STATISTICAL PROPERTIES OF LANGUAGE AFFECTING WORD RECOGNITION

STATISTICAL PROPERTIES OF LANGUAGE AFFECTING WORD RECOGNITION DURING NATURAL READING

By: GAISHA ORALOVA, MA

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2022) Hamilton, Ontario (Cognitive Science of Language)

TITLE: Statistical Properties of Language Affecting Word Recognition During Natural Reading

AUTHOR: Gaisha Oralova
M.A., University of Alberta
B.A., Beijing Language and Culture University

SUPERVISOR: Drs. Victor Kuperman and John F. Connolly

NUMBER OF PAGES: xii, 140

To my parents,
for their love, support, and encouragement

Abstract

Most previous research has explored how words are processed in isolation. However, reading is a complex process where an interplay of various factors affects word identification. Moreover, previous research has mainly focused on alphabetical languages, so extension of the existent findings to non-alphabetical languages is crucial. The current dissertation uses natural reading paradigms to study eye-movements and neurophysiological correlates of the statistical properties of words that affect word recognition during natural reading in English and Chinese.

Chapter 2 concerns the time-courses of word frequency and semantic similarity effects in the reading of English derived words. Previous research pointed to a paradox where behavioral experimental techniques showed earlier signatures of these properties than neuroimaging techniques. By combining eye-tracking and EEG and applying analytical techniques that target the onset of these effects, this study aims at investigating this paradox. Results still show that neurophysiological responses are either largely absent or appear at the same time as shown in eye-movement data.

Chapter 3 shows that the existence of spelling errors negatively impacts the recognition of correct spellings in Chinese. This is revealed by the "spelling entropy effect", which measures the uncertainty about choosing between correct and alternative spelling variants. This is the first study that used co-registration of eye-tracking and EEG to explore the behavioral and neurophysiological signatures of this uncertainty.

Chapter 4 studies how segmentation probabilities influence word segmentation and identification when reading Chinese. The results reveal that space becomes beneficial only when located at places where segmentation probability is considered high. This study is among the first to show beneficial effects of spacing at the sentence level and demonstrates how segmentation probabilities play a crucial role in Chinese word segmentation.

Cumulatively, the results obtained point to the existence of numerous factors involved in word identification in both alphabetic and logographic languages, which should be explored using natural reading experimental paradigms, such as co-registration of EEG and eye-tracking, for obtaining a multifaceted view of word recognition processes.

Acknowledgments

Throughout my doctoral studies at the Cognitive Science of Language Program at McMaster University I have received a great deal of support and assistance. Here, I would like to thank each and every person who made this dissertation a success.

First and foremost, I would like to express my sincere gratitude to my esteemed supervisors, Dr. Victor Kuperman and Dr. John F. Connolly. Dr. John F. Connolly was the person who always believed in me and was confident in my potential to become a great scientist. He was the one who pointed to and inspired me to apply a complex and state-of-the-art methodology for exploring psycholinguistic problems, namely, the co-registration of EEG and eye-tracking. He was also the one who believed that if I could learn Chinese then I could learn anything in my life. Now, I remind myself of this every time I encounter hard problems I need to solve or hard skills I need to acquire.

I would like to offer special thanks to Dr. Victor Kuperman for his guidance throughout my graduate career, for his assistance at every stage of my research projects, for his insightful comments and suggestions, for his unwavering support every time I needed it most. He was very attentive and supportive during hardships in my academic and personal life, which I value a lot. Victor is a great scientist with great research ideas, which he always shares with his students. He is very passionate about research and very knowledgeable in his field. I have learned a lot and continue to learn a lot from him.

Second, I also want to thank Dr. Juhani Järvikivi, who sparked my interest in psycholinguistics during my MA studies at the University of Alberta. He also recommended that I apply to McMaster University for my doctoral studies as he saw that my research interests perfectly fit with research topics investigated at the Reading Lab and the Language, Memory and Brain Lab at the Centre for Advanced Research in Experimental and Applied Linguistics (ARiEAL). I wish to thank Dr. Juhani Järvikivi for inspiring my interest and defining the path of my future research and career.

Third, my appreciation also goes to the members of Reading Lab, who gave me considerable advice and suggestions at the early stages of my projects, on conference presentations, posters and when proofreading my manuscripts. I am grateful to my collaborators, Rober Boshra, Daniel Schmidtke and Aki-Juhani Kyröläinen. I also thank Kaitlyn Battershill for proofreading parts of this thesis.

Finally, I would like to acknowledge my parents, Takhir Oralov and Gulnara Oralova, who always believed in me and were supportive at each stage of my academic career. Without them all this would not have been possible. Also, I would like to thank my husband, Dmitriy for always being there for me during my ups and downs, and my mother-in-law, Larissa, for taking care of my son and her treasured support while I was writing this thesis. Lastly, I am extremely grateful to my son, Ivan, who was born during my PhD studies, and brought an increased sense of meaning into my life. Ivan, you were a breath of fresh air, which motivated me to finish writing this thesis. Thank you.

Table of contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of figures

Chapter 2

Chapter 3

Chapter 4

Declaration of academic achievement

This is a “sandwich” thesis as it is defined by the School of Graduate Studies at McMaster University. It includes three empirical studies, where I am the primary author. Chapter 4 has recently been published as an original research paper and the eye-tracking results of Chapter 3 has been published as part of another paper. The roles of each author for the three of the studies in this dissertation are outlined below.

Chapter 2

This study has been submitted and currently under review in *Journal of Experimental Psychology: Human Perception and Performance* as **Oralova, G**., Schmidtke, D., Boshra, R., Kyröläinen, A.J., Connolly, J.F., Kuperman, V. (submitted). The chicken or the egg? The timeline for lexical and semantic effects in derived word recognition using simultaneous recording of EEG and eye-tracking. The contribution of each author is as follows:

Oralova, G: study design, experiment programming (EEG and eye-tracking), literature review, data collection, data cleaning, EEG and eye-tracking data analysis, manuscript writing, revision, and preparation for publication.

Schmidtke, D: stimuli preparation, statistical analysis (eye-tracking data), generation of plots, writing parts of the manuscript, revision.

Boshra, R: synchronization of eye-tracking and EEG data, EEG data analysis, manuscript revision.

Kyröläinen, A.J.: EEG data modelling, manuscript revision.

Connolly, J.F.: EEG data analysis, manuscript revision.

Kuperman, V: study design, eye-tracking data analysis, manuscript writing and revision.

Chapter 3

This chapter has been submitted to *Journal of Experimental Psychology: Learning, Memory, and Cognition* as **Oralova, G**. & Boshra, R., Kyröläinen, A. J., Connolly, J.F., Kuperman, V. (submitted). Statistics of spelling errors affects brain responses during natural reading of Chinese: Evidence from co-registration of EEG and eye-tracking signals. The eye-tracking results of this study were published in Kuperman, V., Bar-On, A., Bertram, R., Boshra, R., Deutsch, A., Kyröläinen, A.J., Mathiopoulou, B., **Oralova, G**. and Protopapas, A. (2021). Prevalence of spelling errors affects reading behavior across languages. *Journal of Experimental Psychology: General*. The contribution of each author is as follows:

Oralova, G: study design, stimuli preparation, experiment programming (EEG and eye-tracking), literature review, data collection, data cleaning, EEG and eye-tracking data analysis, manuscript writing, revision, and preparation for publication.

Boshra, R: study design, data collection, synchronization of eye-tracking and EEG data, EEG data preprocessing and analysis, revision.

Kyröläinen, A.J.: EEG data modelling, manuscript revision.

Connolly, J.F.: EEG data analysis, manuscript revision.

Kuperman, V: study design, eye-tracking data analysis, manuscript writing and revision.

Chapter 4

This chapter has been published in *Frontiers in Psychology: Language Sciences* as **Oralova, G.** & Kuperman V. (2021). Effects of spacing in sentence reading in Chinese. The contribution of each author is as follows:

Oralova, G: study design, stimuli preparation, experiment programming, literature review, data collection, data cleaning, data analysis, manuscript writing, revision, and preparation for publication.

Kuperman, V: study design, eye-tracking data analysis, manuscript writing and revision.

Additional achievements

In addition to the studies presented in this dissertation, the author contributed to the publication of the following paper:

Ho, A., Boshra, R., Schmidtke, D., **Oralova, G**., Moro, A. L., Service, E., & Connolly, J. F. (2019). Electrophysiological evidence for the integral nature of tone in Mandarin spoken word recognition. *Neuropsychologia*, *131*, 325-332.

