhagen. Written informed consent to participate in this study was provided by participants (if above 18 years old) or by the participants' legal guardian/next of kin.

Author contributions

KS and LK contributed to designing the study. KS was responsible for making the test material and collecting data. BI wrote the code for the analyses, which KS used. The first draft was written by KS. LK and BI commented and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The study was financed by Independent Research Fund Denmark, grant number: 7023-00131B.

Acknowledgments

The authors would like to thank student assistants Julie Johanna Brink Hansen for digitizing the data and Maja Mittag for summarizing L1 corpus data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1124227/full#supplementary-material

References

Ainsworth-Darnell, K., Shulman, H. G., and Boland, J. E. (1998). Dissociating brain responses to syntactic and semantic anomalies: evidence from event-related potentials. *J. Mem. Lang.* 38, 112–130. doi: 10.1006/jmla.1997.2537

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Becker-Christensen, C. (2010). *Dansk syntaks. Indføring i dansk sætningsgrammatik og sætningsanalyse*. København: Samfundslitteratur.

Blom, J. N., and Ejstrup, M. (2019a). "Læsernes holdninger til journalisters stavefejl i de digitale medier" in *17. Møde om Udforskningen af Dansk Sprog*. eds. Y. Goldshtein, I. S. Hansen and T. T. Hougaard (Aarhus: Aarhus Universitet), 113–132.

Blom, J. N., and Ejstrup, M. (2019b). BRÆJKING NEWS: En eksperimentel undersøgelse af fejlobservante læseres holdninger til ukorrekte og korrekte stavemåder opfattet som stavefejl i nyhedsrubrikker på nettet. *NyS* 57, 10–46. doi: 10.7146/nys.v1i57.117114

Blom, J. N., Rathje, M., le Fevre Jakobsen, B., Holsting, A., Hansen, K. R., Svendsen, J. T., et al. (2017). Linguistic deviations in the written academic register of Danish university students. *OSLa* 9, 169–190. doi: 10.5617/osla.5855

Braze, D., Shankweiler, D., Ni, W., and Palumbo, L. C. (2002). Readers' eye movements distinguish anomalies of form and content. *J. Psycholinguist. Res.* 31, 25–44. doi: 10.1023/A:1014324220455

Carassco-Ortiz, H., and Frenck-Mestre, C. (2014). Phonological and orthographic cues enhance the processing of inflectional morphology. ERP evidence from L1 and L2 French. *Front. Psychol.* 5, 1–14. doi: 10.3389/fpsyg.2014.00888

Christensen, M. H., Kristensen, L. B., Vinther, N. M., and Boye, K. (2021). Grammar is background in sentence processing. *Lang. Cogn.* 13, 128–153. doi: 10.1017/langcog.2020.30

Christiansen, M. H., and Chater, N. (2016). The now-or-never bottleneck: a fundamental constraint on language. *Behav. Brain Sci.* 39, e62–e72. doi: 10.1017/S0140525X1500031X

Clifton, C., and Staub, A. (2011). "Syntactic influences on eye movements during reading" in *The Oxford handbook of eye movements*. eds. S. P. Liversedge, I. Gilchrist and S. Everling (Oxford: Oxford University Press), 896–909.

Centre for Language Technology, University of Copenhagen (CST) (2022). Online Part-Of-Speech tagger. Available at: https://cst.dk/online/pos_tagger/index.html (Accessed November 24, 2022).

Dank, M., Deutsch, A., and Bock, K. (2015). Resolving conflicts in natural and grammatical gender agreement: evidence from eye movements. *J. Psycholinguist. Res.* 44, 435–467. doi: 10.1007/s10936-014-9291-9

Deutsch, A., and Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: evidence from ERPs and eye movements. *J. Mem. Lang.* 45, 200–224. doi: 10.1006/jmla.2000.2768

Det Danske Sprog-og Litteraturselskab (DSL) (2022). KorpusDK. Available at: https://ordnet.dk/korpusdk (Accessed 24 November 2022).

Engbert, R., and Kliegl, R. (2011). "Parallel graded attention models of reading" in *The Oxford handbook of eye movements*. eds. S. P. Liversedge, I. Gilchrist and S. Everling (Oxford: Oxford University Press), 788–800.

Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during Reading. *Psychol. Rev.* 112, 777–813. doi: 10.1037/0033-295X.112.4.777

Ferreira, F., Bailey, K. G. D., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* 11, 11–15. doi: 10.1111/1467-8721.00158

Ferreira, F., and Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Lang. Linguist. Compass* 1, 71–83. doi: 10.1111/j.1749-818x.2007.00007.x

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., et al. (2017). Don't underestimate the benefits of being misunderstood. *Psychol. Sci.* 28, 703–712. doi: 10.1177/0956797617690277

Hacker, D. J., Plumb, C., Butterfield, E. C., Quathamer, D., and Heineken, E. (1994). Text revision: detection and correction of errors. *J. Educ. Psychol.* 86, 65–78. doi: 10.1037/0022-0663.86.1.65

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *J. Cogn. Neurosci.* 15, 883–899. doi: 10.1162/089892903322370807

Hahne, A., and Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Brain Res. Cogn. Brain Res.* 13, 339–356. doi: 10.1016/S0926-6410(01)00127-6

Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., and Weber, A. (2012). When one person's mistake is another's standard usage: the effect of foreign accent on syntactic processing. *J. Cogn. Neurosci.* 24, 878–887. doi: 10.1162/jocn_a_00103

Huang, K., and Staub, A. (2021). Using eye-tracking to investigate failure to notice word transpositions in reading. *Cognition* 216:104846. doi: 10.1016/j.cognition.2021.104846

Hyltenstam, K., and Abrahamsson, N. (2003). "Maturational constraints in SLA" in *The handbook of second language acquisition*. eds. C. J. Doughty and M. H. Long (Oxford: Blackwell Publishing), 538–588.

Jacoby, L. L., and Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: effects of reading incorrectly and correctly spelled words. *Can. J. Psychol.* 44, 345–358. doi: 10.1037/h0084259

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Lang. Linguist. Compass* 2, 647–670. doi: 10.1111/j.1749-818X.2008.00072.x

Konieczny, L., Scheepers, C., and Hemforth, B. (1994). "Reanalyses vs. internal repairs: non-monotonic processes in sentence perception" in *First Analysis, Reanalysis and Repair*. eds. B. Hemforth, L. Konieczny, C. Scheepers and G. Strube, vol. *8* (IIG-Berichte: University of Freiburg), 2–22.

Kristensen, L. B., Schack, J., and Søby, K. F. (2023). Om unge der har skulle bøje modalverber, men ikke har turde, ikke har kunne eller ikke har ville. *NfG* 30, 74–90. doi: 10.7146/nfg.v1i29.132901

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). Lmer test package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Larigauderie, P., Guignouard, C., and Olive, T. (2020). Proofreading by students: implications of executive and non-executive components of working memory in the detection of phonological, orthographical, and grammatical errors. *Read. Writ.* 33, 1015–1036.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Lim, J. H., and Christianson, K. (2015). Second language sensitivity to agreement errors: evidence from eye movements during comprehension and translation. *Appl. Psycholinguist.* 36, 1283–1315. doi: 10.1017/S0142716414000290

Lunsford, A. A., and Lunsford, K. J. (2008). 'Mistakes are a fact of life': a National Comparative Study. *Coll. Compos. Commun.* 59, 781–806.

Ministry of Higher Education and Science. (2022). Upper secondary education. Available at: https://ufm.dk/en/education/the-danish-education-system/upper-secondary-education (Accessed 24 November 2022).

Ni, W., Fodor, J. D., Crain, S., and Shankweiler, D. (1998). Anomaly detection: eye-movement patterns. *J. Psycholinguist. Res.* 27, 515–539. doi: 10.1023/A:1024996828734

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456. doi: 10.1006/jmla.1999.2653

Quist, P. (2008). Sociolinguistic approaches to multiethnolect: language variety and stylistic practice. *Int. J. Biling.* 12, 43–61. doi: 10.1177/13670069080120010401

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reichle, E. D. (2011). "Serial-attention models of reading" in *The Oxford handbook of eye movements*. eds. S. P. Liversedge, I. Gilchrist and S. Everling (Oxford: Oxford University Press), 767–786.

Reichle, E. D., Rayner, K., and Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: comparisons to other models. *Behav. Brain Sci.* 26, 445–476. doi: 10.1017/s0140525x03000104

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

Sandra, D., Frisson, S., and Daems, F. (2004). Still errors after all those years ...: Limited attentional resources and homophone frequency account for spelling errors on silent verb suffixes in Dutch. *Writ. Lang. Lit.* 7, 61–77. doi: 10.1075/wll.7.1.07san

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., and Rayner, K. (2014a). Task effects reveal cognitive flexibility responding to frequency and predictability: evidence from eye movements in reading and proofreading. *Cognition* 131, 1–27. doi: 10.1016/j.cognition.2013.11.018

Schotter, E. R., Reichle, E. D., and Rayner, K. (2014b). Rethinking parafoveal processing in reading: serial-attention models can explain semantic preview benefit and N + 2 preview effects. *Vis. Cogn.* 22, 309–333. doi: 10.1080/13506285.2013.873508

Shafto, M. A. (2015). Proofreading in young and older adults: the effect of error category and comprehension difficulty. *Int. J. Environ. Res. Public Health* 12, 14445–14460. doi: 10.3390/ijerph121114445

Smith, P. T., and Groat, A. (1979). "Spelling patterns, letter cancellation and the processing of text. Processing of visible language" In *Proceedings of the first conference on processing of visible language, September 5–8, 1977*, ed. P. A. Koolers, M. E. Wrolstad, and H. Bouma (New York: Plenum), 309–324.

Søby, K. F., and Kristensen, L. B. (2019). Hjælp! Jeg har mistede min yndlings rød taske. Et studie af grammatikafvigelser. *NfG* 26, 89–104. doi: 10.7146/nfg.v0i26.115995

Søby, K. F., and Kristensen, L. B. (to appear). V2 is not difficult to all learners in all contexts - a cross-sectional study of L2 Danish.

Søby, K. F., Milburn, E., Kristensen, L. B., Vulchanov, V., and Vulchanova, M. (2023). In the native speaker's eye: Online processing of anomalous learner syntax. *Appl. Psycholinguist* 44, 1–28. doi: 10.1017/S0142716422000418

Sturt, P., Sanford, A. J., Stewart, A., and Dawydiak, E. (2004). Linguistic focus and good-enough representations: an application of the change-detection paradigm. *Psychon. Bull. Rev.* 11, 882–888. doi: 10.3758/BF03196716

Veldre, A., and Andrews, S. (2018). Beyond cloze probability: Parafoveal processing of semantic and syntactic information during Reading. *J. Mem. Lang.* 100, 1–17. doi: 10.1016/j.jml.2017.12.002

Vinther, N. M., Boye, K., and Kristensen, L. B. (2015). Grammatikken i baggrunden – opmærksomhed under læsning. *NyS* 47, 99–139. doi: 10.7146/nys.v47i47.19877

Warren, T. (2011). "The influence of implausibility and anomaly on eye movements during reading" in *The Oxford handbook of eye movements*. eds. S. P. Liversedge, I. Gilchrist and S. Everling (Oxford: Oxford University Press), 912–923.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 13 September 2023
DOI 10.3389/fpsyg.2023.1212701

OPEN ACCESS

EDITED BY
Marijan Palmovic,
University of Zagreb, Croatia

REVIEWED BY
Ranko Matasović,
University of Zagreb, Croatia
Maria Yuryevna Lebedeva,
Pushkin State Russian Language Institute, Russia

*CORRESPONDENCE
Nina Zdorova
✉ nzdorova@hse.ru

RECEIVED 26 April 2023
ACCEPTED 23 August 2023
PUBLISHED 13 September 2023

CITATION
Zdorova N, Parshina O, Ogly B, Bagirokova I, Krasikova E, Ziubanova A, Unarokova Sh, Makerova S and Dragoy O (2023) Eye movement corpora in Adyghe and Russian: an eye-tracking study of sentence reading in bilinguals.
Front. Psychol. 14:1212701.
doi: 10.3389/fpsyg.2023.1212701

Eye movement corpora in Adyghe and Russian: an eye-tracking study of sentence reading in bilinguals

Nina Zdorova[1,2]*, Olga Parshina[1,3], Bela Ogly[1], Irina Bagirokova[2,4], Ekaterina Krasikova[1], Anastasiia Ziubanova[1], Shamset Unarokova[5], Susanna Makerova[5] and Olga Dragoy[1,2]

[1]Center for Language and Brain, HSE University, Moscow, Russia, [2]Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia, [3]Department of Psychology, Middlebury College, Middlebury, VT, United States, [4]School of Linguistics, HSE University, Moscow, Russia, [5]Laboratory of Experimental Linguistics, Adyghe State University, Maykop, Russia

The present study expands the eye-tracking-while reading research toward less studied languages of different typological classes (polysynthetic Adyghe vs. synthetic Russian) that use a Cyrillic script. In the corpus reading data from the two languages, we confirmed the widely studied effects of word frequency and word length on eye movements in Adyghe-Russian bilingual individuals for both languages. We also confirmed morphological effects in Adyghe reading (part-of-speech class and the number of lexical affixes) that were previously shown in some morphologically-rich languages. Importantly, we demonstrated that bilinguals' reading in Adyghe does differ quantitatively (the effect of language on reading times) and qualitatively (different effects of landing and previous/upcoming words on the eye movements within a current word) from their reading in Russian.

KEYWORDS

eye movement benchmarks, cross-linguistic study, universal patterns of reading, minority language, polysynthetic language, West Circassian

Introduction

Recent eye-tracking studies have been specifically investigating universal patterns of reading across languages in monolingual (English in Cop et al., 2017; 13 languages in Siegelman et al., 2022) and bilingual individuals (Dutch-English in Cop et al., 2015; Chinese-English in Sui et al., 2022) within a corpus-based approach and traditional experimental paradigm [see a comparative study of reading in English, Finnish, and Chinese by Liversedge et al., 2016]. Whereas previous research has been done on major languages of language families, and on bilingual pairs using contrasted orthographies and morphological structures, the present study expands the eye-tracking-while-reading research toward less studied languages of different typological classes (polysynthetic Adyghe vs. synthetic Russian) with the same (Cyrillic) script.

Decades of eye-tracking research have already established psycholinguistic features that affect readers' eye movements and, consequently, their language processing. The most robust lexical effects on eye-movements (i.e., the ones shown consistently across a range of empirical studies) are imposed by word frequency, word length and word predictability (Inhoff and Rayner, 1986; Rayner, 1998; Staub and Rayner, 2007). They were shown to affect both fixation durations and probabilities of skipping, i.e., the probability of a word being skipped and not

fixated during the first-pass reading. Frequent words (Schilling et al., 1998), shorter words (Inhoff and Radach, 1998), and contextually more predictable words (Rayner and Well, 1996) are skipped more often and fixated for a shorter time (Clifton et al., 2007).

Apart from lexical features, morphological and morphosyntactic ones also affect eye movements across languages. For instance, verbs were shown to be read significantly slower than nouns in both Russian-speaking adults (Laurinavichyute et al., 2019) and children (Lopukhina et al., 2022). The latter also showed a difference in skipping rate based on the part-of-speech (POS) with verbs being less likely to be skipped than nouns (Lopukhina et al., 2022). Comparing two bigger groups of word classes (content words vs. function words) in English, Schmauder et al. (2000) reported that function words had a longer total reading time and were reread more frequently than content words. The study also found no evidence for a higher skipping rate on function words when frequency and length were controlled for.

To embrace a holistic approach to language reading with a range of linguistic features taken into account, reading corpora, also known as corpora of eye movements, have become a productive tool in the last decade [see The Multilingual Eye-tracking Corpus of eye movements while reading texts (MECO, Siegelman et al., 2022); Potsdam Sentence Corpus (Kliegl et al., 2004); Ghent corpus of bilingual text reading (Sui et al., 2022); Russian Sentence Corpus (RSC, Laurinavichyute et al., 2019); The child version of the Russian Sentence Corpus (ChiRSC, Lopukhina et al., 2022); The Bilingual Russian Sentence Corpus (BiRSC, Parshina et al., 2021) etc.]. Importantly, the corpora enable us to establish the basic characteristics of eye movements (eye movement benchmarks) and compare them across languages.

Crucially, disregarding the core idea of universality and language specificity that imply linguistic diversity as a necessary prerequisite, the languages in eye-tracking studies (incl. Eye-tracking corpora studies) are, so far, mostly Indo-European languages and the biggest representatives of Uralic, Sino-Tibetian, and Turkic language families, like Finnish, Chinese, Turkish etc. Moreover, the emphasis of the cross-linguistic comparison is primarily based on the differences in orthographies and scripts (English, Chinese, and Finnish in Liversedge et al., 2016; English and Russian in Parshina et al., 2021). Hence, a diversity in reading corpora, a focus shift on morphologically-driven cross-linguistic comparison of eye movements, and a greater attention to smaller representatives of language families, like minority languages is proposed.

The present study covers the eye movement benchmarks while reading in a polysynthetic minority language, Adyghe (also known as Adyghe),[1] which has not been done before. Adyghe is one of the West Caucasian languages spoken in Russia and some Middle East countries. It is an SOV language spoken predominantly in southern Russia, by 81,294 people with 75,793 people using it on an everyday basis (according to the Russian Population Census, 2020).[2] Adyghe uses the Cyrillic script, but includes some language-specific letters, and its orthography is opaque – i.e., the letter-phoneme correspondence is inconsistent and not transparent (Daniel and Lander, 2011; Polinsky, 2020). Adyghe includes the Bzhedugh, Shapsugh, Abadzekh, and Temirgoy dialects (Polinsky, 2020), where the latter is considered the standard variety.

As all Adyghe speakers are also Russian speakers (Polinsky, 2020), their reading data in both languages were collected and compared in within-language and within-group analyses. Russian is a Slavic synthetic SVO language with some analytic trends. It is based on a Cyrillic script, and its phoneme-letter correspondence allocates it to the language with medium-shallow orthography (Rakhlin et al., 2017; Zhukova and Grigorenko, 2019). The last years have seen a major growth in psycholinguistic studies of Russian, including three corpus studies of eye movements in different Russian-speaking populations (Laurinavichyute et al., 2019; Parshina et al., 2021; Lopukhina et al., 2022), that established eye movements benchmarks in Russian (summarized in Table 1) and described the contribution of lexical and morphosyntactic features into reading in Russian.

To sum up, while there is some knowledge about the lexical, morphological, and morphosyntactic effects on eye movements during reading in different populations, languages, and orthographies, little is still known about the reading behavior of bilinguals in understudied, typologically different languages that use the same script. The goals of the study were, therefore, twofold. First, we aimed to establish benchmarks in eye movements while bilinguals were reading in a polysynthetic language (Adyghe) and report the psycholinguistic features that affect eye movements while reading in it. Second, we aimed to explore the differences between reading in two morphologically different languages (polysynthetic Adyghe vs. synthetic Russian), which are both Cyrillic-based.

A within-group comparison of reading bilinguals' data in two languages enabled us to disentangle the effect of language, *per se*, and to shift from a common comparison of bilinguals with monolingual controls (Rothman et al., 2022). However, the discussion of the findings does rely on a meta-comparison with other Russian-speaking groups, like monolinguals (Laurinavichyute et al., 2019), Russian heritage speakers (HSs), and L2 learners of Russian (Parshina et al., 2021).

Materials and methods

Participants

Sixty five bilingual adult speakers of Russian and Adyghe took part in the study (57 women; Mean age = 30.2, SD = 13.5, range 18–60). The mean education level among participants was 14.9 years, SD = 2.3, range 11–20. All participants were recruited in Maykop, the capital of the Republic of Adygea, and they were primarily students of the Adygea State University ($N = 32$). The recruitment unfolded in 2 years: as a first stage of the study in 2021 and the final stage of data collection in 2022.

Whereas the majority indicated both Adyghe and Russian as their mother languages, 23 participants considered Adyghe as their only mother tongue, with Russian as their second language. At the same time, most participants' family languages (i.e., languages spoken by their parents) were, again, both Adyghe and Russian ($N = 57$).

1 In this paper, we stick to the term Adyghe that is widely used in typological literature, including The World Atlas of Language Structures (WALS, Dryer and Haspelmath, 2013).

2 https://eng.rosstat.gov.ru/

TABLE 1 Descriptive statistics of language use according to the shortened LEAP-Q form.

	Adyghe	Russian
Age of reading acquisition onset, years, Mean (SD)	7.2 (2.6)	5.9 (0.9)
Reading skill score, on scale 1 to 5 with 5 as the highest, Mean (SD)	4.0 (0.8)	4.8 (0.5)
Language use per day, %	58.6	41.4
Reading exposure per day, % of participants		
Almost none	4	0
<1 h	64	4
1–2 h	22	16
2–3 h	6	26
3–4 h	2	36
>4 h	2	18
Preferred language to read a text for pleasure, % of participants	16	84

Fourteen participants indicated speaking more than one Adyghe dialect. In this case, we asked them to specify the one mostly used, rather preferred, and/or spoken in the family, which we considered as the dominant dialect. Hence, the distribution of Adyghe dominant dialects among the participants was as follows: Bzhedugh ($N = 22$), Kabardian ($N = 12$), Temirgoy ($N = 11$), and Abadzekh ($N = 5$).

It should be noted that Kabardian is treated among linguists as another Circassic language, – East Circassian (Daniel and Lander, 2011; Polinsky, 2020). At the same time, due to the great proximity and similarity to Adyghe, Kabardian speakers living in Maykop tend to identify themselves as Adyghe speakers of Kabardian variety, and point out that their Kabardian variety differs from the Kabardian language in the Republic of Kabardino-Balkaria. Hence, Kabardian participants were originally included in the present study, based on their self-identification, and on the language of their primary reading exposure – Temirgoy dialect in their former school education, with the latter being especially relevant for a reading study.

To ensure the homogeneity of the data sample, we checked for the differences in reading comprehension accuracy among the speakers of the four dialects. The Kruskal Wallis test (applied due to the non-normal distribution of residuals) showed that comprehension accuracy across the four dialects was different (chi-squared = 173.19, df = 2, $p < 0.01$). A post-hoc pairwise comparison of accuracies with a Dunn test confirmed that mean accuracies of Kabardian speakers differed significantly from speakers of Bzhedugh (adjusted $p < 0.05$) who represented the great majority of the population in Maykop, and of our dataset.

Based on the accuracy data differences, Kabardian speakers ($N = 15$) were excluded from further analysis. The final sample consisted, therefore, of 50 participants (44 women; Mean age = 32.7, SD = 14.1, range 18–60). The mean education level among participants was 15.1 years, SD = 2.1, range 11–20. We summarized the self-reported information about participants' reading acquisition, reading skills, and reading exposure in both languages, from the shortened version of the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian et al., 2007) in Table 1.

All participants had normal or corrected to normal vision. They all signed an informed consent form, and their participation was voluntary. The study was approved by the HSE Committee on Interuniversity Surveys and Ethical Assessment of Empirical Research.

Materials and design

The materials of the study consisted of two corpora of sentences: The Russian Sentence Corpus (RSC, Laurinavichyute et al., 2019) and The Adyghe Sentence Corpus (ASC), which was compiled in an analogous way to the RSC. The first version of the ASC in 2021 included 60 sentences with word annotation, whereas 40 more sentences and target words for a more controlled study design were added later in 2022. Hence, the full version of the ASC included 100 sentences of different syntactic structures typical for Adyghe. Similarly to the RSC, all words in ASC were annotated for parts of speech, word frequency (retrieved from Adyghe Corpus),[3] and word length. Apart from that, the ASC included morpheme annotation (the number of morphemes, and number of roots, number of grammatical and lexical affixes). The parts of speech annotation was performed according to the function of a word in a sentence instead of its actual belonging to a word class and contained bigger classes of words like VERB for all verb-based words including participles, or FUNCTION for all non-content words like prepositions, conjunctions, etc. The distribution of parts of speech in the ASC was as follows: Nouns 38.4%, Verbs 32%, Pronouns 6.7%, Function words 2.7%, Adjectives 8.6%, Adverbs 11.5%.

To enable an experimental design and control data analysis for frequency, word length, and parts-of-speech class, the Adyghe Sentence corpus included target words in eight conditions. The 2 × 2 × 2 design consisted of two parts-of-speech classes (nouns and verbs), two word length classes (short words of 1–7 characters and long words of 8–19 characters), and two word form frequency classes (low frequency < 10 items per million (ipm), high frequency > 20 ipm). Each condition was represented with eight target words in the middle of a sentence (i.e., not the first or the last word), resulting in 64 sentences with a target word. The description of both sentence corpora used in the study is provided in Table 2.

A comprehension question with multiple answer options followed 33% of Russian sentences and 40% of Adyghe sentences. An example of sentences with a question in both languages is provided in Table 3.

Apparatus

Eye movements were recorded using an eye-tracking system EyeLink Portable Duo (SR Research, Canada), with sampling rate of 1,000 Hz. The stimuli were displayed in black Ubuntu Mono font, font size 30 pt., on a light-gray background of the ASUS ROG Zephyrus S GX701GV-EV006 laptop with 1920×1080 screen resolution and

3 http://adyghe.web-corpora.net/

TABLE 2 Descriptive statistics of the two corpora: The Russian Sentence Corpus (taken from Laurinavichyute et al., 2019) and The Adyghe Sentence Corpus.

	The Russian Sentence Corpus (Laurinavichyute et al., 2019)	The Adyghe Sentence Corpus (This study)
Total number of sentences	144	100
Sentence length (in words)	Mean = 9 SD = 1.4 Range: 5–13	Mean = 6.7 SD = 1.8 Range: 2–11
Number of words	1,362 words 1,074 (without first and last words)	625 words 425 (without first and last words)
Word length (in characters)	Mean = 5.7 SD = 3 Range: 1–16	Mean = 7.5 SD = 3.8 Range: 1–28
Word form frequencies (item per million - ipm)	Class 1 (1–10 ipm) – 404 Class 2 (11–100 ipm) – 340 Class 3 (101–1,000) – 192 Class 4 (1,001 – 10,000) – 151 Class 5 (10,001 – max) – 131	Class 1 (1–10 ipm) – 240 Class 2 (11–100 ipm) – 153 Class 3 (101–1,000) – 128 Class 4 (1,001 – 10,000) – 64 Class 5 (10,001 – max) – 27 NA - 13

144 Hz refresh rate. Participants were seated 52 cm from the screen, and 36.5 cm from the camera with their head positioned on a chin rest. Only the right eye was recorded.

Procedure

After signing a consent form, participants filled in a short questionnaire with their demographic data and their language background. Then, they proceeded with an eye-tracking part of the study. Participants from 2021 read the first version of the ASC with 60 sentences, whereas participants from 2022 read both corpora, in Russian and in Adyghe, in their final versions (i.e., 144 and 100 sentences respectively). In their case, the sequence of corpora presentations was counterbalanced. Participants were given both an oral and a written instruction about the experiment's procedure. The eye-tracking-while-reading task started with a 9-point calibration (with an average error < = 0.5 and a maximum error < = 1.0), continued with the instruction for the experiment on the screen and was followed with the practice trials (5 in the RSC and 3 in ASC). Each trial started with a drift correction point on the position of the first letter in the first word of the sentence. If no fixation was detected within 500 msec, a recalibration was performed. Once a drift correction was successful, a sentence appeared in the middle of the screen. Participants were instructed to read sentences silently at their normal pace, and fixate on a red point in the right lower corner of the screen once they finished reading a sentence (see Figure 1 picturing how a trial was unfolding). After that, either a comprehension question, or a new trial appeared.

TABLE 3 Examples of stimuli in Russian (from RSC) and Adyghe (from ASC).

Stimuli	
Russian	
Sentence	*Взяв с собой фотоаппарат, вся семья поехала в парк на пикник.*
English translation	Taking a camera with them, the whole family went to a picnic in a park.
Glossing	Vsya-v s soboy fotoapparat, vsya semya poexa-l-a v park na piknik. Taking with themselves camera, the whole family went to park to a picnic.
Question	*Куда поехала семья на пикник?* Where did the family go?
Answer options	*В парк* *В лес* *В сад* To the park, to the forest, to the garden
Correct answer	*В парк* To the park
Adyghe	
Sentence	*ЧыжьэкIэ, псыхъом ушъхьэдэплъымэ, ордэ унашъхьэр къэлъагъощтыгъ.* far away river if you look across big rooftop could be seen
English translation	Far away, across the river, the roofs of an ancient castle could be seen.
Glossing	č-ẑe-ç'je psəxo-m wə-ŝx^{h}ə-də-pʃə-m-ə ordə wənə-šx^{h}əxə-r č'ə-ʎa-ʁoš'-təʁ
Question	*Сыда къэлъагъощтыгъэр?* What could be seen?
Answer options	*Ордэ унашъхь* *Ежь замокыр* The rooftop of an ancient castle, An ancient castle itself
Correct answer	*Ордэ унашъхь* The rooftop of an ancient castle

Participants answered with a mouse click, choosing from the options presented. While reading one corpus, short breaks of 1–3 min were introduced, whereas a longer, up to 15 min break, was held between the two corpora. A re-calibration was performed after each break. Experimental procedure with.