Chapter 1

The *what*, *how* and *when* of visual word processing

In the age of shared technology where written language is pervasive, reading seems to be the most fundamental skill of a well-educated person. In the last several decades, reading research has seen a profound interest and a vast amount of knowledge about this activity has been contributed by educators and psychologists, and more recently by psycholinguists and neuroscientists. Many cognitive processes have been uncovered showing that perceptual, memorial, and linguistic factors are all intertwined together to allow reading. All this research inevitably narrows down to a basic unit of meaning, namely, words. As once mentioned by Balota, 1994: “The word is as central to psycholinguists as the cell is to biologists.” Word recognition research has been central to developing computational models (Coltheart, Curtis, Atkins, & Haller, 1993; McClelland & Rumelhart, 1981; Morton, 1969, 1980; Seidenberg & McClelland, 1989), advancing reading acquisition research (e.g., Verhoeven, Reitsma, & Siegel, 2011), and understanding the neural mechanisms behind language processing (Pulvermüller, Shtyrov, & Ilmoniemi, 2005). This concern with word recognition research brought many bright minds in the field of psychology of language to a rich collection of questions about the cognitive processes involved in and the factors that influence the process of word recognition. Through this research, a vast amount of information has accumulated not only regarding the statistical properties of words, including word frequency, concreteness, and valence, but also how these properties influence word identification processes during reading. Subsequently, various computational models have been created to model word recognition processes based on these properties (Davis, 2010; McClelland & Rumelhart, 1981; Norris, 2006; Reichle, Warren, & McConnell, 2009). In order to construct a viable model of word recognition, researchers aimed to understand the *what*, *how* and *when* of visual word processing. In particular, the research has been concentrating on the following questions:

- *What* are the properties of words that influence their recognition performance during reading?
- *How* do these word properties influence word recognition?
- *When* do these word properties influence word recognition?

In general, this thesis has also put its focus on uncovering the *what*, *how* and *when* of statistical factors that affect word recognition processes and solves a number of issues related to these processes through further exploration.

Although much research has already been done on word recognition, there are two problematic trends that have emerged in the literature. First, most of the visual word recognition models are based on isolated word reading (for an extensive review, see Norris, 2013). Yet, reading is a complex process, which incorporates a number of perceptual and cognitive processes, including low level “bottom-up” processes that take up information from foveal and parafoveal vision (Baccino & Manunta, 2005; Schotter, Angele, & Rayner, 2012; White, Rayner, & Liversedge, 2005), high-level top-down processes, which integrate semantic information into the preceding context (Kutas & Hillyard, 1980; Molinaro, Conrad, Barber, & Carreiras, 2010), oculomotor programming that brings fixations to the word of interest (e.g., Nuthmann & Henderson, 2012), and attention allocation processes that direct limited attentional resources between words in fovea and parafovea (e.g., Becker, 1976; Kennedy, 2000; McCann, Folk, & Johnston, 1992). Not surprisingly, all these cognitive activities overlap with each other in time. Previous research has addressed the question of how isolated word reading differs from natural reading. For instance, Kornrumpf, Niefind, Sommer and Dimigen (2016) found that word reading in context is an interactive process that incorporates many sources of information and differs from isolated word reading substantially. Given the differences in how words are read in isolation and in context, research on words during normal and active reading is needed to further advance our understanding of natural reading processes. This thesis emphasizes the usage of natural word reading paradigms and employs experimental techniques, which allow free eye movements and active reading behavior.

Second, much research is done on English or on other languages with alphabetical writing system. Some aspects of word reading have universal principles regardless of the writing system, such as Universal Phonological Principle (UPP), according to which reading engages phonology at the smallest unit of a language, be it a phoneme, syllable, or the whole word (Perfetti, Zhang, & Berent, 1992). Nevertheless, there are differences in the “specific implementation of reading” (Perfetti & Liu, 2005, p.195). For instance, when activating phonology, Chinese does not activate sub-syllabic connections due to their absence in the language itself. Phonology in this language is activated by syllables, which are whole characters mapped to spoken syllables (see Perfetti, Liu, & Tan, 2005). Subsequently, a model of Chinese word identification will contain these language specific characteristics. This illustrates the importance of research on other languages, especially on writing systems that drastically differ from English. Two of the studies in this thesis (Chapters 3 & 4) explore probabilistic factors influencing Chinese word recognition.

In summary, there has been much experimental work on cognitive processes underlying visual word recognition. Notwithstanding, many models of word recognition were mainly based on research on the reading of isolated words and mostly in the English language. In this regard, this thesis aims at addressing specific gaps in the previous literature by utilizing natural reading paradigms and exploring statistical factors influencing word recognition in both English and Chinese. In particular, the research in this thesis falls into three subtopics, which explore the *what*, *how* and *when* of word recognition processes by seeking answers to the following research questions:

1. What is the time-course of lexical and semantic effects during English derived word recognition?
2. Do the statistics of spelling errors in Chinese affect recognition of correctly spelled words?
3. Do segmentation probabilities play a role in Chinese word segmentation and identification when reading spaced and unspaced texts?

The remainder of this chapter briefly introduces each of the studies that explore the questions above and lays out the outline of the thesis.

1 What is the time-course of lexical and semantic effects during English derived word recognition?

Although there has been a great amount of work investigating statistical properties of words and their effect on word processing, results from previous research often brought contradictory findings with regard to the existence of the effects and their timing during word recognition. Strikingly, some research showed that the existence of a certain effect can be contingent on experimental task, as is the case with the effect of word frequency. For instance, neuroimaging studies that used sentence reading tasks, where there is preceding context before a target word, have demonstrated a lack of this effect (Degno, Loberg, Zang, Zhang, Donelli, & Liversedge, 2019; Kretzschmar, Schlesewsky, & Staub, 2015; Solomyak & Marantz, 2010). On the other hand, other research on isolated words clearly indicated its apparent existence, especially when reading monomorphemic words (Assadollahi & Pulvermüller, 2003; Fruchter & Marantz, 2015; Penolazzi, Hauk, & Pulvermüller, 2007). Moreover, there are numerous eye movement studies that clearly point to the existence of the frequency effect (Inhoff & Rayner, 1986; Juhasz & Rayner, 2006; Kuperman & Van Dyke, 2013).

More confusion arises from research that concerns the timing of lexical and semantic effects during word recognition. Recent literature highlighted the apparent discrepancy in the timeline of the effects affecting complex word recognition as evidenced by neurophysiological (EEG/MEG) and behavioral (reaction time, eye-tracking) studies (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011; Kretzschmar et al., 2015; Schmidtke & Kuperman, 2019; Schmidtke, Matsuki, & Kuperman, 2017). The discrepancy lies in the responses obtained from behavioral experiments preceding the responses shown by brain electrical activity. Recently, this phenomenon was coined as the "paradox of brainless behavior" (Schmidtke & Kuperman, 2019).

The main goal of this study is to shed light on this paradox by combining both EEG and eye-tracking methodologies to explore two well-established effects frequently found in the derived word recognition research, namely, whole word frequency and semantic similarity. Whole word frequency is the measure of how often a derived word is used in the language. Semantic similarity measures similarity in meaning between the stem and the whole derived word, for instance, how the stem *vaccine* of the derived word *vaccination* is similar to the meaning of the whole word *vaccination*. These effects are explored in both lexical decision and sentence reading tasks to uncover task specific effects on their timing during recognition of words presented in context and in isolation. The expectations follow the assumptions outlined in the eye-mind link hypothesis (Just & Carpenter, 1980), according to which neurophysiological responses should precede those observed in behavior.

Moreover, when investigating the time when certain effects occur during word recognition, the majority of studies relied on the analysis of central tendency, which is not suitable for uncovering the onset of psycholinguistic effects on word recognition performance. Thus, we applied specific analytical techniques that are able to point to the beginning of the effects, such as quantile regression analysis for the analysis of eye-tracking data and generalized additive mixed modelling for the analysis of EEG data.

The main contribution of this study is to attempt to solve the "paradox of brainless behavior" in a more methodologically rigorous way. Yet, Chapter 2 findings show that paradox is still present: EEG analysis does not reveal any effects when words were read naturally in text. It is only in the lexical decision experiment, where words were read in isolation one after another, that a semantic similarity effect was observed and showed an earlier occurrence than in eye-tracking. In contrast, eye-tracking data revealed the existence of whole word frequency and semantic similarity effects in both experiments. The sentence reading experiment showed earlier occurrence of these effects than it is observed in EEG. In what follows, Chapter 2 concludes that the usage of one EEG technique, specifically ERP/FRP, in exploration of the time-course of certain experimental effects in a paradigm that involves natural movements of reader's eyes seems to be problematic. In fact, eye-tracking proved to be very useful in investigating the timing of the linguistic effects under natural reading conditions. This study calls for more sophisticated analysis techniques of FRP data as research using co-registration of EEG and eye-tracking is still in its infancy. Additionally, according to eye-tracking results obtained from both experiments, the timeline of the effects in the lexical decision experiment tended to be later than in the sentence reading experiment. This points to the effectiveness of using natural reading paradigms for the exploration of psycholinguistic effects, especially when it concerns their timing.

2 Do the statistics of spelling errors in Chinese affect recognition of correctly spelled words?

When reading social media or other unedited texts readers inevitably encounter numerous words written in their incorrect orthographic forms. How does every occurrence of a spelling error affect the recognition of that word in its correct spelling? According to several theories of learning, the strength of association between form and meaning of a certain word is determined by the frequency of simultaneous exposure to form and meaning of that word (Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; Ramscar, Dye, & McCauley, 2013). For example, the more you see the orthographic form *elefant* instead of *elephant* as the name of an animal, the stronger the connection of *elefant* and the weaker the connection of *elephant* with the meaning of the word *elephant*. Subsequently, a frequent encounter of erroneous orthographic form creates a separate orthographic representation in the mental lexicon that co-exists with the correct orthographic representation of a word. Upon the next encounter of that word, the mental representation of the erroneous spelling for that word starts to compete for activation with the representation of the correct spelling.

How does this competition affect visual word recognition processes? A recent study by Rahmanian and Kuperman (2019) hypothesized that competition between spelling variants could pose a difficulty for recognizing correctly spelled words. In their study, they used eye-tracking sentence reading and lexical decision experiments to measure eye fixation durations on correctly spelled English words that are often misspelled in the language. The competition between spelling variants was measured via *spelling entropy*, where words with higher spelling error frequency had higher spelling entropy values. Rahmanian and Kuperman (2019) found longer looking times and delayed reaction times for words that had higher spelling entropy. Later, this hypothesis was tested on other languages, such as Finnish, Hebrew and Greek, and showed similar results (Kuperman, Bar-On, Bertram, Boshra, Deutsch, Kyröläinen, Mathiopoulou, Oralova, & Protopapas, 2021). As mentioned earlier, the majority of studies were conducted on alphabetical languages, and these findings need to be extended to non-alphabetical languages to test theories for universality across writing systems. Chapter 3 focuses on this issue and investigates if the frequency of spelling errors also negatively impacts the recognition of a correctly spelled word in a language with a logographic writing system – Chinese.

Rahmanian and Kuperman (2019) and Kuperman et al. (2021) used eye-tracking or lexical decision experiments to reveal behavioral evidence for the spelling entropy effect. However, behavioral experiments are based on responses from readers that register the final state of the cognitive effort and do not provide the dynamics of that effort during fixation duration or during lexical decision latencies before a button press. Chapter 3 uses simultaneous recording of EEG and eye-tracking to explore both behavioral and neural responses when frequently misspelled words are first fixated during natural sentence reading.

In Chapter 3, eye-tracking results for Chinese confirm previous findings obtained in alphabetical languages that words with higher spelling entropy have prolonged fixation durations. EEG data analysis reflected average amplitude differences at the 150 – 300 ms window, where words with higher spelling entropy, or higher uncertainty, elicited negative going amplitude values. These findings confirm the existence of the spelling entropy effect in logographic languages such as Chinese, and point to the universality of the spelling entropy effect across writing systems. Moreover, as confirmed by the analysis of neurophysiological data, this study is the first to show that this effect happens early during word recognition, where the orthographic processing stage is reported in the previous literature (e.g., Newman & Connolly, 2004; Sauseng, Bergmann, & Wimmer, 2004).

3 Do segmentation probabilities play a role in Chinese word segmentation and identification when reading spaced and unspaced texts?

Alphabetical languages with conventional inter-word spacing heavily rely on spacing information during reading and word recognition, so that if spaces are eliminated, reading rate in those languages slows down by 30-50% (Rayner & Pollatsek, 1996). On the contrary, Chinese is a non-spaced language, in which readers lack visual cues for word segmentation. Strikingly, it was found that Chinese readers often disagree on what constitutes a word and where word boundaries should be placed (e.g., Hoosain, 1992; Liu, Li, Lin, & Li, 2013). In order to understand what constitutes a segmentation unit for Chinese readers and investigate where spaces are most beneficial, if at all, spacing was artificially introduced into a normally unspaced text in numerous experimental studies (e.g., Bai, Yan, Liversedge, Zang, & Rayner, 2008; Bassetti, 2009; Cui, Drieghe, Bai, Yan, & Liversedge, 2014; Inhoff, Liu, Wang &, Fu, 1997). Previous eye-tracking studies that used spacing manipulation to find evidence for spacing advantage measured eye fixation durations on a single word and sentence reading times when participants read spaced and unspaced Chinese sentences (e.g., Bai, Liang, Blythe, Zang, Yan, & Liversedge, 2013; Zang, Liang, Bai, Yan, & Liversedge, 2013). All studies were inconclusive regarding beneficial effects of spacing information at the word and sentence levels. Specifically, they found word level advantages (through shortened word reading times in the word-spaced condition) despite the null effects at the sentence level: word-spaced and unspaced sentences were read in an identical amount of time.

Another line of research explored how Chinese readers segment unspaced texts and tested the possibility that readers use statistical probabilities between character transitions that guide their segmentation decisions. Indeed, one of the studies has found that the probability of a character to be used as a single character word plays an important role in the decision to preview other characters to the right of fixation (Zang, Wang, Bai, Yan, Drieghe, & Liversedge, 2016). This statistic signals the reader that a word boundary should be placed at this position and there is no need to preview the next character to concatenate the fixated character with the one to the right of a fixation to form a word. In another study by Yen, Radach, Tzeng, and Tsai (2012), it was found that the probability of a character to serve as the beginning or the end of a word signifies the word boundaries. Cumulatively, these studies point to an important role of transitional (or segmentation) probabilities that readers consider when segmenting continuous text into words.

Given the incongruous body of previous literature regarding the advantage of spacing information at the word and sentence levels and given that Chinese readers do not always agree on where word boundaries should be placed, Chapter 4 hypothesized that spaces could be beneficial only at places where most readers agree on a word boundary. In other words, a space may be advantageous only at places where a segmentation probability for the word boundary is high. Chapter 4 utilized segmentation judgments for word boundaries reported in Liu et al. (2013) and Wang, Huang, Yao, and Chan (2015) and used their stimulus sentences to create three experimental conditions: (i) a natural unspaced condition; (ii) a heavily spaced condition, where spaces were inserted at every possible character transition (the transitions where at least 5% of raters agreed on a word boundary); and (iii) a lightly spaced condition, where spaces were inserted only in highly probable transitions (the transitions where at least 90% of raters agreed on a word boundary). The two spaced conditions were compared against the default unspaced condition for the exploration of the spacing effect at the word and sentence levels.

Similar to previous findings, Chapter 4 shows that heavily spaced sentences took identical time to read as their unspaced counterparts, however, on the other hand, words demarcated by spaces in the heavily spaced condition were read faster than unspaced words. In contrast, lightly spaced sentences, where spaces were placed only at highly probable word boundary positions, were read with shorter amounts of time compared to unspaced sentences. Additionally, word-level analysis also revealed shorter fixation times and increased skipping rates in the spaced condition. These findings point to the selective nature of the beneficial spacing effect: it is advantageous to reading behavior only when spaces are located at highly probable word transitions, where the majority of readers agree on a word boundary. The observed differences between lightly and heavily spaced conditions in Chapter 4 explain the discrepant findings in the earlier literature and highlight the use of segmentation probabilities as an important factor when studying Chinese reading. Nevertheless, testing effects of spacing at both extremes of segmentation probabilities does not seem to show spacing as an effective cue for Chinese word segmentation: despite the word level advantage, the beneficial effect is small or completely cancelled out at the sentence level.

Outline of the thesis

Using the co-registration of EEG and eye-tracking signals, Chapter 2 contributes its findings to the hotly debated topic about the timing of psycholinguistic effects frequently found during derived word recognition, whole word frequency and semantic similarity. Chapter 3 explores another effect that influences word recognition processes, namely, spelling entropy, and extends the findings from English to another language with a distinct writing system, namely, Chinese. Also, this study goes further in the analysis and provides the timeline of this effect using simultaneous recording of EEG and eye-tracking. Chapter 4 reports an eye-tracking study, which shows the practicality of using segmentation probabilities when studying the effects of spacing in Chinese reading and explains discrepant findings found in previous research. Furthermore, it points to the importance of segmentation probabilities as another factor that influences word reading. Chapter 5 summarizes all research findings in this thesis and discusses how they contribute to the existent literature on visual word recognition.