Data analysis

Statistical analysis was performed in R (R Core Team, 2020). Analysis of eye movements predominantly followed the protocol in Kliegl et al. (2006) and Laurinavichyute et al. (2019). Thus, the first and the last words in each sentence were removed. First fixation durations shorter than 60 ms were excluded from the analysis, as they were not likely to reflect lexical processing yet (see Sereno and Rayner, 2003). No upper cut-off limits were applied. The following

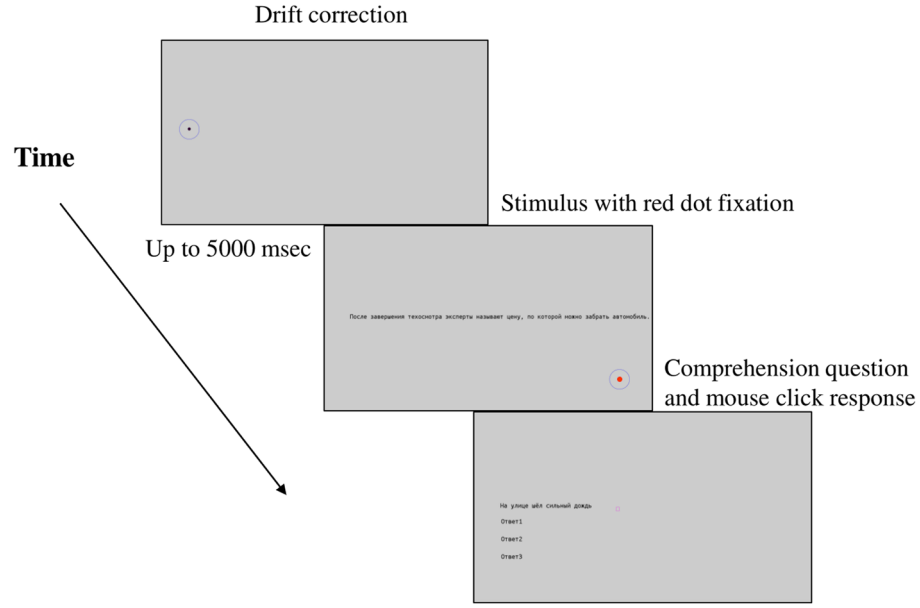


FIGURE 1
Experimental trial unfolding.

9 measurements of eye movements were chosen as dependent variables:

i. first fixation duration (FFD);
ii. single fixation duration (SFD);
iii. Gaze duration (GD);
iv. total reading time (TT);
v. probability of skipping the word (P0);
vi. probability of fixating the word only once (P1);
vii. Probability of fixating the word more than once (P2+);
viii. Probability of the word being an origin of a regressive saccade (RO);
ix. probability of the word being a goal of a regressive saccade (RG).

The listed measurements reflect both, early (FFD, SFD, GD, P0, P1) and late language processing (TT, P2+, RO, RG) - even though the same cognitive processes might overlap in different eye-movement measures (Holmqvist et al., 2011), early measures tend to be primarily associated with lexical activation, early information integration, and early morphological decomposition (Holmqvist et al., 2011, p. 385; Vasishth et al., 2013), while late measures reflect rather post-lexical processing including syntactic integration, and reanalysis (Boston et al., 2008; Holmqvist et al., 2011).

Continuous eye-movement outcome measures (FFD, SFD, GD, TT) were log-transformed and were fit with separate linear mixed-effects models. Binary variables (P0, P1, P2+, RO, RG) were fit with separate generalized linear mixed-effects models. Random effects for both model types included participants' id, sentence number, and words. For modeling, lme4 package, version 1.1–31 (Bates et al., 2015) was used. Significant effects were adjusted for multiple comparisons with Bonferroni correction. Tables with the models' output were created with sjPlot package, version 2.18.12 (Lüdecke, 2017), and are provided in the Supplementary materials. Figures were plotted with ggplot2 package, version 3.4.0 (Wickham, 2016).

The full list of independent variables was as follows:

a. word frequency
b. word length
c. part-of-speech class (POS)
d. word frequency of a previous word
e. word frequency of a next word
f. word length of a previous word
g. word length of a next word
h. word's relative position in a sentence
i. landing position (how far from the word beginning the first fixation landed)
j. number of lexical affixes (for ASC only)
k. self-reported reading skill score in Adyghe (for ASC only)

Following Laurinavichyute et al. (2019), all word frequencies were log-transformed, word length was centered, but not scaled. POS was a factor variable with 6 levels [VERB, NOUN, ADJ(ective), ADV(erb), PRONOUN, FUNCTION], with verbs being the basis for comparison. The number of lexical affixes, as well as the reading skill score were centered, but not scaled.

The data analysis of eye movements included two parts in line with the aims of the study. First, to establish benchmarks in eye movements while reading in Adyghe and report psycholinguistic features that affect reading in this language, we performed analysis of eye movements in ASC. This analysis included subparts of all-word analysis in the final sample of 50 participants, and target-word analysis in 38 participants from 2022 data collection (when target words were introduced in the final version of the ASC). Second, to disentangle the effect of language *per se* on reading in bilinguals with two

morphologically different languages, we conducted a within-group analysis of eye movements on all words while reading two corpora: ASC and RSC ($N = 38$).

Thus, taking into account the different linear models depending on the eye-tracking measure in focus, and on the analysis type (all-word vs. target-word), there were several model structures. The full structure of the models for continuous eye-tracking measures in all-word analysis was as follows: *continuous eye-tracking measure ~ reading skill in Adyghe + word frequency + word length + next word's length + next word's frequency + previous word's length + previous word's frequency + word's relative position + POS + number of lexical affixes + landing + (1 | participant) + (1 | sentence number) + (1 | word).* The full structure of the models for binary eye-tracking measures in all-word analysis was as follows: *binary eye-tracking measure ~ reading skill in Adyghe + word frequency + word length + number of lexical affixes + (1 | participant) + (1 | sentence number) + (1 | word).*

The full structure of the models in target-word analysis (for both continuous and binary eye-tracking measures) was shortened to the controlled independent variables only: *continuous/binary eye-tracking measure ~ word frequency + word length + POS + (1 | participant) + (1 | sentence number).*

The full structure of the models for continuous eye-tracking measures in within-group analysis was as follows: *continuous eye-tracking measure ~ lang*(reading skill in Adyghe + reading skill in Adyghe + word frequency + word length + next word's length + next word's frequency + previous word's length + previous word's frequency + word's relative position + landing) + (1 | participant) + (1 | sentence number) + (1 | word).* The full structure of the models for binary eye-tracking measures in within-group analysis was shortened to the very basic word features only: *binary eye-tracking measure ~ lang*(word frequency + word length) + (1 | participant) + (1 | sentence number) + (1 | word).*The code is freely available at Open Science Framework (OSF) platform, DOI 10.17605/OSF.IO/5UR8D.[4]

Results

All model outputs with significant effects reported in this section (i.e., after Bonferroni correction) are provided in Supplementary Tables S1–S6.

The benchmarks of eye movements in reading in Adyghe

The descriptive measures are summarized in Table 4 below.

TABLE 4 Descriptive statistics of eye-movements in reading ASC, Mean (SD).

Measure	Measurement	
FFD	msec	282.5 (48.12)
SFD		308 (50.4)
GD		662 (194)
TT		956 (301)
P0	%	1 (0.01)
P1		18 (0.08)
P2+		80 (0.08)
RO		23 (0.1)
RG		17 (0.08)
Fixation count	N	3.74 (1.01)
Landing position	%	31 (0.08)

Word frequency

A significant effect of a word form frequency was observed across all basic fixation duration measures: in FFD (Est. = −0.01, SE = 0.00, $t = -3.90$, $p = 0.002$), in SFD (Est. = −0.02, SE = 0.01, $t = -3.47$, $p = 0.08$), in GD (Est. = −0.03, SE = 0.00, $t = -6.01$, $p < 0.001$), and in TT (Est. = −0.03, SE = 0.01, t = −6.27, $p < 0.001$). The direction of the effect was as expected: the fixation duration decreased with a higher word form frequency as illustrated in Figure 2. More frequent words were significantly more likely to be fixated only once (P1: Log odds = 0.06, SE = 0.02, $t = 3.20$, $p = 0.007$), and were less likely to be fixated two or more times (P2+: Log odds = −0.07, SE = 0.02, $t = -3.93$, $p < 0.001$). Additionally, the probability of a word being a goal of regression decreased with higher word frequency (RG: Log odds = −0.05, SE = 0.02, $t = -3.10$, $p = 0.01$). In target word analysis, the more frequent words elicited longer fixation durations in TT only (Est. = −0.03, SE = 0.01, $t = -3.17$, $p = 0.006$).

Word length

Longer words significantly increased GD (Est. = 0.10, SE = 0.00, $t = 22.53$, $p < 0.001$) and TT (Est. = 0.10, SE = 0.00, $t = 20.66$, $p < 0.001$) - see Figure 3. Longer words were shown to be less likely skipped (P0: Est. = −0.34, SE = 0.03, $t = -10.38$, $p < 0.001$) or fixated once only (P1: Est. = −0.50, SE = 0.02, $t = -23.21$, $p < 0.001$), whereas they were highly likely to be fixated more than twice (P2+: Est. = 0.54, SE = 0.02, $t = 24.98$, $p < 0.001$). Longer words were also significantly more likely to be a goal of regression (RG: Est. = −0.06, SE = 0.01, $t = -4.75$, $p < 0.001$).

In target word analysis, the longer words elicited longer fixation durations in GD (Est. = 0.10, SE = 0.01, $t = 14.91$, $p < 0.001$) and TT (Est. = 0.10, SE = 0.01, $t = 16.22$, $p < 0.001$). The effects of fixation probabilities remained stable: in P0 (Est. = −0.78, SE = 0.13, $t = -5.98$, $p < 0.001$), in P1 (Est. = −0.57, SE = 0.05, $t = -12.33$, $p < 0.001$), in P2+ (Est. = 0.60, SE = 0.05, $t = 13.14$, $p < 0.001$), and in RG (Est. = −0.08, SE = 0.03, $t = -2.68$, $p = 0.029$).

Morphological features: POS class and the number of lexical affixes

Nouns were read significantly faster than verbs (TT: Est. = −0.13, SE = 0.03, $t = -4.20$, $p < 0.001$). However, other POS did not differ significantly from verb reading. Moreover, the target-word analysis with verbs and nouns did not show significant effects of POS either. Figure 4 shows predicted values of total reading times across all parts of speech. The number of lexical affixes significantly increased TT (Est. = 0.20, SE = 0.06, $t = 3.17$, $p = 0.025$).

4 https://osf.io/5ur8d/?view_only=432e327cd0e64b5ca062be7e7e56b9b3

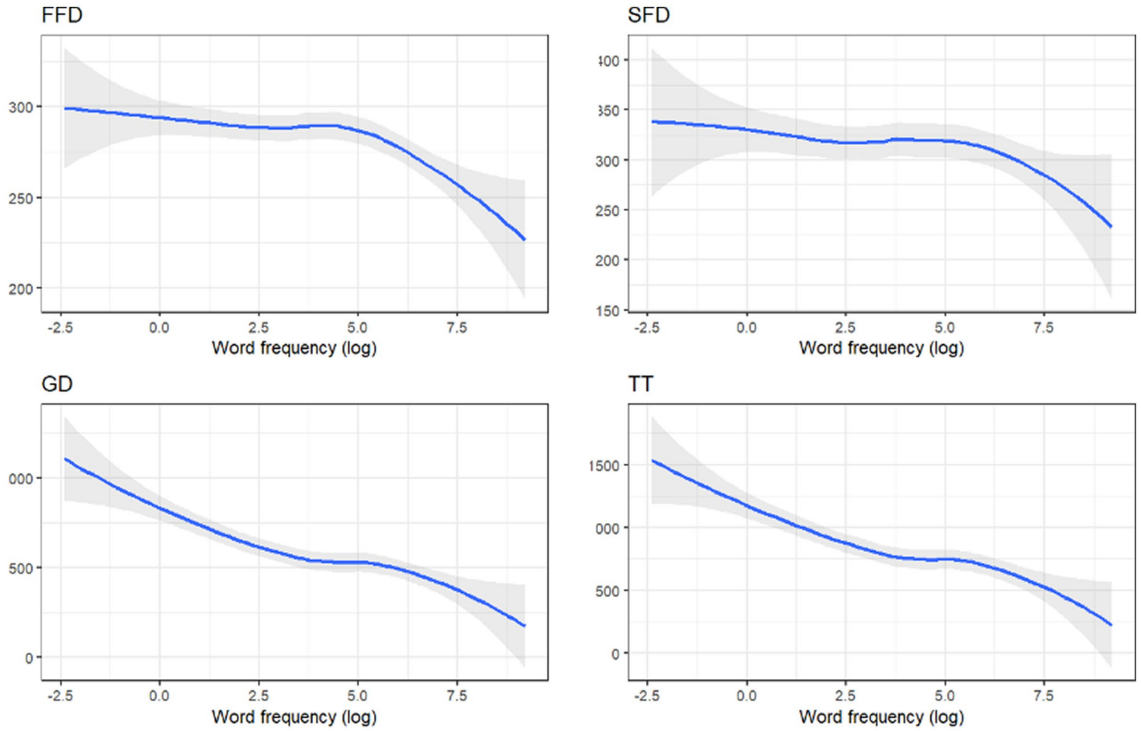


FIGURE 2
Estimated fixation durations depending on the word frequency.

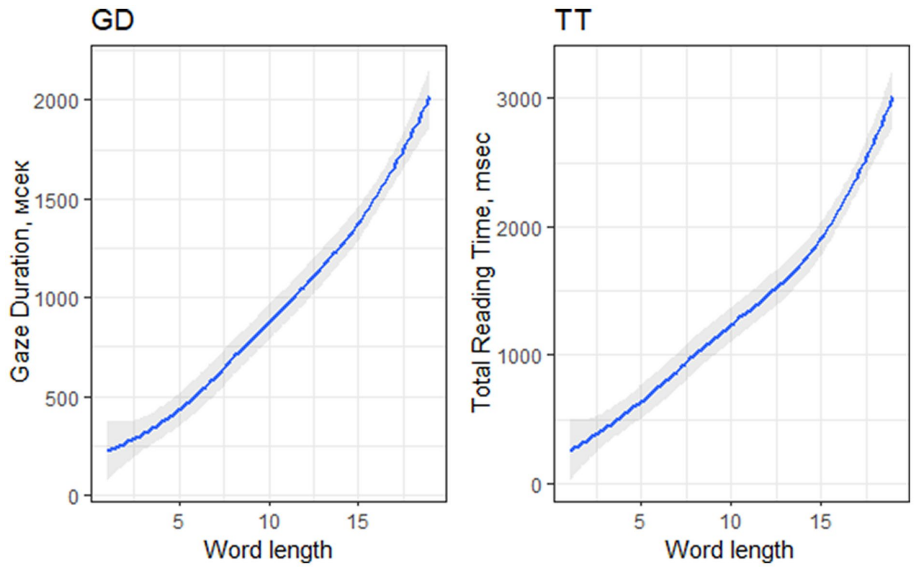


FIGURE 3
Word length effect on GD and TT in all-word analysis.