References

Assadollahi, R., & Pulvermüller, F. (2003). Early influences of word length and frequency: a group study using MEG. *Neuroreport*, *14*(8), 1183-1187.

Baccino, T., & Manunta, Y. (2005). Eye-fixation-related potentials: Insight into parafoveal processing. *Journal of Psychophysiology*, *19*(3), 204-215.

Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of experimental psychology: Human perception and performance*, *34*(5), 1277.

Bai, X., Liang, F., Blythe, H. I., Zang, C., Yan, G., & Liversedge, S. P. (2013). Interword spacing effects on the acquisition of new vocabulary for readers of Chinese as a second language. *Journal of Research in Reading*, *36*, S4-S17.

Bassetti, B. (2009). Effects of adding interword spacing on Chinese reading: A comparison of Chinese native readers and English readers of Chinese as a second language. *Applied Psycholinguistics*, *30*(4), 757-775.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on I discriminative learning. *Psychological review*, *118*(3), 438.

Becker, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 556.

Cui, L., Drieghe, D., Bai, X., Yan, G., & Liversedge, S. P. (2014). Parafoveal preview benefit in unspaced and spaced Chinese reading. *Quarterly Journal of Experimental Psychology*, *67*(11), 2172-2188.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological review*, *100*(4), 589.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological review*, *117*(3), 713.

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., & Liversedge, S. P. (2019). Parafoveal previews and lexical frequency in natural reading: Evidence from eye movements and fixation-related potentials. *Journal of Experimental Psychology: General*, *148*(3), 453.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011).

Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of experimental psychology: General*, *140*(4), 552.

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

Fruchter, J., & Marantz, A. (2015). Decomposition, lookup, and recombination: MEG evidence for the full decomposition model of complex visual word recognition. *Brain and language*, *143*, 81-96.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, *30*(4), 1383-1400.

Hoosain, R. (1992). Psychological reality of the word in Chinese. In *Advances in psychology* (Vol. 90, pp. 111-130). North-Holland.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & psychophysics*, *40*(6), 431-439.

Inhoff, A. W., Liu, W., Wang, J., & Fu, D. J. (1997). Use of spatial information during the reading of Chinese text. *Cognitive research on Chinese language*, 296-329.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846-863.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Kennedy, A. (2000). Attention allocation in reading: Sequential or parallel? In *Reading as a perceptual process* (pp. 193-220). North-Holland.

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1648.

Kuperman, V., Bar-On, A., Bertram, R., Boshra, R., Deutsch, A., Kyröläinen, A. J., ... & Protopapas, A. (2021). Prevalence of spelling errors affects reading behavior across languages. *Journal of Experimental Psychology: General*.

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 802.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203-205.

Liu, P. P., Li, W. J., Lin, N., & Li, X. S. (2013). Do Chinese readers follow the national standard rules for word segmentation during reading?. *PloS one*, *8*(2), e55440.

McCann, R. S., Folk, C. L., & Johnston, J. C. (1992). The role of spatial attention in visual word processing. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1015.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, *88*(5), 375.

Molinaro, N., Conrad, M., Barber, H. A., & Carreiras, M. (2010). On the functional nature of the N400: Contrasting effects related to visual word recognition and contextual semantic integration. *Cognitive Neuroscience*, *1*(1), 1-7.

Morton, J. (1969). Interaction of information in word recognition. *Psychological review*, *76*(2), 165.

Morton, J. (1980). The logogen model and orthographic structure. *Cognitive processes in spelling*.

Newman, R. L., & Connolly, J. F. (2004). Determining the role of phonology in silent reading using event-related brain potentials. *Cognitive Brain Research*, *21*(1), 94-105.

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychological review*, *113*(2), 327.

Nuthmann, A., & Henderson, J. M. (2012). Using CRISP to model global characteristics of fixation durations in scene viewing and reading with a common mechanism. *Visual Cognition*, *20*(4-5), 457-494.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic bulletin & review*, *16*(1), 1-21.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological psychology*, *74*(3), 374-388.

Perfetti, C. A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a "universal" phonological principle. In *Advances in psychology* (Vol. 94, pp. 227-248). North-Holland.

Perfetti, C. A., & Liu, Y. (2005). Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing*, *18*(3), 193-210.

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: some implications of research on Chinese for general theories of reading. *Psychological review*, *112*(1), 43.

Pulvermüller, F., Shtyrov, Y., & Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition. *Journal of cognitive neuroscience*, *17*(6), 884-892.

Rahmanian, S., & Kuperman, V. (2019). Spelling errors impede recognition of correctly spelled word forms. *Scientific Studies of Reading*, *23*(1), 24-36.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of" mouses" in adult speech. *Language*, 760-793.

Rayner, K., & Pollatsek, A. (1996). Reading unspaced text is not easy: Comments on the implications of Epelboim et al.'s (1994) study for models of eye movement control in reading. *Vision research*, *36*(3), 461-465.

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and aging*, *21*(3), 448.

Sauseng, P., Bergmann, J., & Wimmer, H. (2004). When does the brain register deviances from standard word spellings?—An ERP study. *Cognitive brain research*, *20*(3), 529-532.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1793.

Schmidtke, D., & Kuperman, V. (2019). A paradox of apparent brainless behavior: The time-course of compound word recognition. *Cortex*, *116*, 250-267.

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, *74*(1), 5-35.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042-2057.

Verhoeven, L., Reitsma, P., & Siegel, L. S. (2011). Cognitive and linguistic factors in reading acquisition. *Reading and writing*, *24*(4), 387-394.

Wang, S., Huang, C. R., Yao, Y., & Chan, A. (2015, July). Create a manual Chinese word segmentation dataset using crowdsourcing method. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing* (pp. 7-14).

White, S. J., Rayner, K., & Liversedge, S. P. (2005). Eye movements and the modulation of parafoveal processing by foveal processing difficulty: A reexamination. *Psychonomic bulletin & review*, *12*(5), 891-896.

Yen, M. H., Radach, R., Tzeng, O. J. L., & Tsai, J. L. (2012). Usage of statistical cues for word boundary in reading Chinese sentences. *Reading and writing*, *25*(5), 1007-1029.

Zang, C., Liang, F., Bai, X., Yan, G., & Liversedge, S. P. (2013). Interword spacing and landing position effects during Chinese reading in children and adults. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 720.

Zang, C., Wang, Y., Bai, X., Yan, G., Drieghe, D., & Liversedge, S. P. (2016). The use of probabilistic lexicality cues for word segmentation in Chinese reading. *Quarterly Journal of Experimental Psychology*, *69*(3), 548-560.

Chapter 2

The chicken or the egg? The timeline for lexical and semantic effects in derived word recognition using simultaneous recording of EEG and eye-tracking

This study has been submitted and currently under review in *Journal of Experimental Psychology: Human Perception and Performance* as Oralova, G., Schmidtke, D., Boshra, R., Kyröläinen, A.J., Connolly, J.F., Kuperman, V. (submitted). The chicken or the egg? The timeline for lexical and semantic effects in derived word recognition using simultaneous recording of EEG and eye-tracking.

Abstract

Several studies on the time-course of word recognition highlighted a paradox: neural signatures for morphological and semantic effects reported in the literature tend to either not show up or lag behind the behavioral signatures of the same effects. We further explored this paradox by co-registering EEG and eye-tracking signals within participants while they read a series of derived English words (e.g., *government*) either embedded in sentences or shown in isolation as part of the lexical decision task. We orthogonally manipulated whole word frequency and semantic similarity between the derived word and its base (*govern*) and focused on determining the onsets of these effects on the simultaneously-recorded eye-movements and neural activity. Quantile regression analysis of eye-tracking data showed a consistent difference between high- and low-frequency words at 175 ms for words read in sentences, and at 741 ms for words shown in isolation. Generalized additive mixed modeling of fixation-related potentials (FRP) did not show a frequency effect whether the word was read in context or in isolation. Furthermore, eye fixation durations revealed a reliable contrast between transparent and opaque derived words at 355 ms in sentence reading experiment and at 855 ms in lexical decision experiment. On the contrary, FRP results showed an earlier effect of semantic similarity at 365 ms in the lexical decision experiment and no effect in sentence reading. In sum, the within-participants co-registration study of sentence reading replicated the behavior-before-brain paradox. A more intuitive brain-then-behavior sequence was partly observed in the lexical decision experiment. We discuss methodological implications of this finding for the validity of experimental paradigms commonly used in studies of complex word recognition.

1 Introduction

A central theme of the last two decades of lexical processing research has been the time-course of cognitive processes involved in word recognition (see among many others Dambacher, Kliegl, Hofmann, & Jacobs, 2006; Grainger, & Holcomb, 2009; Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006; Kliegl, Grabner, Rolfs, & Engbert, 2004; Sheridan and Reichle, 2016). In some subfields – including recognition of written morphologically complex words – precise characterization of this time-course is critical for adjudicating competing theories (see Feldman, Milin, Cho, Moscoso del Prado Martín, & O'Connor, 2015; Lavric, Elchlepp, & Rastle, 2012; Rastle, Davis, Marslen-Wilson, & Tyler, 2000; Solomyak & Marantz, 2010). An important obstacle to charting the time-course of recognizing complex words in print is the robust and paradoxical discrepancy in the relevant empirical evidence obtained by two different experimental techniques. The two techniques that are in the center of this debate, and of this paper, are eye-tracking (registration of eye-movements) and electroencephalography (EEG, the registration of electrical impulses emitted by the brain). Both techniques offer a fine-grained temporal resolution, on the millisecond scale, when registering behavioral or neural responses to stimuli. The timing of linguistic effects on word recognition that emerges in the signals that these techniques record is taken as a temporal signature of the cognitive processes leading to those effects (e.g., Bertram, 2011; Leminen, Smolka, Duñabeita, Pliatsikas, 2019). The abovementioned discrepancy lies in the robust observation that effects detected at a certain timepoint in the eye-movement record with high statistical confidence are either completely absent from the EEG record or emerge in that record much (100-400 ms) later, often when the behavioral response has already ended (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011; Kretzschmar, Schlesewsky, & Staub, 2015; Sereno & Rayner, 2003). This contradiction is reported both in simplex word recognition (Assadollahi & Pulvermüller, 2003; Hauk et al., 2006; Penolazzi, Hauk, & Pulvermüller, 2007; Pulvermüller, 2002) and, closer to the present theme, during recognition of morphologically complex derived (Schmidtke, Matsuki, & Kuperman, 2017) and compound words (Schmidtke & Kuperman, 2019).

This discrepancy is paradoxical because it suggests that behavioral responses predate the brain activity that must underlie the initiation of these responses in the first place. Since this is a logical impossibility, Schmidtke et al. (2017) and Schmidtke and Kuperman (2019) outlined several potential explanations for the paradox. First, most EEG studies of word processing report the timing of the peak amplitude as a temporal signature of a given effect (e.g., Lavric, Clapp, & Rastle, 2007; Rastle, Lavric, Elchlepp, & Crepaldi, 2015; Smolka, Gondan, & Rösler, 2015), with only some studies also reporting the timing when a certain percentage (50% or 75%) of the peak amplitude is reached (Pylkkänen, Feintuch, Hopkins, & Marantz, 2004). Since the peak amplitude indicates the maximum magnitude of a response rather than its onset, this practice is not meaningful for charting the time-course of word processing. A solution to this would be the use of an analytic technique that identifies the onset of an effect in the neural activity: this paper introduces such a technique below. Second, temporal estimates of lexical effects in the eye-movement record are typically derived from fixation duration, even though the effect is expected to emerge before the fixation terminates. Similar to demands on EEG analysis, a statistical method is necessary to establish the point of onset of an effect *within* the duration of the fixation. Schmidtke et al. (2017) used survival analysis to narrow down the temporal window within which an effect is reliably detectable. This paper pursues the same methodological objective using a different statistical method, e.g., quantile regression. A final, third, possibility is that the present approaches to collecting and analyzing EEG data do not have the required precision to offer meaningful estimates of the time-course of lexical processing.

Prior work (Schmidtke et al., 2017; Schmidtke & Kuperman, 2019) have examined some of the options above. They reviewed existing eye-tracking and EEG/MEG studies of morphological processing in derived words and compounds in English and analyzed new eye-tracking and lexical decision data. Applying survival analysis to the distributions of fixation durations, Schmidtke and colleagues were able to narrow down the time-window of most effects of linguistic form (e.g., word length, bigram frequency) and meaning (semantic similarity, psychological positivity) pertaining to the whole forms of derived words and compounds to a window between 120-220 ms, i.e., shorter than the average reading time for the word. The neurobehavioral literature reported the peak amplitudes and even sub-peak amplitudes for the same effects in the 300-600 ms window, after the word was read and the eyes moved on to the next word (for similar reports see co-registration studies by Dimigen et al., 2011; Kretzschmar et al., 2015). These results are diagnostic of the paradox wherein behavior predates brain activity. Yet Schmidtke et al. and Schmidtke and Kuperman indicated that their comparisons across experimental paradigms are incomplete. First, the behavioral and neural results under comparison were obtained from different participants, rather than in the co-registration paradigm. Second, these results shed light on the reading of connected texts, whereas virtually all neurobehavioral studies of complex word recognition used single word recognition with a meta-linguistic task as their "carrier" paradigm. In this paradigm, words are presented in isolation rather than in linguistic context and thus the reader cannot benefit either from a word's predictability in context or from the parafoveal preview of upcoming words, which provides early processing benefits when those upcoming words are fixated on (Hyönä, Bertram, & Pollatsek, 2004; Rayner, Ashby, Pollatsek, & Reichle, 2004). Due to this difference in tasks – text reading vs isolated word recognition (Kuperman, Drieghe, Keuleers, & Brysbaert, 2013)– it is presently unclear how the time-course captured by the eye-tracking and EEG record compares within and between the tasks.

The present study aimed to determine and compare the time-courses of behavioral and neural activity of the lexical processing of derived words in a methodologically complete way. We co-registered eye-tracking and EEG records from the same participants, eliminating the possibility of cross-sample variability. Furthermore, we conducted both the lexical decision task and the sentence reading task, while recording both the eye-movements and the brain responses. This enabled identification of task-specific demands and their implications for temporal estimates of lexical effects. Third, we implemented analytical techniques that enabled us to detect the onset of an effect – rather than its maximum – in both the eye-movement data (quantile regression) and in the

EEG data (generalized additive mixed models). See detailed descriptions in the Methods section.

Our goal was to track the emergence and time-course of two well-established lexical effects on complex word recognition in the behavior and the brain, i.e., word frequency and semantic similarity (defined below). An expected and intuitive order of events would be for an effect to first be statistically detectable in the neural record; the resulting change in brain activity would lead to an initiation of a behavioral response and, in turn, to the detectable onset of the effect of that predictor on reading behavior. Outcomes in which a behavioral response to a given predictor precedes a neural response to that predictor or occurs in the absence of the response would reiterate the paradox and highlight the methodological drawbacks of the experimental paradigms commonly used in word processing research. Below we justify our choice of two lexical variables used as detect access to form and meaning of derived words and review prior reports of respective effects in the human behavior and brain.

1.1 Whole word frequency

1.1.1 Behavioral evidence. Frequency with which a word occurs in language is a robust predictor of the cognitive effort of word processing, as evidenced by many behavioral paradigms, including lexical decision and sentence/text reading studies (see review in Brysbaert, Mandera, & Keuleers, 2018). In eye-movement studies of monomorphemic word reading, for instance, the word frequency effect has been detected in first fixation duration - the earliest durational eye-movement measure of lexical processing lasting an average of 248 ms for high frequency (HF) words and 298 ms for low frequency (LF) words (e.g., Inhoff & Rayner, 1986: HF = 248 ms, LF = 264 ms; Kretzschmar et al., 2015: HF:231 ms; LF: 244 ms ; Rayner, Ashby, Pollatsek, & Reichle, 2004: HF = 256 ms, LF = 282 ms; Rayner, Liversedge, White, & Vergilino-Perez, 2003: HF = 257 ms; LF = 298 ms; Rayner & Raney, 1996: HF = 248 ms, LF = 273 ms). Further distributional analyses brought forward the estimate of the earliest influence of word frequency on fixation durations substantially to an estimated time interval of between 180 ms and 200 ms for first fixation and gaze durations (e.g., Staub, White, Drieghe, Hollway, & Rayner, 2010). Therefore, when reading simplex words in connected text, readers begin to demonstrate sensitivity to word frequency within 200 ms of fixating on the target word**.**

Within the theories of complex word processing, the whole-word frequency effect bears additional significance. Often referred to as surface frequency or whole word frequency (whole word frequency hereafter), the frequency effect has served as a signature of access to a whole (complex) word representation, accessed either in conjunction with, or subsequent to, morphological decomposition (e.g., Fruchter & Marantz, 2015; Niswander, Pollatsek, & Rayner, 2000; Schreuder & Baayen, 1995; Taft, 1979; for a review see Amenta & Crepaldi, 2012).