Word properties of previous and next words

Word length but not word frequency of a previous word significantly affected TT of reading a current word (Est. = −0.02, SE = 0.00, $t = -3.92$, $p = 0.001$). In turn, neither word length, nor word frequency of a next word affected eye movements while reading the current word.

Relative position and landing

Words in the middle and closer-to-final positions were first fixated longer (seen in FFD increase: Est. = 0.08, SE = 0.03, $t = 3.13$, $p = 0.028$), but they were read significantly faster in total reading time than word in the initial positions (TT: Est. = −0.28, SE = 0.06, $t = -4.98$, $p < 0.001$ – see Figure 5). Landing position further from the word beginning

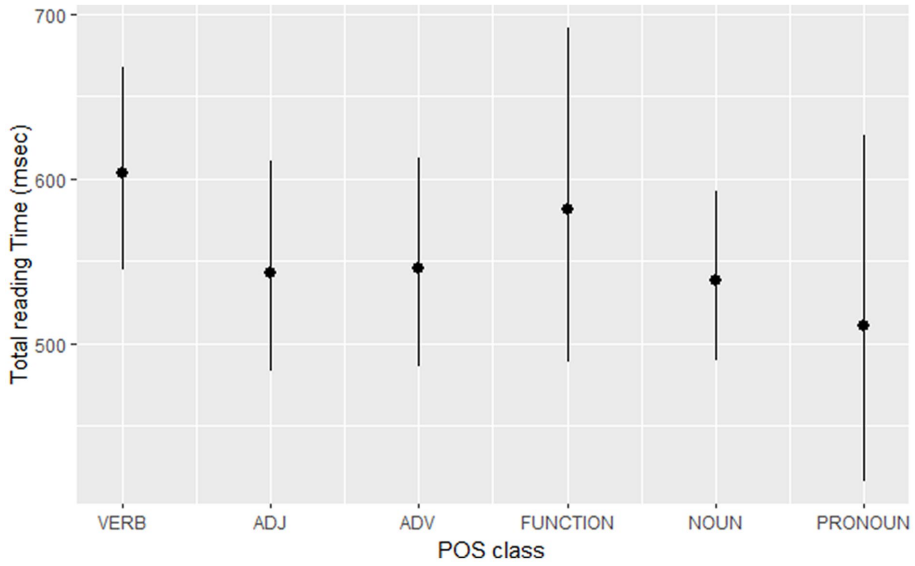


FIGURE 4
The predicted values of TT across parts of speech.

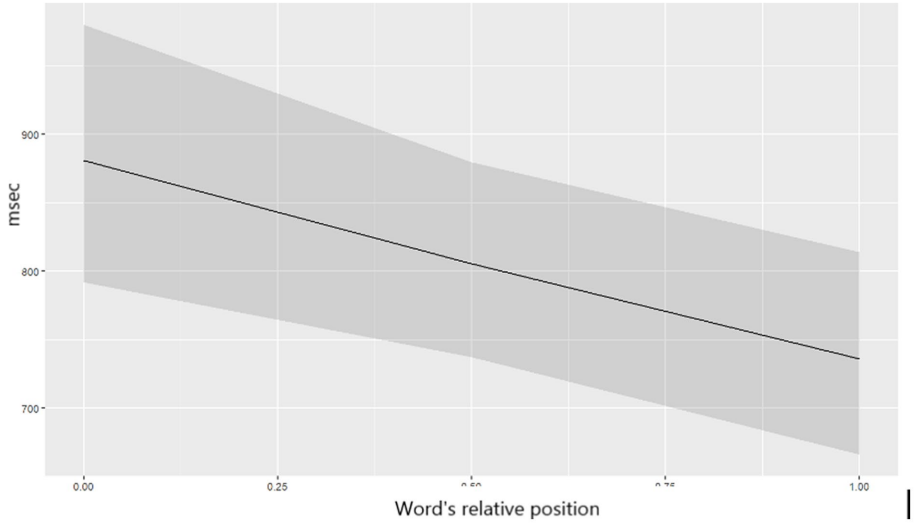


FIGURE 5
The predicted values of TT depending on the word's relative position in a sentence.

elicited longer FFD (Est. = 0.22, SE = 0.02, $t = 12.25$, $p < 0.001$) and SFD (Est. = 0.09, SE = 0.03, $t = 3.06$, $p = 0.036$), whereas it shortened GD (Est. = −0.12, SE = 0.02, $t = -5.42$, $p < 0.001$) and TT (Est. = −0.23, SE = 0.02, $t = -12.22$, $p < 0.001$) (Figure 6).

Reading skill in Adyghe

The self-reported reading skill score in Adyghe significantly affected reading, which was seen in late fixation durations measures (GD and TT). With an increasing level of reading skills both measures decreased: GD with Est. = −0.16, SE = 0.04, $t = -4.17$, $p < 0.001$, and TT with Est. = −0.19, SE = 0.04, $t = -4.52$, $p < 0.001$.

Within-group analysis of reading in two languages

To guarantee that a within-group analysis across two languages can be run, and reading in two languages is comparable in the group under study, we first analyzed reading comprehension in both languages. Comprehension accuracy in Russian was, on average, 0.9, SD = 0.07, range 0.69–1, and comprehension accuracy in Adyghe was, on average, 0.88, SD = 0.09, range 0.67–0.99. The Shapiro test showed that accuracy data distribution departed from normality ($p < 0.001$), which is why a non-parametric test was used. The Wilcoxon signed

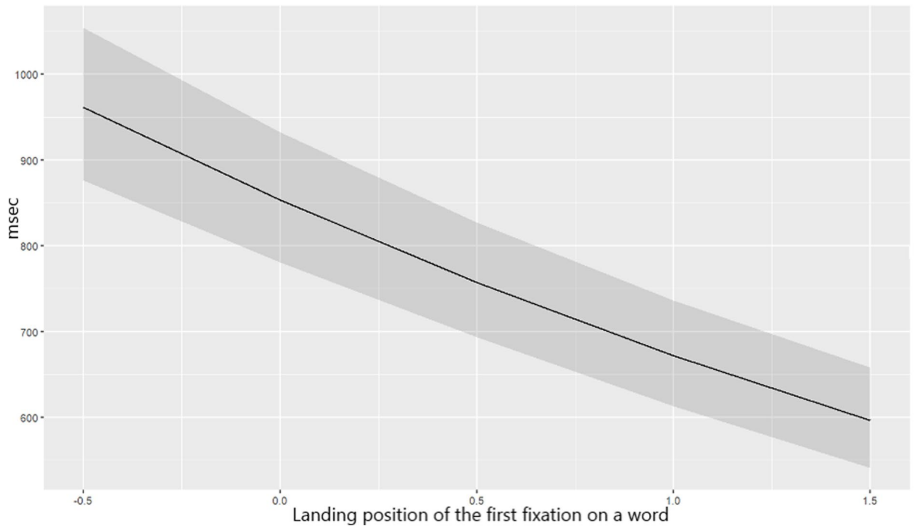


FIGURE 6
The predicted values of TT depending on the landing position of a first fixation on a word.

TABLE 5 Descriptive statistics of eye-movement measures in two languages ($N = 38$), Mean (SD).

Measure	Measurement	Adyghe	Russian
FFD	msec	290 (47.6)	211 (23.6)
SFD		317 (50.9)	226 (31.1)
GD		670 (167)	250 (37.7)
TT		936 (292)	300 (53.7)
P0	%	2 (0.02)	43 (0.1)
P1		20 (0.08)	40 (0.06)
P2+		77 (0.09)	16 (0.08)
RO		19 (0.09)	12 (0.07)
RG		17 (0.08)	12 (0.06)
Fixation count	N	3.52 (0.94)	0.82 (0.24)
Landing position	%	31 (0.07)	48 (0.04)

rank test resulted in non-significant differences between reading comprehension in both languages ($V = 306$, $p < 0.5$).

These results are essential for further data analysis and interpretation, as they validate a within-group comparison of reading in two languages, and eliminate the effect of longer reading times, due to poorer comprehension. Table 5 summarizes descriptive measures of eye movements while reading in two languages.

There were two significant effects observed consistently across all basic measures (FFD, SFD, GD, and TT): the effect of language (Adyghe), and word frequency. All reading times were higher in Adyghe compared to reading times in Russian ($p < 0.001$). Probability measures substantiated more effortful processing in Adyghe with lower probabilities of skipping and single fixations on words (both p's < 0.001), and higher probabilities of 2+ fixations ($p < 0.001$) and regressions from the current word ($p = 0.023$).

A higher word frequency decreased all reading times (FFD, SFD, GD, and TT) with $p < 0.001$, increased skipping rate ($p < 0.001$), and decreased probabilities of more than one fixation ($p < 0.001$), re-fixations ($p < 0.001$), and regressions from the word ($p = 0.012$).

Word length and landing position were another two variables with consistent significant effects ($p < 0.001$) in SFD and GD. Additionally, an increased word length increased FFD ($p = 0.003$), TT ($p < 0.001$), and probability of more than one fixation ($p < 0.001$), whereas it decreased the probability of skipping and fixating a word once only (both p's < 0.001). Landing position further from the word's beginning increased not only SFD and GD, but also FFD ($p < 0.001$).

The main effects of parafoveal words (either frequency or length) were not significant. Word's relative position further from the sentence beginning increased FFD (Est. = 0.03, SE = 0.01, $t = 3.70$, $p = 0.005$). The main effects of reading skills in Adyghe and Russian did not reach significance in any measures.

There were some interactions of language with other variables. We are reminded that Russian was taken as a baseline level for comparison, and it is, therefore, implied in the models' intercept. Primarily, reading skills in both languages significantly affected reading in Adyghe, compared to reading in Russian, with $p < 0.001$ in all duration measures (FFD, SFD, GD, and TT). However, the direction of the effect was different. Higher reading skill in Adyghe accelerated reading in it compared to reading in Russian, whereas higher reading skills in Russian slowed down reading in Adyghe compared to reading in Russian.

No significant interaction of language and word frequency was found. Significant effects of word length in the interaction with language were found in late measures (GD and TT) and in fixation probabilities. Namely, longer words were read significantly longer in Adyghe (GD: Est. = 0.05, SE = 0.00, $t = 10.45$, $p < 0.001$; TT: Est. = 0.04, SE = 0.01, $t = 7.19$, $p < 0.001$) than words of the same length were read in Russian. Compared to Russian, longer words were less likely to be fixated only once (P1: Est. = −0.35, SE = 0.03, $t = -13.54$, $p < 0.001$), and were more likely to be fixated more than twice (P2+: Est. = 0.20, SE = 0.03, $t = 6.83$, $p < 0.001$).

The effects of the parafoveally located words' properties (frequency and length) and current word's relative position were not significant

with the exception for the length of a previous word. Longer words on the left side decreased the total reading time of a current word (Est. = −0.02, SE = 0.00, $t = -3.39$, $p = 0.015$). Landing position further from the word's beginning interacted with language in both early (FFD) and late measures (GD, TT). Namely, it took more time during the first fixation to process the word, and less time to read it in the next fixations compared to the same landing position in Russian.

Discussion

The present study aimed to answer two research questions: (1) what are the benchmarks of eye movements while reading in a polysynthetic language (Adyghe), and (2) how does its reading differ from reading in a synthetic language (Russian) that is based on the same script? To answer these questions, we collected eye-movement data while reading two corpora: the Russian Sentence Corpus (RSC, Laurinavichyute et al., 2019) and the Adyghe Sentence Corpus (ASC). The analysis of eye movements included two parts in line with the research questions. First, an analysis (of all words and target words exclusively) in a larger data sample ($N = 50$) while reading ASC was performed. Second, we conducted a within-group analysis ($N = 38$) of eye movements comparing reading in two languages.

Benchmarks of eye movements while reading in Adyghe

Overall, the most robust universal effects of word frequency and word length on eye movements found in previous research across different languages (Inhoff and Rayner, 1986; Rayner, 1998; Staub and Rayner, 2007) were confirmed in our study in a polysynthetic language. Simultaneously, the finding of word frequency not being a significant effect across a range of measures contradicts the previously studied effects across languages and might imply some inconsistencies in the Adyghe Corpus, which is a constantly developing source of word frequencies in Adyghe. Presumably, the word frequencies extracted at the moment of the study did not fully reflect actual language use and might need to be updated.

On the other hand, this peculiarity brings us to the underlying question of the definition of a word and its units in polysynthetic languages. Lexical affixes might be confused with roots, and a word form reflects not just the form variations of a lemma but new "words" in its common notion. The blurred word boundaries (Haspelmath, 2018) make it possible that we need a shift toward other frequency measures. It will likely be more efficient to include morpheme frequency and/or initial bigram frequency similar to the analysis conducted in Yan et al. (2014) in Uighur.

We also confirmed another universal effect in eye-tracking-while-reading research - the effect of word length. It was consistently observed in late duration measures, as well as in probability measures. Importantly, no effect of word length in early measures (FFD and SFD) resembles reading in Russian among monolingual adults in Laurinavichyute et al. (2019) and HSs in Parshina et al. (2021). No effect in RO and lower probability of regressions to the longer word (RG) are compatible with those in German (Kliegl et al., 2006), but not in Russian (Laurinavichyute et al., 2019; Parshina et al., 2021).

The effects of a previous/upcoming word in Adyghe partially resemble those in German monolinguals (Kliegl et al., 2006) and in high proficient Russian HSs (Parshina et al., 2021) but not in Russian monolinguals (Laurinavichyute et al., 2019). Specifically, longer previous words accelerated the total reading time of a current word in Adyghe, whereas longer upcoming words did not show any effect. This outcome seems logical, taking into account the higher average word length in Adyghe (*cf.* 7.5 letters in ASC vs. 5.7 letters in RSC) which does not enable their readers to extract lexical information from the right side in the parafoveal processing.

The influence of POS class on eye movements in Adyghe was in line with previous research in Russian (Laurinavichyute et al., 2019): verbs were read significantly slower than nouns (in TT), whereas other POS did not differ significantly from verb reading. This finding corresponds to the notion, across different fields of linguistics, that verbs are more complex units and are more difficult to acquire and process than nouns (Bassano, 2000; Mätzig et al., 2009; Crepaldi et al., 2011).

Finally, we observed a morphological effect of lexical affixes on eye movements in a polysynthetic language. Essentially, this finding confirmed that a higher number of lexical affixes increases cognitive load and is a relevant lexical feature to be controlled for. However, we have to acknowledge its limited distribution: only total reading times in all-word analysis, but not in target-word analysis, were affected. Presumably, either the limited distribution of the effect in a sentence or less controlled materials might account for these results.