Available sentence reading eye-movement studies have reported reliable effects of the whole word frequency on compound and derived word reading. Only a few studies have examined the whole word frequency effects of suffixed derived words in naturalistic eye-movement studies during reading. In a study of Dutch derived words, Kuperman, Bertram and Baayen (2010) found whole word frequency effects in single fixation durations (mean = 245 ms) and gaze durations (mean = 270 ms). Furthermore, Amenta, Marelli, & Crepaldi (2015) reported whole word frequency effects in Italian derived words in first fixation durations (mean = 243 ms), see also Niswander, Pollatsek, & Rayner (2000). Thus, the simplest statistical tools of analyzing the central tendency place the derived word frequency effect in the range of 240-270 ms. More recently, in three separate eye-movement studies of English derived word processing, Schmidtke et al. (2017) applied a distributional analysis to first fixation durations and reported that the earliest discernible effect of whole word frequency on first fixation durations occurred between 150 ms (study 5) and 169 ms (study 6).

While this study focuses on derived words, relevant and converging evidence exists in eye-tracking studies of compound reading. A recent large-scale corpus of eye-movements to over 900 English compounds during sentence reading from 440 participants (Schmidtke, Van Dyke, & Kuperman, 2020) reported a reliable effect of whole word frequency on first fixation durations (mean = 234 ms). This aligns well with the mean first fixation durations in other studies reporting compound frequency effect (Marelli & Luzzatti, 2012; mean duration = 231 ms), Dutch (Kuperman, Schreuder, Bertram, & Baayen, 2009; mean duration = 270 ms), Finnish (Bertram & Hyönä, 2003; Experiment 1 mean duration = 232 ms; Experiment 2 mean duration = 194 ms) and English (Juhasz, 2016; mean duration = 259 ms) (see Schmidtke & Kuperman, 2019 for a review). Furthermore, in a distributional analysis of the same corpus data, Schmidtke & Kuperman (2019) reported that word frequency reliably began to exert an influence on first fixation durations as early as 144 ms (Study 2) and as late as 219 ms (Study 5).

1.1.2 Neurophysiological evidence. Most neurophysiological studies addressing the time course of frequency effects have been conducted on simple monomorphemic words. For instance, a MEG study by Assadollahi and Pulvermüller (2003) observed a difference between high and low word surface frequency in the electromagnetic activity of the brain at 120-170 ms for short words and 225-250 ms for long words. Low frequency words led to stronger brain responses than high frequency words for both, short and long words. Similarly, in an ERP experiment, Penolazzi, Hauk, & Pulvermüller (2007) found neurophysiological signatures of word frequency between 120-180 ms, in an interaction with word length. However, there are studies showing much earlier effects of frequency. For instance, in an ERP lexical decision study of English by Sereno, Rayner and Posner (1998), the effect of word frequency was registered as early as 144 ms with low frequency words eliciting larger negative going amplitudes. Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson (2006) utilized regression analysis of EEG data and documented an even earlier effect of word frequency of English words at 110 ms. Recently, Sereno, Hand, Shahid, Mackenzie, & Leuthold (2020) found this effect even earlier, at 80-120 ms with enhanced negativity in anterior regions and enhanced positivity in posterior regions of the scalp.

Frequency effects for complex words, including derived words, show up much later in neurophysiological studies of the word recognition time-course. In a MEG lexical decision task, Fruchter & Marantz (2015) found an effect of whole word frequency in left middle temporal sites at a 431-500 ms window (for an extensive review of neurophysiological indices of derived and inflected words, please see Leminen, Smolka, Dunabeitia, & Pliatsikas, 2019).

Interestingly, there are neurophysiological studies that show no effect of word frequency during monomorphemic or derived word recognition. Perhaps the most powerful example comes from Kretzschmar et al., (2015) explored word predictability and word frequency effects of monomorphemic word in a sentence reading experiment using co-registration of EEG and eye-tracking from same participants. For the analysis of EEG data, or to be precise, fixation related potentials (FRPs), which is a signal time-locked to the onset of the first fixation, they used consecutive analysis of 50 ms windows from 150 ms to 700 ms at midline and lateral electrode sites. While there were robust effects of word frequency on multiple, including early, eye-movement measures, no reliable effect of the word frequency was registered in the brain activity. This was the first eye-tracking and EEG co-registration study that used natural sentence reading paradigm to explore the word frequency effect. In another sentence reading co-registration study, Degno, Loberg, Zang, Zhang, Donelli, & Liversedge (2019) also explored lexical frequency and parafoveal preview benefit effects on monomorphemic words. Similarly, no significant effect of lexical frequency was reported on FRP components. Furthermore, in a single trial correlational lexical decision and MEG study of derived words, Solomyak & Marantz (2010) report that the whole word frequency effect was not significant in either the M170 or the later M350 component. In fact, many additional EEG or MEG studies reported null or reduced effects of word frequency when words were presented with a preceding sentence context (Van Petten & Kutas, 1990; Dambacher, Kliegl, Hofmann, & Jacobs, 2006).

In sum, neurophysiological evidence for the word frequency effect in single-word paradigms suggests that for monomorphemic words the effect emerges between 110 ms and 180 ms, and within a window of 350 ms to 500 ms for complex words. Moreover, in some studies when words are shown in context, frequency effects tend to be weak or absent. Strikingly, the timelines for the word frequency effect for complex words differ across eye-tracking and neurophysiological studies, with eye-tracking always showing the effect earlier than neurophysiological paradigms.

1.2 Semantic similarity

Semantic similarity, or also called semantic transparency, refers to the extent to which the meaning of the complex word string is predictable from the meaning of its constituent morphemes. For example, the derived word *vaccination* is considered semantically transparent or has a high semantic similarity because its meaning is semantically predictable and similar to the meaning of its base form *vaccine*. This is not the case for a semantically opaque derived word, which has low semantic similarity, such as *department*, whereby *depart* has an unclear semantic relationship with the whole word. While operationalized in more than one way (see Gagné, Spalding, & Nisbet, 2016; Auch, Gagné, & Spalding, 2020), semantic similarity has been examined as a predictor of complex word processing across many behavioral (e.g., primed lexical decision latencies Rastle & Davis, 2008; unprimed lexical decision latencies: Rastle & Davis, 2008; eye-movements during reading: Marelli & Luzzatti, 2012) and neurophysiological paradigms (EEG: Morris, Frank, Grainger, & Holcomb, 2007; MEG: Brooks & Cid de Garcia, 2015).

Derived words of high semantic similarity, where the meaning of the whole word is similar to the meaning of its base, are expected and often found to be processed faster than words of low semantic similarity. It is argued that the conjunctive activation of the semantic representations of a complex word and its constituents is the mechanism responsible for this effect (El-Bialy, Gagné, & Spalding, 2013; Libben, 1998; Libben, Gibson, Yoon, & Sandra, 2003; Zwitserlood, 1994). That is, complex word processing benefits from a stronger semantic association between the whole word form and its constituent(s). Importantly, an effect of semantic similarity implies that, at some stage of the time-course of word recognition, the meaning of the whole complex word is processed.

1.2.1 Behavioral evidence. A few eye-movement studies have examined the semantic similarity or transparency effect in derived word processing, the object of this study. Marelli, Amenta, Morone, & Crepaldi (2013) found a transparency effect on first fixation duration when Italian derived words were presented without context in a priming lexical decision experiment combined with eye tracking (mean fixation duration = 256 ms). In eye-movement experiments where participants read English derived words embedded within sentences, Schmidtke et al. (2017) found that the earliest effects of semantic similarity (measured using Latent Semantic Analysis, defined below) occurred at 189 ms (Study 5; mean fixation duration = 231 ms), 219 ms (Study 6; mean fixation duration = 225 ms) and 207 ms (Study 7; mean fixation duration = 233 ms).

Converging time-course evidence comes from eye-tracking research on compound word reading (see reviews in Schmidtke, Van Dyke, & Kuperman, 2018 and Schmidtke & Kuperman, 2019). For instance, Juhasz (2007) found a main effect of semantic transparency on gaze duration in English compounds (mean gaze durations: opaque condition; 441 ms vs. transparent condition: 417 ms). Schmidtke et al. (2020) reported a main effect of semantic similarity between the right constituent and the whole compound word on first fixation durations (mean fixation duration = 234 ms) in English compounds. Marelli & Luzzatti (2012) reported effects of semantic transparency in first fixation durations during Italian compound word reading in isolation, as early as 231 ms. Going beyond central tendencies, Schmidtke & Kuperman (2019) conducted a distributional analysis of lexical effects on compound word reading and estimated the earliest reliable effects of left-whole semantic similarity (between the left constituent and the whole compound word) in English compound word reading on first fixation durations to 142 ms (Study 1; mean fixation duration = 229 ms), 142 ms (Study 2; mean fixation duration = 225 ms) and 161 ms (Study 4; mean fixation duration = 222 ms). The earliest onsets of the effect of right-whole semantic similarity (between the right constituent and the whole compound word) were detected somewhat later, at 173 ms (Study 1), 183 ms (Study 3; mean fixation duration = 231 ms) and 167 ms (Study 4).