A limited distribution of the morphological effect (monomorphemic vs. inflected words) was earlier observed in agglutinative languages (Finnish and Turkish). In Finnish, isolated words were affected by morphological complexity, whereas words in a sentence context were not (Hyönä et al., 2002). In Turkish, this effect in sentence reading was observed in probability measures but not in early measures like SFD (Özkan et al., 2021). Having said that, a preserved effect of morphological complexity in a sentence context was reported in Yan et al. (2014) on the materials of a highly agglutinative language (Uighur) in both early (FFD) and late (GD) measures.

Reading in polysynthetic Adyghe vs. reading in synthetic Russian

Whereas reading in both languages seem to be affected similarly by word frequency with more frequent words being read faster, it seems to be affected differently by word length. Namely, two significant interactions of word length with language while reading in Adyghe demonstrated that longer words in Adyghe are read slower than words of the same length in Russian. This might account either for morphological differences between languages (longer Adyghe words might have a more complex morphemic structure which loads processing, whereas long Russian words are not necessarily polymorphemic) or for differences in participants' reading skills in the two languages.

The opposite effects of reading skills in Adyghe and Russian during reading in Adyghe reflect a common debate regarding language interference in bilinguals (Kaushanskaya and Marian, 2007; Libben and Titone, 2009). A higher reading skill in one language accelerated processing in that language but inevitably impeded processing in

another one. Hence, participants with a higher self-assessed reading skill in Adyghe read Adyghe sentences faster than Russian ones, whereas participants with a higher self-assessed reading skill in Russian read Adyghe sentences slower than Russian ones.

Non-significant main effects of the neighboring words characteristics (frequency and length) together with their non-significant interactions with language leads us to conclude that speakers of a polysynthetic language do not rely on information about neighboring words. On the contrary, different Russian-speaking groups (monolinguals in Laurinavichyute et al., 2019; HSs and L2 learners in Parshina et al., 2021) do extract some information from upcoming words, even though it is predominantly observed on late measures. Apparently, bilingual speakers of a polysynthetic language transfer this processing pattern to their other language (Russian, in this case), which distinguishes their reading in Russian from other Russian-speaking populations.

Noteworthy are the differences in the preferred landing position across the two languages. Statistical analysis showed that a further landing position on an Adyghe word will result in more efficient word processing (with a longer FFD but shorter GD and TT) compared to reading in Russian if landing on the same position. Descriptively, Adyghe bilinguals tend to land closer to the word beginning (on the first 31% of the word letters) when reading in Adyghe and closer to the word's center (on the first 48% of the word letters) when reading in Russian.

Limitations and further research

We must admit some limitations of the study. A corpus study has the pitfall of using less controlled materials, which can lead to multicollinearity among predictors. We partially addressed this issue in the target-word analysis, where three variables were controlled (frequency, length, and POS with verbs and nouns as levels), and in the all-word analysis, where the variance of the inflation factor (VIF) of the predictors was always less than 2. Most morphemic features in our data (except for the number of lexical affixes) were highly correlated with word length, which restricted us to one morphemic variable in the analysis and limited our investigation of morphological effects on reading in Adyghe.

Consequently, the primary suggestion for further research is either an orthogonally-designed experimental study on reading in Adyghe or a further exploitation of the ASC from a different perspective. For instance, the number of morphemes, together with the number of lexical and grammatical affixes, could be considered for another controlled-condition study. The great variety of dialects in Adyghe is an area for further corpus research. Not only was dialectal variation not the focus of our study, but we also had to exclude speakers of Kabardian from the analysis to ensure comparability with other dialects. Their data are, in turn, freely available together with other materials of the study at OSF, DOI 10.17605/OSF.IO/5UR8D (see footnote 4) and can be used in further research.

Apart from that, we see a potential to investigate in more detail the transfer of reading patterns that bilinguals make from one language to another. In our study, we observed this kind of transfer regarding the neighboring words: Adyghe-Russian bilinguals do not rely on their characteristics while reading in any language, whereas other Russian-speaking populations do when they read in Russian. The list of independent variables used to study eye movements from this perspective could be extended, and a different type of analysis (e.g., a scanpath analysis) could shed more light on reading patterns in the two languages and their interaction.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/5ur8d/?view_only=432e327cd0e64b5ca062be7e7e56b9b3.

Ethics statement

The studies involving humans were approved by the HSE Committee on Interuniversity Surveys and Ethical Assessment of Empirical Research. The participants provided their written informed consent to participate in this study.

Author contributions

OD and OP contributed to the conception of the study, curatorship of the data collection, and manuscript editing. NZ, OP, BO, IB, ShU, and SM were responsible for the stimuli creation. NZ, OP, BO, EK, AZ, and SM contributed to the participants recruitment and data collection. NZ contributed to the coding, data analysis, manuscript writing and editing. All authors contributed to the article and approved the submitted version.

Funding

This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1212701/full#supplementary-material

References

Arkhangelskiy, T., Bagirokova, I., Lander, Y., and Lander, A. Adyghe Corpus. Available at: http://adyghe.web-corpora.net/index_en.html (Accessed April 26, 2023).

Bassano, D. (2000). Early development of nouns and verbs in French: exploring the interface between lexicon and grammar. *J. Child Lang.* 27, 521–559. doi: 10.1017/S0305000900004396

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Boston, M., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: an evaluation using the Potsdam sentence Corpus. *J. Eye Mov. Res.* 2, 1–12. doi: 10.16910/jemr.2.1.1

Clifton, C.Jr., Staub, A., and Rayner, K. (2007). "Eye movements in reading words and sentences" in *Eye movement research: Insights into mind and brain*. eds. R. V. Gompel, M. Fisher, W. Murray and R. L. Hill (New York: Elsevier), 341–371.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting GECO: an eyetracking corpus of monolingual and bilingual sentence reading. *Behav. Res.* 49, 602–615. doi: 10.3758/s13428-016-0734-0

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

Crepaldi, D., Berlingeri, M., Paulesu, E., and Luzzatti, C. (2011). A place for nouns and a place for verbs? A critical review of neurocognitive data on grammatical-class effects. *Brain Lang.* 116, 33–49. doi: 10.1016/j.bandl.2010.09.005

Daniel, M., and Lander, Y. (2011). "The Caucasian languages" in *The languages and linguistics of Europe. A comprehensive guide*. eds. B. Kortmann and J. V. Auwera, vol. *1* (Berlin, Boston: De Gruyter Mouton), 125–157.

Data and data analysis at Open Science Framework. Available at: https://osf.io/5ur8d/?view_only=432e327cd0e64b5ca062be7e7e56b9b3

Dryer, M. S., and Haspelmath, M. (2013). WALS Online (v2020.3) [Data set]. Zenodo. Available at: https://wals.info (Accessed April 26, 2023).

Haspelmath, M. (2018). The last word on polysynthesis: a review article. *Linguist. Typol.* 22, 307–326. doi: 10.1515/lingty-2018-0011

Holmqvist, K., Nyström, N., and Andersson, R., (2011). in R. Dewhurst, H. Jarodzka and J. Van de Weijer (Eds.) *Eye tracking: A comprehensive guide to methods and measures*, Oxford, UK: Oxford University Press.

Hyönä, J., Vainio, S., and Laine, M.. (2002). A morphological effect obtains for isolated words but not for words in sentence context. *Eur. J. Cogn. Psychol.* 14, 417–433. doi: 10.1080/09541440143000131

Inhoff, A. W., and Radach, R. (1998). "Definition and computation of oculomotor measures in the study of cognitive processes" in *Eye guidance in reading and scene perception*. ed. G. Underwood (Elsevier Science Ltd.), 29–53.

Inhoff, A. W., and Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: effects of word frequency. *Percept. Psychophys.* 40, 431–439. doi: 10.3758/BF03208203

Kaushanskaya, M., and Marian, V. (2007). Bilingual language processing and interference in bilinguals: evidence from eye tracking and picture naming. *Lang. Learn.* 57, 119–163. doi: 10.1111/j.1467-9922.2007.00401.x

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* 16, 262–284. doi: 10.1080/09541440340000213

Kliegl, R., Nuthmann, A., and Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *J. Exp. Psychol. Gen.* 135, 12–35. doi: 10.1037/0096-3445.135.1.12

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., and Kliegl, R. (2019). Russian sentence Corpus: benchmark measures of eye movements in reading in Russian. *Behav. Res. Methods* 51, 1161–1178. doi: 10.3758/s13428-018-1051-6

Libben, M. R., and Titone, D. A. (2009). Bilingual lexical access in context: evidence from eye movements during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 381–390. doi: 10.1037/a0014875

Liversedge, S. P., Drieghe, D., Li, X., Yan, G., Bai, X., and Hyönä, J. (2016). Universality in eye movements and reading: a trilingual investigation. *Cognition* 147, 1–20. doi: 10.1016/j.cognition.2015.10.013

Lopukhina, A., Zdorova, N., Staroverova, V., Ladinskaya, N., Kaprielova, A., Goldina, S., et al. (2022). Benchmark measures of eye movements during reading in Russian children. PsyArXiv [Preprint]. doi: 10.31234/osf.io/2x5pk

Lüdecke, D. (2017). sjstats: statistical functions for regression models. R package version 0.12.0. Available at: https://CRAN.R-project.org/package=sjstats

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language pro-files in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Mätzig, S., Druks, J., Masterson, J., and Vigliocco, G. (2009). Noun and verb differences in picture naming: past studies and new evidence. *Cortex* 45, 738–758. doi: 10.1016/j.cortex.2008.10.003

Özkan, A., Beken Fikri, F., Kırkıcı, B., Kliegl, R., and Acartürk, C. (2021). Eye movement control in Turkish sentence reading. *Q. J. Exp. Psychol.* 74, 377–397. doi: 10.1177/1747021820963310

Parshina, O., Laurinavichyute, A., and Sekerina, I. (2021). Eye-movement benchmarks in heritage language reading. *Biling. Lang. Congn.* 24, 69–82. doi: 10.1017/S136672892000019X

Polinsky, M. (ed.) (2020). "Introduction" in *The Oxford handbook of languages of the Caucasus*. (Oxford: Oxford University Press), 1–25.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.r-project.org/

Rakhlin, N. V., Kornilov, S. A., and Grigorenko, E. L. (2017). "Learning to read Russian" in *Learning to read across languages and writing systems*. eds. L. Verhoeven and C. Perfetti (Cambridge University Press), 393–415.

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

Rayner, K., and Well, A. D. (1996). Effects of contextual constraint on eye movements in Reading: a further examination. *Psychon. Bull. Rev.* 3, 504–509. doi: 10.3758/BF03214555

Rothman, J., Bayram, F., DeLuca, V., di Pisa, G., Duñabeitia, J., Gharibi, K., et al. (2022). Monolingual comparative normativity in bilingualism research is out of "control": arguments and alternatives. *Appl. Psycholinguist.* 44, 316–329. doi: 10.1017/S0142716422000315

Russian Population Census (2020). Available at: https://eng.rosstat.gov.ru/ (Accessed April 26, 2023).

Schilling, H. H., Rayner, K., and Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Mem. Cogn.* 26, 1270–1281. doi: 10.3758/bf03201199

Schmauder, A. R., Morris, R. K., and Poynor, D. V. (2000). Lexical processing and text integration of function and content words: evidence from priming and eye fixations. *Mem. Cogn.* 28, 1098–1108. doi: 10.3758/BF03211811

Sereno, S. C., and Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends Cogn. Sci.* 7, 489–493. doi: 10.1016/j.tics.2003.09.010

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H. D., Alexeeva, S., Amenta, S., et al. (2022). Expanding horizons of cross-linguistic research on reading: the multilingual eye-movement Corpus (MECO). *Behav Res* 54, 2843–2863. doi: 10.3758/s13428-021-01772-6

Staub, A., and Rayner, K. (2007). "Eye movements and on–line comprehension processes" in *The Oxford handbook of psycholinguistics*. ed. M. G. Gaskell (Oxford: Oxford University Press), 325–342.

Sui, L., Dirix, N., Woumans, E., and Duyck, W. (2022). GECO-CN: Ghent eye-tracking Corpus of sentence reading for Chinese-English bilinguals. *Behav. Res.* doi: 10.3758/s13428-022-01931-3

Vasishth, S., von der Malsburg, T., and Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *WIREs Cogn. Sci.* 4, 125–134. doi: 10.1002/wcs.1209

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Yan, M., Zhou, W., Shu, H., Yusupu, R., Miao, D., Krügel, A., et al. (2014). Eye movements guided by morphological structure: evidence from the Uighur language. *Cognition* 132, 181–215. doi: 10.1016/j.cognition.2014.03.008

Zhukova, M., and Grigorenko, E. (2019). "Developmental dyslexia in Russia" in *Developmental dyslexia across languages and writing systems*. eds. L. Verhoeven, C. Perfetti and K. Pugh (Cambridge University Press), 133–151.

Zdorova, N. (2023). Data and data analysis at Open Science Framework. [Data set]. Available at: https://osf.io/5ur8d/?view_only=432e327cd0e64b5ca062be7e7e56b9b3

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 September 2023
DOI 10.3389/fpsyg.2023.1145638

OPEN ACCESS

EDITED BY
Marijan Palmovic,
University of Zagreb, Croatia

REVIEWED BY
Lorna C. Quandt,
Gallaudet University, United States
Omid Khatin-Zadeh,
University of Electronic Science and Technology of China, China

*CORRESPONDENCE
Anna K. Laurinavichyute
✉ anna.laurinavichyute@uni-potsdam.de

RECEIVED 16 January 2023
ACCEPTED 31 August 2023
PUBLISHED 20 September 2023

CITATION
Ziubanova AA, Laurinavichyute AK and Parshina O (2023) Does early exposure to spoken and sign language affect reading fluency in deaf and hard-of-hearing adult signers?
Front. Psychol. 14:1145638.
doi: 10.3389/fpsyg.2023.1145638

Does early exposure to spoken and sign language affect reading fluency in deaf and hard-of-hearing adult signers?

Anastasia A. Ziubanova[1], Anna K. Laurinavichyute[2]* and Olga Parshina[3]

[1]Center for Language and Brain, HSE University, Moscow, Russia, [2]Department of Linguistics, University of Potsdam, Potsdam, Germany, [3]Psychology Department, Middlebury College, Middlebury, VT, United States

Introduction: Early linguistic background, and in particular, access to language, lays the foundation of future reading skills in deaf and hard-of-hearing signers. The current study aims to estimate the impact of two factors – early access to sign and/or spoken language – on reading fluency in deaf and hard-of-hearing adult Russian Sign Language speakers.

Methods: In the eye-tracking experiment, 26 deaf and 14 hard-of-hearing native Russian Sign Language speakers read 144 sentences from the Russian Sentence Corpus. Analysis of global eye-movement trajectories (scanpaths) was used to identify clusters of typical reading trajectories. The role of early access to sign and spoken language as well as vocabulary size as predictors of the more fluent reading pattern was tested.

Results: Hard-of-hearing signers with early access to sign language read more fluently than those who were exposed to sign language later in life or deaf signers without access to speech sounds. No association between early access to spoken language and reading fluency was found.

Discussion: Our results suggest a unique advantage for the hard-of-hearing individuals from having early access to both sign and spoken language and support the existing claims that early exposure to sign language is beneficial not only for deaf but also for hard-of-hearing children.

KEYWORDS

reading fluency, deaf, hard-of-hearing, sign language, multimodal bilingualism, scanpaths, eye movements

1. Introduction

Although able to reach high reading proficiency, deaf readers are on average less skilled than hearing ones (Goldin-Meadow and Mayberry, 2001; Luckner et al., 2005; Kelly and Barac-Cikoja, 2007). Poorer reading in deaf individuals was initially attributed to spoken language phonology deficit (Hanson, 1989), but later research indicated that phonological activation is not necessary for proficient reading (Mayberry et al., 2011; Bélanger et al., 2012, 2013; Clark et al., 2016; Thierfelder et al., 2020; *cf.* Blythe et al. (2018) arguing for phonological recoding and Yan et al. (2015) as well as Yan et al. (2021) arguing for phonological preview benefit). More recently, reading skills in deaf people have been associated with different social integration background and educational methods, personal cognitive and social strengths (Marschark et al., 2015), exposure to written language (Tomasuolo et al., 2019), silent lipreading (Kyle et al., 2016), and, most importantly, early language development (Padden and Ramsey, 2000; Mayberry, 2007;

Freel et al., 2011; Lederberg et al., 2013; Clark et al., 2016; Tomasuolo et al., 2019).

The foundation of early language development is access to language. In deaf and hard-of-hearing people, access to language can take different paths, be that access to sign language, to spoken language, or both. The precise role of each route for reading proficiency is under debate. Mayberry and Lock (2003; see also Clark et al., 2016) claim that it is early sign language acquisition that is essential for later reading abilities (based on data from children with severe and profound hearing loss, who have no access to the sounds of spoken language). Early acquisition of sign language is crucial not only for future proficiency in the sign language itself (in particular, for grammaticality judgments, Cormier et al., 2012; syntax, Boudreault and Mayberry, 2006; Henner et al., 2016; vocabulary, Caselli et al., 2021; Berger et al., 2023), but also for the later processing of written language (Clark et al., 2016). In particular, knowledge of American Sign Language syntax is correlated with the knowledge of English syntax (Chamberlain and Mayberry, 2008; Pinar et al., 2017; Hoffmeister et al., 2022); large vocabulary in Dutch Sign Language is correlated with large vocabulary in written Dutch (Hermans et al., 2008); better antonym knowledge in American Sign Language is correlated with better reading in English (Novogrodsky et al., 2014); better knowledge of American Sign Language is correlated with better comprehension of written English (Freel et al., 2011). Perhaps most convincingly, proficiency in American Sign Language was the single significant predictor of performance on nationally standardized measures of reading comprehension, English language use, and mathematics (Hrastinski and Wilbur, 2016).

However, early acquisition of sign language might be not the only road to proficient reading. Tomasuolo et al. (2019) found that deaf children of deaf parents, deaf oral monolinguals, and people with normal hearing had similar fixation durations during reading, and all outperformed deaf children of hearing parents who learned sign language only after the age of six. Tomasuolo and colleagues concluded that competence in either sign or spoken language is crucial for skilled reading in deaf. In a similar vein, Bertone and Volpato (2009) claim the critical role of (partial) access to spoken language: orally-trained children with access to speech sounds (cochlear implantation) outperformed all other groups of deaf children in a picture-matching task. To summarize, there is currently no consensus on whether it is access to sign or spoken language, or both that is important for future reading skills.

The first factor – early access to sign language – primarily depends on the hearing status of the child's parents, since deaf parents tend to be signers, and hearing parents tend to either learn sign language together with their child (which might help children to gain age-appropriate SL vocabulary, see Berger et al., 2023) or opt for oral communication and education without any use of sign language. Deaf children born to deaf parents are likely to have early access to sign language and successfully acquire it as their first language. They are usually referred to as native signers, defined as having at least one deaf parent (here, we follow Tomasuolo et al., 2019; Hoffmeister et al., 2022, and others). In contrast, deaf children born to hearing parents may be deprived of sign language input – in fact, of any language input – as infants, which may hinder overall language development (Goldin-Meadow and Mayberry, 2001; Mayberry, 2007).

The second factor – early access to spoken language – depends on the degree of hearing loss of the child assuming other factors such as the quality of caretaker-child interactions, socioeconomic status, peer socialization, and cultural and individual differences are equal. For infants with some level of hearing loss, the severity of their hearing loss typically determines the amount of spoken language input they receive during infancy. Slight to moderately severe degrees of hearing loss correspond to the speech sound range the individuals perceive (see Table A2 in Appendix), and individuals with slight to moderate hearing loss have partial access to spoken language sounds. Hard-of-hearing children who have access to speech sounds from birth (e.g., from one or both parents, siblings, or other caretakers who use spoken language) are likely to acquire spoken language early. Children with severe and profound deafness are minimally exposed to spoken language sounds (only via lip-reading) and start learning spoken language later, already at school or at pre-school correction classes. Later exposure to spoken language may lead to lower spoken language proficiency (Bertone and Volpato, 2009).

The current study aims to add to the existing evidence on reading fluency in deaf and hard-of-hearing (DHH) signers: in addition to the early access to sign language, we also consider the access to spoken language approximated by the degree of hearing loss as a factor that can potentially influence reading fluency. While early access to sign language is clearly beneficial for reading skills of deaf individuals, it is less clear what role early access to spoken and/or sign language plays for hard-of-hearing individuals with partial access to speech sounds.

2. The present study

To investigate global reading fluency in DHH Russian signers, we focus not on the isolated measures related to individual word reading, such as fixation durations and skipping rates, but rather on the global trajectories of eye movements in reading the entire sentences (von der Malsburg and Vasishth, 2011). While the analysis of word-level eye movement characteristics is indispensable for studying how individual word properties affect reading, the analysis of scanpaths (i.e., sequences of eye movements) focuses on the bigger picture. Scanpath analysis combines fixation locations and their durations during reading the entire sentence into one continuous measure and allows the researchers to quantify the similarity between eye movement trajectories of different people.

To illustrate the concept of a scanpath, Figure 1 visualizes a trajectory of eye movements made while reading a sentence. The x-axis marks words in the sentence and the y-axis shows time in seconds. In this case, the reader fixated on the first word for about 400 ms and then continued to read the sentence word by word, skipped the 5th and the 6th words, fixated on the 7th and 8th words, skipped the 9th word and fixated on the 10th word, then made a regression to the 7th word, etc. This trajectory is an example of non-fluent reading: the scanpath includes six regressions and one atypically long fixation – the last word in the sentence was fixated for more than a second.

Utilizing the scanpath method to compare reading strategies in English-Russian bilingual and Russian-speaking monolingual speakers, Parshina et al. (2021a,b) found that monolingual adult readers followed the fluent reading strategy (fast sentence reading times, high word skipping rates, and almost no regressions), suggesting no difficulties in word recognition or syntactic and semantic information integration. Bilingual readers with early exposure to the second language and earlier exposure to the print language (e.g., heritage speakers of Russian who immigrated to the USA later in childhood) preferred the intermediate strategy (longer sentence reading times, lower word skipping rates, and

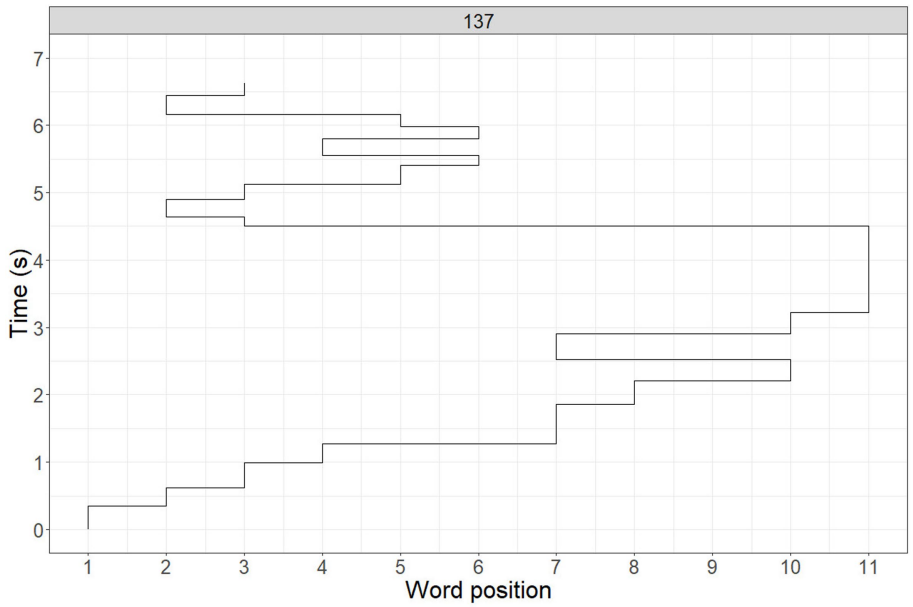


FIGURE 1
The example of gaze trajectory while sentence reading. The *y* axis shows sentence reading time in seconds and the *x* axis shows word position in the sentence.

more backward saccades to reread the words), indicative of delays in word recognition. Finally, bilingual readers with less exposure to spoken and written Russian (e.g., heritage speakers born in the USA) read according to the beginner strategy (even longer sentence reading times, more word and whole sentence rereadings), which the authors suggested reflects challenges not only in word recognition but also in the integration of morphosyntactic and semantic information.

In the present study (based on these findings and results in other studies, see above), we expect that early exposure to any type of language (sign or spoken) should be associated with greater reading fluency in DHH signers. That is, we expect higher reading fluency in both deaf and hard-of-hearing signers who have deaf parents and, therefore, were exposed to sign language from birth, and in hard-of-hearing signers who were exposed to spoken language from birth. We hypothesize that these readers will adopt a more fluent pattern of reading compared to DHH readers with less exposure to language (sign or spoken).

Admittedly, reading fluency *per se* is not a direct index of reading skill or successful comprehension: A text can be skimmed fast but poorly understood (Strukelj and Niehorster, 2018). Moreover, eye movements while reading depend not only on reading skill but also on reading goals and task demands (Mézière et al., 2021, 2022). For these reasons, we approximate reading skill through a combination of two measures: scanpaths, a combined measure capturing eye movements while reading, and questions probing sentence comprehension. A combination of skilled eye movement reading patterns and high question response accuracy would therefore index a better reading skill.

3. Methods

3.1. Participants

In Russia, deaf individuals are predominantly orally educated: they are taught to use monolingual spoken Russian as the primary means of production and lipreading for oral comprehension (Bazoev, 2016). This means that all DHH participants of the present study know spoken Russian and Russian print to some degree. Moreover, at the time of testing, all participants were daily users of Russian sign language (RSL; mean subjective assessment of proficiency = 8.97, SD = 1.46)[1], which means that all participants were bilingual and bimodal in RSL, spoken Russian, and Russian print. Participants were recruited from the Head Educational, Research and Methodological Center for Vocational Rehabilitation of persons with disabilities at Bauman University in Moscow. All participants were compensated with 500 Rub. The study was approved by the HSE ethics committee.

The study included 40 DHH signers: 26 participants with complete hearing loss ($M_{age} = 31$, $SD = 9$) and 14 hard-of-hearing participants ($M_{age} = 26$, $SD = 11$). The individual characteristics of each participant can be found in Table A1 in Appendix. The group of deaf participants included people with severe and profound hearing loss. The hard-of-hearing group of participants included people whose level of hearing loss ranged from slight to moderately severe. The degree of hearing loss was self-reported based on the diagnosis by a medical practitioner (established on the basis of either otoacoustic emissions testing (OAE) or pure-tone audiometry).

Fifteen out of twenty-six deaf participants were born to deaf parents (recall that such individuals are considered to be native signers) and had hereditary deafness, while 11 were born in hearing families

1 Only subjective assessment of RSL proficiency (How proficient would you say you are in RSL on a scale from 1 to 10?) is available because there is no standardized proficiency test for RSL. Self-reported proficiency strongly correlates with objective proficiency measures (Shameem, 1998; Marian et al., 2007).

TABLE 1 Characteristics of participants in each group.

	Deaf participants, hearing parents	Hard-of-hearing participants, hearing parents	Deaf participants, deaf parents	Hard-of-hearing participants, deaf parents	Stat. comparison
Demographics					
Total N	11	7	15	7	n.s.
Female participants	7	3	10	4	*
Vocabulary	33,272 (19,652)	52,571 (19,738)	51,466 (34,350)	49,571 (19,518)	n.s.
Age	28 (7.63)	25 (4.79)	33 (9.3)	27 (15.4)	n.s.
Start of RSL use	6.5 (3.1)	11.28 (5.49)	4 (1.4)	3.7 (2)	*
Years of education	16.54 (3.58)	16.42 (2.50)	17 (2.8)	13.6 (1.9)	n.s.
RSL proficiency (self-reported)	9.27 (1.55)	7.28 (1.49)	9.6 (0.8)	8.85 (1.2)	*
Characteristics of reading					
Accuracy	0.69 (0.46)	0.76 (0.43)	0.76 (0.43)	0.80 (0.40)	n.s.
Sentence reading times, ms	4,721 (1554)	4,692 (988)	4,611 (2117)	3,368 (979)	n.s.
Average fixation duration, ms	246 (128)	229 (118)	240 (122)	221 (106)	n.s.
Number of fixations on a sentence	19.2 (10)	20.5 (9.12)	19.2 (11)	15.3 (5.52)	n.s.

Statistical comparisons are based on the mixed-effects or linear models that the reader can find in the supplementary code. In the Statistical comparison column, n.s. stands for no significant differences. The significant differences between groups are as follows: Deaf participants and children of at least one deaf parent reported both earlier start of RSL use and higher RSL proficiency. In addition, there were significantly more females among deaf children of hearing parents than in other groups.
Individuals with at least one deaf parent are considered to be native signers. Numbers without parentheses represent counts or group means, numbers in parenthesis represent standard deviations. For the "Start of RSL use," we encoded the starting age of those participants who said that they use RSL from childhood as 5 years, which is a conservative estimate.

and had hearing loss due to other causes (see Table 1). Seven out of fourteen hard-of-hearing participants had deaf parents, the other seven had hearing parents. One participant from the hard-of-hearing group had a deaf mother and a hearing father and was classified as having deaf family due to access to sign language from birth.