Taken together, the studies mentioned above demonstrate that semantic similarity affects eye-movements during complex word recognition as early as 142 ms of the first fixation, and that *on average*, effects are observed at or close to 250 ms in first fixation durations.

1.2.2 Neurophysiological evidence. Most of the neurolinguistic research on the effect of semantic similarity has been conducted on English derived word recognition and has been conducted using EEG and MEG. Either technique has been used with tasks involving priming in combination with lexical decision or semantic categorization. Much of this body of work is reviewed in Leminen et al. (2019) and we draw upon relevant articles here.

In an experimental design which contrasted the priming effect of transparently related words (*hunter* - *HUNT*) with unrelated word pairs (*shovel* - *HUNT*), Morris, Frank, Grainger, & Holcomb (2007) and Lavric, Clapp, & Rastle (2007) reported differences across both conditions in the ERP signal at 250 ms (N250) in a lexical decision task. Transparent primes elicited larger negativities on unrelated targets. Using the same experimental design, Morris, Grainger, & Holcomb (2008) found the N250 effect in a semantic categorization task. Indeed, a series of priming-with-lexical decision ERP studies (Lavric, Rastle, & Clapp, 2011; Morris, Porter, Grainger, & Holcomb, 2011; Morris, Grainger, & Holcomb, 2013) have adopted similar experimental designs, i.e., contrasting transparent prime-target pairs (*voltage* - *VOLT*) with complex but unrelated prime-target pairs (*painter* - *VOLT*). Morris et al. (2013) reported a large effect of transparent complex words at 150-200 ms, yet Morris et al. (2011) reported an effect in the N250 component, and Lavric et al. (2011) found the effect later, in the N400 component (see also Kielar & Joanisse, 2011). More recently, Jared, Jouravlev, & Joanisse (2017) conducted a masked priming lexical decision task (Experiment 1b) which explored the effect of semantic transparency in four conditions (transparent, e.g., *foolish-FOOL*; quasi-transparent, e.g., *bookish-BOOK*, opaque, e.g., *vanish-VAN*, orthographic control, where a target word did not contain suffix but its prime did overlap in letters with the target *e.g., bucket-BUCK*). Once again, they reported a significant difference in mean amplitudes as a function of transparency in the N250 window (200-250 ms): target words that were primed by transparent words showed larger priming effects with larger negativities shown to unrelated targets. The same effect was observed in the N400 window (350-500 ms) where opaque primes did not facilitate the recognition of target words: no amplitude difference was observed between the opaque and orthographic control conditions.

Studies using MEG to study the effects of semantic transparency in visual complex word recognition have adopted lexical decision and lexical decision with priming experimental designs. Cavalli, Cole, Badier, Zielinski, Chanoine, & Ziegler (2016) reported semantic priming effects at M250. Further, Fruchter & Marantz (2015) conducted a lexical decision experiment of derived English words and reported that increased magnetic field activity associated with the “semantic fit” of the whole word and the stem, gauged via Latent Semantic Analysis score trended towards significance (p = 0.053) in the 300 to 500 ms window. An MEG study conducted by Lehtonen, Monahan and Poeppel (2011) investigated masked priming of derived words in English across three conditions: semantically transparent (*cleaner-CLEAN*), opaque (*corner-CORN*) and orthographic-only (*brothel-BROTH*) prime–target pairs. They found the same effects for transparent and opaque complex words at 220 ms post stimulus-onset (see Zweig, & Pylkkänen, 2009 for a similar design reporting semantic activity at 260 ms post onset for English derived words).

In sum, according to EEG/MEG studies, the effect of semantic similarity or transparency can be registered in the brain on average at 250 ms, within a window of time from 150 ms to 400 ms. Compared to the whole word frequency effect, the time course of similarity effects is more in line with the eye-movement record. However, it is still clear that, overall, there is a discrepancy in the temporal estimates obtained by behavioral and neuroimaging studies, with the former often preceding the latter.

To conclude, two temporally sensitive methodologies (EEG/MEG and eye-tracking) and two tasks (lexical decision and sentence reading) frequently used in word recognition research give rise to estimates of lexical activity that are mixed and hard to reconcile. While previous work has addressed some aspects of this discrepancy, the present study directly tests visual recognition of derived words in both paradigms and both tasks.

1.3 The Present Study

This study presents readers with English suffixed words either in isolation for lexical decision or in sentence context for silent reading for comprehension. In both tasks, we record the readers' eye-movements and the EEG brain activity, minimizing variability due to different samples, stimuli, and testing conditions. The critical question of this study is the temporal order in which effects of two well-established predictors of complex word recognition – whole word frequency and semantic similarity – emerge in the neural record and behavioral activity. The expectation, in line with the eye-mind link (Just & Carpenter, 1980), is that the former precedes the latter: for the oculomotor system to initiate a response influenced by a word property, that property first needs to affect the activity of the brain. It is also possible that the brain response does not materialize in a discernible behavioural change. Any other outcome signals a potential issue with methodological validity of either or both experimental paradigms, see above. Since the focus of the investigation is on the earliest detection of lexical effects, we complemented conventional statistical techniques for analyzing the central tendency in responses with the techniques enabling more detailed analyses of how the responses unfold over time. The description of these techniques and analyses of the data obtained with their help are reported below.

We address the contradictions in time-courses of lexical effects across neuroimaging and behavioural studies in several critical ways. First and foremost, we combine eye-tracking and EEG techniques. Most neurophysiological studies of sentence or text reading utilized the rapid serial visual presentation technique (RSVP), which employs word by word presentation, a condition that disregards crucial aspects of natural reading process, such as fluent reading of connected texts for comprehension. The RSVP technique suppresses parafoveal viewing of upcoming words and minimizes saccadic movements due to presenting each word in the same location. Eye-tracking, on the other hand, allows the exploration of cognitive processes during natural reading behavior. Thus, by combining these two techniques, we emphasize the beneficial aspects of both: exploring the exact timeline of cognitive events of interest under natural reading conditions while extracting underlying neurophysiological activity locked to each word fixation. Furthermore, as mentioned earlier, different experimental techniques often bring contradictory results in terms of the timeline of events during word recognition. This could be partially due to the use various experimental tasks, different set of stimuli, and different groups of participants, so it becomes difficult to compare results across studies. Here, we attempt to eliminate maximally these differences by running lexical decision and sentence reading tasks using the same set of stimuli and the same set of participants within each task.

As shown by Dimigen et al. (2011) and Kretschmar et al. (2015) studies, co-registration of EEG and eye-tracking alone may not rule out the timing controversies created when studies utilize these two methods in separate experiments. To analyze data collected from the co-registered eye-movements and EEG, we implement statistical tools that are geared toward establishing the earliest point in time when a certain variable has a discernible impact on eye-movement behavior and EEG. As argued by Schmidtke et al. (2017) and Schmidtke & Kuperman (2019), traditional methods of data analysis, such as ANOVA, *t*-test and linear mixed-effects regression, are blunt instruments for the task of modelling the time-course of lexical effects from a continuous random variable, such as eye-movement fixations or EEG amplitudes. For example, methods that analyze the peak or mean amplitude of an event-related signal or mean values of fixation durations cannot readily identify the earliest time when the effect of a particular word property begins to influence word recognition. As reviewed above, distributional analysis methods have detected the onset of whole word frequency and semantic similarity effects on eye-movement behavior within 100 to 400 ms of fixating on a morphologically complex word.

In the present analysis of eye-movements we apply quantile regression (Koenker & Bassett, 1978). While standard least squares regression techniques focus on the mean, quantile regression describes the entire conditional distribution of the dependent variable (Mosteller & Tukey, 1977). In the context of the present study, we calculate coefficient estimates for the effect of lexical variables at various quantiles of the conditional distribution, which allows us to establish the points in time at which a lexical variable reliably exerts an influence on reading times. As for for the EEG data analysis, we analysed neural time-series data using Generalized Additive Mixed Modelling (GAMM), which avoids the analysis of averaged amplitude values for a pre-defined time window and takes the whole length of the epoch for the analysis. By evaluating the difference curve in a GAMM model it can point to the time when two conditions differ from each other, which will allow us to determine when a certain lexical variable starts to influence the amplitude of brain responses.

Our expectations about the time course of effects of whole word frequency and semantic similarity across methodologies are guided by the assumptions of the eye-mind link hypothesis (Engbert, Longtin, & Kliegl, 2002; Reichle, Pollatsek, Fisher, & Rayner, 1998), which is that eye-movements, i.e., a behavioral outcome, is initiated by lexical processing occurring in the brain, i.e., neural activity. Therefore, if reliably present, we expect a lexical effect shown in EEG to predate effects exhibited in eye movements.

2 Methods

2.1 Participants

A total of sixty-six McMaster students with native English speakers participated in the sentence reading experiment (N = 33; mean age = 20.66; 27 female participants) and in the lexical decision experiment (N = 33; mean age = 21.03; 24 female participants). As confirmed by the health screening questionnaire, none of the participants had a head injury, were on medication affecting their central nervous system, or had a history of language, vision or hearing disorders. Six participants self-reported left-handedness, which was further confirmed by the Edinburgh Handedness Inventory Questionnaire (Oldfield, 1971).

2.2 Stimuli

Our target words were borrowed from the pool of stimuli in Schmidtke et al. (2017) and consisted of 160 unique derived words (5-14 letters) with the following 8 suffixes: *-ity, -er, -ness, -ful, -ion, -ive, -ment, -able.* In Experiment 1 (sentence reading with eye-tracking), these words were embedded into sentences with a semantically neutral context before the derived word. The target word did not occur as the first and last word in a sentence (e.g., *The models were very competitive about their waistline size.).* Each sentence consisted of maximum of 16 words and did not occupy more than one line on the computer screen. In Experiment 2 (lexical decision), another 160 pseudo-derived words were added to the 160 derived target words, generating a total of 320 items in the stimulus list. For each derived word, a pseudo-derived word was generated with Wuggy software, version 0.2.0b2 (Keuleers & Brysbaert, 2010) keeping the same number of letters and same set of suffixes as in a real derived word. The full list of materials with word frequency and semantic similarity properties for the two experiments is available in Supplementary materials S1, for the description of variables see below.

In addition, we made use of Author Recognition Test (ART), where participants were presented with the list of 65 author names and 65 foils and asked to identify author names (Acheson, Wells, & MacDonald, 2008). This test taps into individual exposure to print and has been proven to predict reading skills and other reading related variables in native speakers of a language, including spelling ability, word recognition, and reading fluency (Stanovich and West, 1989; West, Stanovich, & Mitchel, 1993; Gordon, Moore, Choi, Lowder, 2020; McCarron & Kuperman, 2021). The score for each subject was calculated as a sum of all correctly identified author names minus foil names wrongly identified as authors. No scores were subtracted when an existing author was not chosen.

2.3 Apparatus and Recording

Eye movements were recorded with the EyeLink 1000 Plus eye-tracker (SR Research Ltd., Kanata, Ontario, Canada) at a sampling rate of 1000 Hz. All experimental stimuli were presented by Experiment Builder version 2.1.140 software (SR Research Ltd., Kanata, Ontario, Canada). The eye movement data was recorded from one eye only, either left of right.

Electrical brain activity of participants reading the stimuli was registered by the BioSemi ActiveTwo system. Participants wore an elastic cap used to attach 64 Ag/AgCl electrodes to the scalp according to an extended 10-20 system. Five additional electrodes were placed externally: two were positioned above and over the outer canthus of the left eye, another three were placed to record activity from the two mastoid processes and from the tip of the nose for potential use during re-referencing offline. All data were recorded at a sampling rate of 512 Hz, referenced online to the driven right leg circuit, and bandpass filtered at 0.01 to 100 Hz.

2.4 Data Synchronization

Data synchronization between signals of EEG and eye-tracking was done with the help of Transistor-Transistor Logic (TTL) triggers sent from the eye-tracker to the continuous EEG recording during stimulus presentation. For both experiments, TTL triggers were sent at the beginning and end of the experiment, and at the beginning and end of each trial presentation. In the sentence reading experiment, an additional trigger was sent when participants crossed an invisible boundary before the word of interest in the sentence. Offline alignment of EEG and eye-tracking data was performed by the EYE-EEG extension of the EEGLAB toolbox (Dimigen et al., 2011; http://www2.hu-berlin.de/eyetracking-eeg/index.php). Analysis of brainwaves (i.e., FRPs) started from the first fixation that landed on the target word after the invisible boundary was crossed.

2.5 EEG Pre-processing

After recording of EEG data, offline signals were re-referenced to the averaged mastoids. The resulting signal was then filtered with a band-pass filter of 0.1-30 Hz. After the synchronization procedure (described above), raw data inspection was performed to remove signals with muscle artifacts. All ocular artifacts (horizontal and vertical), blinks and saccades, were corrected using a procedure optimized for co-registration of EEG and eye-tracking signals using Extended Infomax Independent Component Analysis (ICA) (Meyberg, Sommer, & Dimigen, 2017; Bell & Sejnowski, 1995). To eliminate slow drifts for ICA training and decomposition, the data were further filtered (15-30 Hz) and epoched into segments with a 0.1- 0.5 range around each fixation. The removal of ocular artifact components was confirmed visually and with the help of the EEG-EYE extension.

Epochs with a uniform length 1100 ms (-100 ms to 1000 ms) were cut around the first fixation on target words. The artifact rejection procedure was used to remove any trials with values of -100 or +100 *m*V. Baseline correction was performed using the 100 ms before fixation onset. The resulting data were then down-sampled to 128 Hz (from 512 Hz) to ease the computing load during statistical computations. Epoched data were imported into the R statistical software for further processing and statistical analysis. All trials that were skipped as confirmed by eye-tracking data were removed. Our analysis focused on the first 900 ms of the epoch (-100 ms to 800 ms).

2.6 Procedure

Before the start of either experiment, all participants signed a consent form documenting their willingness to participate. First, participants were instructed to complete a health screening form with their basic demographic information (age, gender, education, and native language), previous medical history (head injuries, psychological disorders, language disorders, and hearing or vision loss), and their current physiological state (current medications, hours of sleep before date of testing, and degree of alertness at the time of testing). Second, all participants performed the Author Recognition Task (ART), see above. Finally, after completing the forms and tests, participants completed either Experiment 1 or Experiment 2. The participants' eye-movements during reading were recorded with the eye-tracker and electrical brain activity was simultaneously registered by EEG.

2.6.1 Experiment 1: Sentence reading. In Experiment 1, participants were shown 160 sentences presented one-by-one, each occupying one line on the screen. Before each trial, a drift correction procedure was performed with a fixation point, i.e., a black dot placed over the first word of a sentence on the left side of the screen. After finishing reading a sentence, participants had to move their eyes to a box in the bottom right corner of the screen and fixate it for 200 ms to initiate next trial. One third of sentences were followed by yes-no comprehension questions. The answers to these were recorded by fixating over 250 ms on respective Yes and No text boxes shown below the question on the screen. Before the beginning of the experiment, participants read 4 practice sentences to familiarize themselves with the flow of the experiment. Each participant in this experiment read the same set of sentences, which were randomized in order individually. Sentences were presented in one line, in black font color against white background using a monospace font, Courier New, size 22.

2.6.2 Experiment 2: Lexical decision. In Experiment 2, participants performed a lexical decision task where they indicated if the character string that they saw in the middle of the screen was a real English word (n = 160) or a non-word (n = 160) by moving their eyes to corresponding text boxes at bottom corners of the screen (Yes in the bottom right corner and No in the bottom left corner). A fixation of at least 200 ms on either text box registered the response and terminated the trial. Each trial began with a drift correction procedure with a black fixation point placed at the middle of the word to appear. The words were presented in the middle of screen one at a time in black color against white background in size 22 Courier New font. Prior to the beginning of the experiment, participants practiced making lexical decisions on 2 words and 3 non-words. Each participant read same set of words and non-words with a randomized order.

During both experiments, participants were seated in a dimly lit room 60 cm away from a 24-inch monitor with screen resolution of 1920x1080 pixels and a refresh rate of 60 Hz. With electrodes placed on their heads, participant's head was stabilized, and head movements were minimized by the chin rest. Before an eye-tracking recording, calibration procedure measured characteristics of a participant's eyes to calculate an eye model for the gaze data to be recorded. Calibration of eye movements was done with the help of a 13-point calibration, during which a participant is asked to look at specific points on the screen. The validation procedure followed immediately to assess the quality of calibration. We aimed for calibration accuracy to fall below 0.5 degrees of visual angle to proceed with testing. All participants could rest between trials, if needed. If participants requested a break, the same