The aim of the present study is to establish whether reading fluency of DHH signers is correlated with their parents' hearing status and the individual degree of hearing loss. The mapping from these predictors to the main factors of interest, early access to sign and spoken language, is as follows: parents' hearing status maps directly onto early access to sign language, which may benefit both deaf and hard-of-hearing individuals. In contrast, early access to spoken language maps onto the degree of participant's individual hearing loss with more severe loss leading to lesser spoken input the individual receives during infancy. We also hypothesize that the degree of hearing loss might interact with the parents' hearing status: the individuals with some access to speech sounds and at least one hearing parent are likely to have more early access to spoken sounds compared to individuals with deaf primary caretakers.

Materials. As reading materials, we used 144 sentences from the Russian Sentence Corpus developed as benchmark set of materials for assessing eye movements while reading in Russian (Laurinavichyute et al., 2019). The corpus is comprised of natural sentences randomly selected from the Russian National Corpus (https://Ruscorpora.ru) and normed for acceptability. Sentences had different syntactic structures: narratives, exclamations, and interrogatives, as well as sentences with non-standard word order. Sentences spanned from five to twelve words (with the average sentence length of 9 words) and were selected for being syntactically and lexically accessible. The Russian Sentence Corpus has been successfully read by advanced L2 learners and heritage speakers of Russian (Parshina et al., 2021b).

Originally, only 33% of the sentences in the corpus were followed by comprehension questions. To assess comprehension of DHH signers with higher precision, we introduced more questions: in the present study, 58% of sentences were followed by comprehension questions with three possible response options, see Example (1):

(1)			
	Sentence	***Дорога ведет в глухой лес, петляя по склонам.*** 'The road leads into the deep forest, winding along the slopes.'	
	Question	*Куда ведет дорога?* 'Where does the road lead?'	
	Correct answer	*В лес*	'Into the forest'
	Incorrect answer 1	*В огород*	'Into the garden'
	Incorrect answer 2	*В деревню*	'Into the village.

In addition, approximate vocabulary size of print Russian was measured for each participant using an online computerized adaptive testing tool (Golovin, 2014; Andreev et al., 2016; Ashkinazi and Golovin, 2016). During the test, participants see a word or a non-word and have to indicate whether they know its meaning. If participants indicate that they know the meaning of the word, they may with some probability be asked to select a correct interpretation of the meaning or a correct synonym out of four options. If a participant knows infrequent words, even less

frequent words are selected for further testing to estimate their vocabulary size with more precision.

3.2. Procedure

Stimuli were presented on the ASUS VG248QE monitor (resolution: 1,920 × 1,080 pix, response time: 1 ms, frame rate: 144 Hz, font face: 22-point Courier New). Eye movements were recorded at the rate of 1,000 Hz with desktop eye-tracker EyeLink 1,000+ using a chinrest. Eye-to-camera distance was 60 cm, eye-to-screen distance was 90 cm.

The experiment started with 9-dot camera calibration. After the calibration, a black dot was presented at the position of the first letter of the first word in the sentence. After the camera registered a fixation on the black dot, the sentence appeared. Participants were asked to read the sentence without signing (silent reading). If no fixation was registered on the black dot within 2 s, calibration was repeated. After having read the sentence, participants had to look at the red dot in the lower right corner of the screen. Fixation on the red dot triggered the next trial.

If the sentence was followed by a question, then after a fixation on the red dot was detected, the question appeared in place of the sentence. The response options were presented below the question. To select an answer, participants had to click on the response. The experiment started with three practice sentences and continued with 6 blocks, 24 experimental sentences in each. Between blocks, participants could have a break followed by a recalibration. The order in which the sentences appeared was randomized.

3.3. Analysis

To answer the main research question of the study, i.e., whether more proficient sentence reading trajectories in DHH participants are associated with early exposure to language, sign and/or spoken, we followed the steps in analysis in Parshina et al. (2021a,b). First, gaze trajectories (scanpaths) were recorded for all sentences for each participant. Trajectories with similar spatial and temporal characteristics (calculated using the Levenshtein distance) were then automatically grouped into clusters. To that end, we applied Gaussian mixture modeling (using the mclust package for R; Fraley and Raftery, 2007) that allowed us to identify the optimal number of clusters in each sentence. The advantage of using Gaussian mixture modeling over other clustering techniques (e.g., k-means clustering) is the method's ability to detect clusters even in the presence of overlapping parameters. The median number of clusters for the entire corpus was 2 clusters, ranging from 1 to 9 clusters in each sentence. To facilitate interpretation and to avoid capturing random variation in reading patterns we proceeded to fit the models with the fixed number of 2 Gaussians for all sentences in the corpus. Any participant could read some sentences more fluently, and others more effortfully, so the same person's reading trajectories for different sentences could be placed in different clusters. However, we expected that for each participant, one cluster would be dominant.

To find out whether early exposure to language affects cluster placement in DHH participants, we used a generalized mixed-effects model with the cluster as a dependent variable and parents' hearing status, participant's degree of hearing loss as predictors. We additionally used participants' vocabulary size, age, and gender as covariates, as these factors are known to affect reading (Baumann, 2014; von der Malsburg et al., 2015; Reilly et al., 2019). The model also included the age at which participants started learning RSL, as RSL proficiency might play a role in reading (Hrastinski and Wilbur, 2016). The model was fit using `lme4` package (Bates et al., 2014), with dummy-coded categorical fixed effects (hearing parents coded as 0, deaf parents as 1; hard of hearing participants coded as 0, deaf participants as 1). Vocabulary size, age, and the age at which participants started learning RSL were centered and scaled; gender was coded as 1 for female, −1 for male participants. The random effects structure included random intercepts for participants and sentences, as well as by-sentence random slopes for the fixed effects of participants' hearing status, their parents' hearing status, and the interaction of these effects. Correlations between random slopes were not estimated.

The data and analysis code are openly available at: https://osf.io/je8du/. The readers are encouraged to reproduce our analysis and to apply any other analyses they see fit to the data set.

4. Results

Based on the eye-movement characteristics, the two clusters earlier identified via Gaussian mixture modeling were labeled as more fluent and less fluent reading clusters (see Figure 2 for an example of typical gaze trajectories corresponding to the less-fluent and more-fluent reading clusters).

The less-fluent cluster was characterized by longer sentence reading times, longer fixation durations, greater number of fixations and regressions, and lower question response accuracy (see Table 2; note that in contrast to the eye-tracking measures, response accuracies were not used to compute the clusters). Both the less-fluent and the more-fluent reading clusters differed from the typical reading pattern of fluent monolingual Russian speakers reading the same materials [as reported in Parshina et al. (2021a,b)]. For comparison, monolingual Russian speakers had, on average, reading time of 2.1 s, and made 1.3 fixations per word (Parshina et al., 2021a,b).

We now turn to the main question of the study, namely whether parents' hearing status and participants' degree of hearing loss affect the reading patterns of DHH signers. Mixed-effect model demonstrated that the participant's degree of hearing loss did not affect cluster membership, whereas parents' hearing status did, and these two factors interacted (see Table 3; Figure 3): reading patterns of hard-of-hearing children of deaf parents were more likely (estimated 87% probability) to belong to the more fluent cluster than those of hard-of-hearing children of hearing parents (estimated 52% probability) or deaf children of deaf parents (estimated 49% probability). In addition, greater vocabulary size was strongly associated with placement to the more fluent cluster. Gender, age, and the age at which participants started to learn RSL did not affect the probability of cluster placement.

5. Discussion

The present study aimed to find out whether early exposure to sign and/or spoken language affects reading fluency in deaf and,

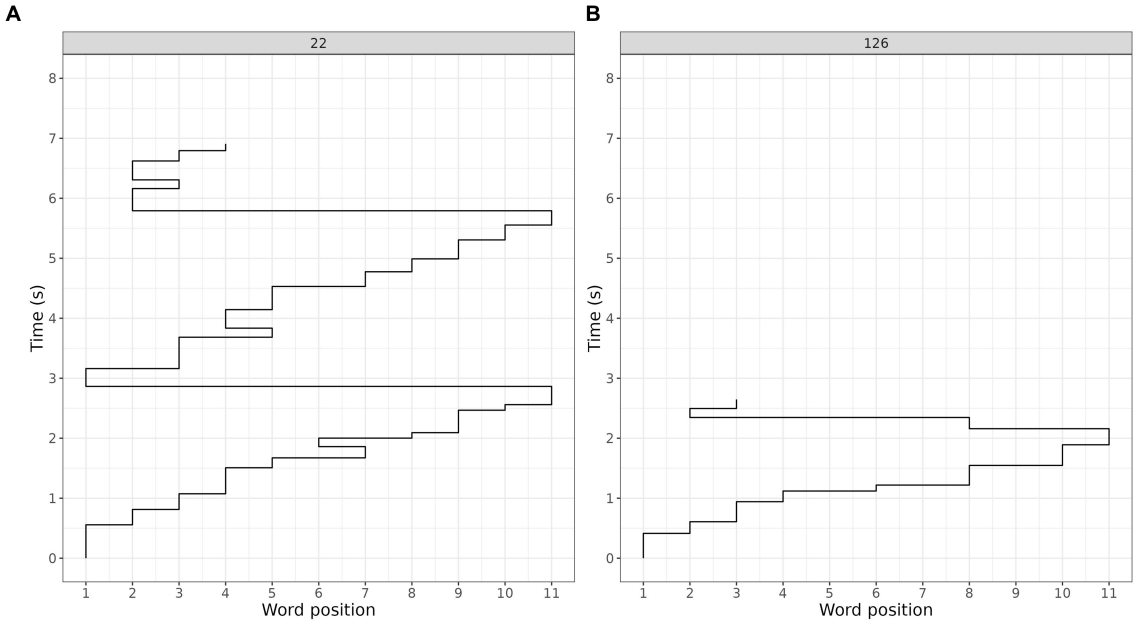


FIGURE 2
Gaze trajectories for reading the same sentence corresponding to typical reading patterns in the less-fluent **(A)** and more-fluent **(B)** clusters.

TABLE 2 Comparison of eye-movement measures and question response accuracies in the less-fluent vs. more-fluent cluster.

	Less-fluent reading	More-fluent reading	*p*-value
Accuracy, *M* (*SD*)	0.69 (0.46)	0.79 (0.41)	**0.002**
Number of fixations/ sentence, *M* (*SD*)	35 (15)	23 (10)	**<0.001**
Number of fixations/ word, *M* (*SD*)	2.9 (2.2)	1.9 (1.4)	**<0.001**
Sentence reading time, *M* (*SD*), s	9.1 (4)	5.5 (3)	**<0.001**
Fixation duration*, *M* (*SD*), ms	260 (174)	245 (156)	**<0.001**

Bold values indicate statistical significance.

especially, hard-of-hearing signers. While early access to sign language has been shown to benefit the reading skills of deaf individuals, it is less clear what role early access to spoken and/or sign language plays for hard-of-hearing individuals with partial access to speech sounds.

Our results suggest a unique advantage for the hard-of-hearing individuals from having early access to both sign and spoken language: native bilingual signers with access to speech sounds were much more likely to have more fluent reading patterns than any other group of participants. Early access to spoken language in hard-of-hearing signers with hearing parents did not correlate with reading fluency. Our results partially support the conclusions of Clark et al. (2016) who claimed that it is early sign language acquisition that is important for later reading fluency. However, early access to sign language seems to affect different groups of participants differentially: participants with partial access to speech sounds benefit from it the most in terms of reading. It seems that hard-of-hearing children born to deaf parents can have the best of both worlds: early access to sign language ensures timely language development, and on top of that, partial access to speech sounds further helps in mastering the spoken and print language system and vocabulary. Our results support the existing claims that early exposure to sign language is beneficial not only for deaf but also for hard-of-hearing children from infancy on (Mayberry, 2007; Freel et al., 2011; Humphries et al., 2014; Hall et al., 2019), and are broadly compatible with claims that bimodal education is effective for proficiency in written language (Lange et al., 2013; Henner et al., 2015).

The role of early access to spoken language is less clear: the lack of significant association between reading fluency and early access to spoken language does not mean that no link between the two exists. Conducting a follow-up study exclusively focused on investigating the impact of early access to spoken language on reading fluency in hard-of-hearing and deaf adult non-signers would provide valuable insights into this debate. However, the results of the current study suggest that for bilingual signers individuals access to spoken language may play a relatively smaller role in reading fluency compared to early access to sign language.

6. Conclusion

The current study aimed to evaluate whether and to what degree early access to sign language and early access to spoken language affect reading fluency in adult signers. We found that hard-of-hearing signers with early access to sign language and partial access to spoken language read more fluently than those who were exposed to sign language later in life. No association between early access to spoken language and reading fluency was found. If future studies confirm the greater role of early access to sign language for reading proficiency in hard-of-hearing signers, this could have deep impact on the social and educational policies ensuring the well-being of DHH individuals.

TABLE 3 Parameter estimates for the generalized mixed-effects model for the cluster distribution.

Predictors	Estimate (Log-Odds)	95% CI	*p*-value
(Intercept)	0.10	−0.80–1.01	0.822
Parents' hearing status (deaf)	1.88	0.54–3.22	**0.012**
Degree of hearing loss (profound)	0.07	−0.95–1.10	0.889
Vocabulary size	0.74	0.42–1.07	**<0.001**
Gender (female)	0.13	−0.20–0.47	0.435
Age	0.03	−0.31–0.37	0.863
Age of Start of RSL usage	−0.09	−0.52–0.34	0.691
Parents' hearing status × degree of hearing loss	−2.03	−3.43– −0.62	**0.010**
Random effects			
σ2	3.29		
τ00 item.id	2.01		
τ00 participant.id	0.82		
τ11 item.id.DeafParents:Deaf	10.95		
τ11 item.id.Deaf	1.96		
τ11 item.id.DeafParents	10.33		
N participants	40		
N item.id	144		
Observations	155,448		
Marginal R^2/Conditional R^2	0.125/0.656		

*p-values in the table are multiplied by two because we adjusted the alpha level by the factor of two. The reason for the adjustment is that we have performed the analysis after having collected data from 37 participants and then decided to proceed with data collection to have data from at least 40 participants.
Intercept corresponds to the baseline probability of placement to the more fluent cluster.
Bold values indicate statistical significance.

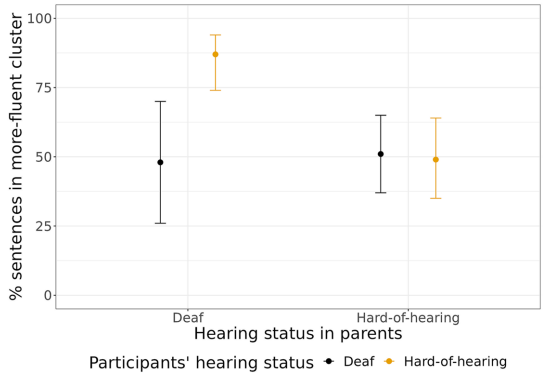


FIGURE 3
The estimated probability of placement to the more fluent cluster depending on the degree of hearing loss and parents' hearing status. Error bars represent 95% confidence intervals.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/je8du/.

Ethics statement

The studies involving humans were approved by the HSE Committee on Interuniversity Surveys and Ethical Assessment of Empirical Research. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

AZ and AL contributed to conception and design of the study. AZ programmed the procedure and performed data collection and wrote the first draft of the manuscript. OP, AZ, and AL performed the statistical analysis. AL wrote the revised version of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University). AL was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 317633480, SFB 1287. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 491466077.

Acknowledgments

We would like to thank Michael V. Mozgovoj from Bauman Moscow State Technical University for encouraging the DHH students from The Head Educational, Research and Methodological Center for Vocational Rehabilitation to participate in our experiment, and all the DHH RSL signers who participated in the experiment.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1145638/full#supplementary-material

References

Andreev, A., Ashkinazi, L., Golovin, G., Avanesyan, K., Babich, N., Brodsky, Y., et al. (2016). The futures we want: global sociology and the struggles for a better world. View from Russia [electronic resource]: collected papers. The 3rd ISA Forum of Sociology «The Futures We Want: Global Sociology and the Struggles for a Better World».

Ashkinazi, L. A., and Golovin, G. (2016). Studying passive vocabulary by means of the internet. Available at: https://www.myvocab.info/articles/studying-passive-vocabulary-by-means-of-the-internet

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models Usinglme4. *J. Stat. Softw.* 67. doi: 10.18637/jss.v067.i01

Baumann, J. F. (2014). "Vocabulary and reading comprehension: the nexus of meaning" in *Handbook of research on reading comprehension*. Eds. S. E. Israel and G. G. Duffy (London: Routledge).

Bazoev, V. Z. (2016). "Bilingualism and the education of deaf students: modern tendencies" in *Vestnik Leningradskogo gosudarstvennogo universiteta im*. ed. A. S. Pushkina (In Russian) 320–330.

Bélanger, N. N., Baum, S. R., and Mayberry, R. I. (2012). Reading difficulties in adult deaf readers of French: phonological codes, not guilty! *Sci. Stud. Read.* 16, 263–285. doi: 10.1080/10888438.2011.568555

Bélanger, N. N., Mayberry, R. I., and Rayner, K. (2013). Orthographic and phonological preview benefits: Parafoveal processing in skilled and less-skilled deaf readers. *Q. J. Exp. Psychol.* 66, 2237–2252. doi: 10.1080/17470218.2013.780085

Berger, L., Pyers, J., Lieberman, A., and Caselli, N. (2023). Parent American sign language skills correlate with child–but not toddler–ASL vocabulary size. *Lang. Acquis.* 1–15. doi: 10.1080/10489223.2023.2178312

Bertone, C., and Volpato, F. (2009). Oral language and sign language: possible approaches for deaf people's language development. *Cadernos de Saúde* 2, 51–62. doi: 10.34632/cadernosdesaude.2009.2976

Blythe, H. I., Dickins, J. H., Kennedy, C. R., and Liversedge, S. P. (2018). Phonological processing during silent reading in teenagers who are deaf/hard of hearing: an eye movement investigation. *Dev. Sci.* 21:e12643. doi: 10.1111/desc.12643

Boudreault, P., and Mayberry, R. I. (2006). Grammatical processing in American sign language: age of first-language acquisition effects in relation to syntactic structure. *Lang. Cogn. Process.* 21, 608–635. doi: 10.1080/01690960500139363

Caselli, N., Pyers, J., and Lieberman, A. M. (2021). Deaf children of hearing parents have age-level vocabulary growth when exposed to American sign language by 6 months of age. *J. Pediatr.* 232, 229–236. doi: 10.1016/j.jpeds.2021.01.029

Chamberlain, C., and Mayberry, R. I. (2008). American sign language syntactic and narrative comprehension in skilled and less skilled readers: bilingual and bimodal evidence for the linguistic basis of reading. *Appl. Psycholinguist.* 29, 367–388. doi: 10.1017/S014271640808017X

Clark, M. D., Hauser, P. C., Miller, P., Kargin, T., Rathmann, C., Guldenoglu, B., et al. (2016). The importance of early sign language acquisition for deaf readers. *Read. Writ. Q.* 32, 127–151. doi: 10.1080/10573569.2013.878123

Cormier, K., Schembri, A., Vinson, D., and Orfanidou, E. (2012). First language acquisition differs from second language acquisition in prelingually deaf signers: evidence from sensitivity to grammaticality judgement in British sign language. *Cognition* 124, 50–65. doi: 10.1016/j.cognition.2012.04.003

Fraley, C., and Raftery, A. (2007). Model-based methods of classification: using the mclust software in chemometrics. *J. Stat. Softw.* 18, 1–13. doi: 10.18637/jss.v018.i06

Freel, B. L., Clark, M. D., Anderson, M. L., Gilbert, G. L., Musyoka, M. M., and Hauser, P. C. (2011). Deaf individuals' bilingual abilities: American sign language proficiency, reading skills, and family characteristics. *Psychology* 2, 18–23. doi: 10.4236/psych.2011.21003

Goldin-Meadow, S., and Mayberry, R. I. (2001). How do profoundly deaf children learn to read? *Learn. Disabil. Res. Pract.* 16, 222–229. doi: 10.1111/0938-8982.00022

Golovin, G. (2014). Тест словарного запаса [Vocabulary Test]. Available at: https://www.myvocab.info/

Hall, M. L., Hall, W. C., and Caselli, N. K. (2019). Deaf children need language, not (just) speech. *First Lang.* 39, 367–395. doi: 10.1177/0142723719834102

Hanson, V. L. (1989). Phonology and reading: evidence from profoundly deaf readers. *Phonol. Read. Disabil.* 15, 199–207. doi: 10.3758/BF03197717

Henner, J., Caldwell-Harris, C. L., Novogrodsky, R., and Hoffmeister, R. (2016). American sign language syntax and analogical reasoning skills are influenced by early acquisition and age of entry to signing schools for the deaf. *Front. Psychol.* 7:1982. doi: 10.3389/fpsyg.2016.01982

Henner, J., Hoffmeister, R., Fish, S., Rosenburg, P., and DiDonna, D. (2015). *Bilingual instruction works even for deaf children of hearing parents*. Washington: American Educational Research Association.

Hermans, D., Knoors, H., Ormel, E., and Verhoeven, L. (2008). The relationship between the reading and signing skills of deaf children in bilingual education programs. *J. Deaf. Stud. Deaf. Educ.* 13, 518–530. doi: 10.1093/deafed/enn009

Hoffmeister, R., Henner, J., Caldwell-Harris, C., and Novogrodsky, R. (2022). Deaf children's ASL vocabulary and ASL syntax knowledge supports English knowledge. *J. Deaf. Stud. Deaf. Educ.* 27, 37–47. doi: 10.1093/deafed/enab032

Hrastinski, I., and Wilbur, R. B. (2016). Academic achievement of deaf and hard-of-hearing students in an ASL/English bilingual program. *J. Deaf. Stud. Deaf. Educ.* 21, 156–170. doi: 10.1093/deafed/env072

Humphries, T., Kushalnagar, P., Mathur, G., Napoli, D. J., Padden, C., and Rathmann, C. (2014). Ensuring language acquisition for deaf children: what linguists can do. *Language* 90, e31–e52. doi: 10.1353/lan.2014.0036

Kelly, L. P., and Barac-Cikoja, D. (2007). "The comprehension of skilled deaf readers" in *Children's comprehension problems in Oral and written language: a cognitive perspective*. Eds. K. Cain and J. Oakhill, 244–280.

Kyle, F. E., Campbell, R., and MacSweeney, M. (2016). The relative contributions of speechreading and vocabulary to deaf and hearing children's reading ability. *Res. Dev. Disabil.* 48, 13–24. doi: 10.1016/j.ridd.2015.10.004

Lange, C. M., Lane-Outlaw, S., Lange, W. E., and Sherwood, D. L. (2013). American sign language/English bilingual model: a longitudinal study of academic growth. *J. Deaf. Stud. Deaf. Educ.* 18, 532–544. doi: 10.1093/deafed/ent027

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., and Kliegl, R. (2019). Russian sentence Corpus: benchmark measures of eye movements in reading in Russian. *Behav. Res. Methods* 51, 1161–1178. doi: 10.3758/s13428-018-1051-6

Lederberg, A. R., Schick, B., and Spencer, P. E. (2013). Language and literacy development of deaf and hard-of-hearing children: successes and challenges. *Dev. Psychol.* 49, 15–30. doi: 10.1037/a0029558

Luckner, J. L., Sebald, A. M., Cooney, J., Young, J. III, and Muir, S. G. (2005). An examination of the evidence-based literacy research in deaf education. *Am. Ann. Deaf* 150, 443–456. doi: 10.1353/aad.2006.0008

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Learn.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Marschark, M., Shaver, D. M., Nagle, K. M., and Newman, L. A. (2015). Predicting the academic achievement of deaf and hard-of-hearing students from individual, household, communication, and educational factors. *Except. Child.* 81, 350–369. doi: 10.1177/0014402914563700

Mayberry, R. I. (2007). When timing is everything: age of first-language acquisition effects on second-language learning. *Appl. Psycholinguist.* 28, 537–549. doi: 10.1017/S0142716407070294

Mayberry, R. I., Del Giudice, A. A., and Lieberman, A. M. (2011). Reading achievement in relation to phonological coding and awareness in deaf readers: a meta-analysis. *J. Deaf Stud. Deaf Educ.* 16, 164–188. doi: 10.1093/deafed/enq049

Mayberry, R. I., and Lock, E. (2003). Age constraints on first versus second language acquisition: evidence for linguistic plasticity and epigenesis. *Brain Lang.* 87, 369–384. doi: 10.1016/S0093-934X(03)00137-8

Mézière, D., Yu, L., McArthur, G., Reichle, E., and von der Malsburg, T. (2022). Scanpath regularity as an index of Reading comprehension. Scientific Studies of Reading.

Mézière, D., Yu, L., Reichle, E., von der Malsburg, T., and McArthur, G. (2021). Using eye-tracking measures to predict Reading comprehension. doi: 10.31234/osf.io/v2rdp

Novogrodsky, R., Caldwell-Harris, C., Fish, S., and Hoffmeister, R. J. (2014). The development of antonym knowledge in American sign language (ASL) and its relationship to reading comprehension in English. *Lang. Learn.* 64, 749–770. doi: 10.1111/lang.12078

Padden, C., and Ramsey, C. (2000). American sign language and reading ability in deaf children. *Lang. Acquisit. Eye* 1, 65–89.

Parshina, O., Laurinavichyute, A. K., and Sekerina, I. A. (2021a). Eye-movement benchmarks in heritage language reading. *Biling. Lang. Congn.* 24, 69–82. doi: 10.1017/S136672892000019X

Parshina, O., Sekerina, I. A., Lopukhina, A., and von Der Malsburg, T. (2021b). Monolingual and bilingual Reading processes in Russian: an exploratory Scanpath analysis. *Read. Res. Q.* 57, 469–492. doi: 10.1002/rrq.414

Pinar, P., Carlson, M. T., Morford, J. P., and Dussias, P. E. (2017). Bilingual deaf readers' use of semantic and syntactic cues in the processing of English relative clauses. *Biling. Lang. Congn.* 20, 980–998. doi: 10.1017/S1366728916000602

Reilly, D., Neumann, D. L., and Andrews, G. (2019). Gender differences in reading and writing achievement: evidence from the National Assessment of educational Progress (NAEP). *Am. Psychol.* 74, 445–458. doi: 10.1037/amp0000356

Shameem, N. (1998). Validating self-reported language proficiency by testing performance in an immigrant community: the Wellington indo-Fijians. *Lang. Test.* 15, 86–108. doi: 10.1177/026553229801500104

Strukelj, A., and Niehorster, D. C. (2018). One page of text: eye movements during regular and thorough reading, skimming, and spell checking. *J. Eye Mov. Res.* 11, 1–22. doi: 10.16910/jemr.11.1.1

Thierfelder, P., Wigglesworth, G., and Tang, G. (2020). Orthographic and phonological activation in Hong Kong deaf readers: an eye-tracking study. *Q. J. Exp. Psychol.* 73, 2217–2235. doi: 10.1177/1747021820940223

Tomasuolo, E., Roccaforte, M., and Di Fabio, A. (2019). Reading and deafness: eye tracking in deaf readers with different linguistic background. *Appl. Linguis.* 40, 992–1008. doi: 10.1093/applin/amy049

von der Malsburg, T., Kliegl, R., and Vasishth, S. (2015). Determinants of scanpath regularity in reading. *Cogn. Sci.* 39, 1675–1703. doi: 10.1111/cogs.12208

von der Malsburg, T., and Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *J. Mem. Lang.* 65, 109–127. doi: 10.1016/j.jml.2011.02.004

Yan, G., Lan, Z., Meng, Z., Wang, Y., and Benson, V. (2021). Phonological coding during sentence reading in Chinese deaf readers: an eye-tracking study. *Sci. Stud. Read.* 25, 287–303. doi: 10.1080/10888438.2020.1778000

Yan, M., Pan, J., Bélanger, N. N., and Shu, H. (2015). Chinese deaf readers have early access to parafoveal semantics. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 254–261. doi: 10.1037/xlm0000035

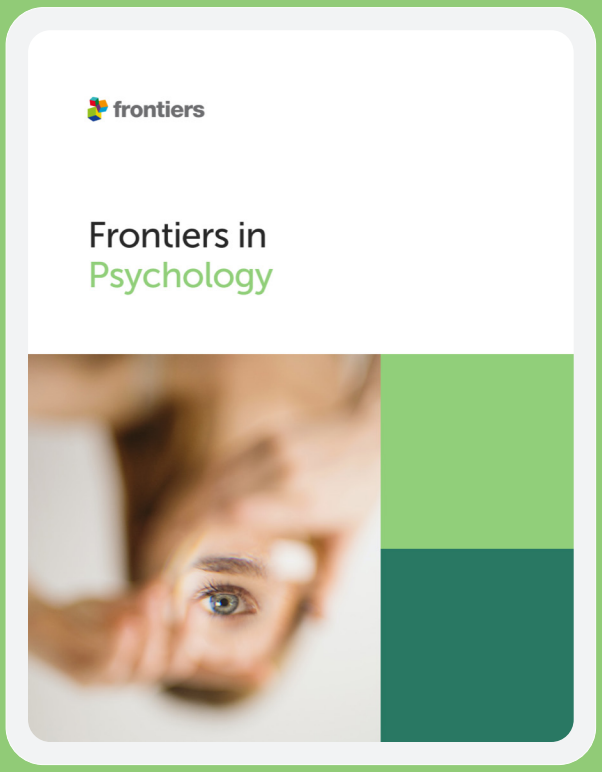
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact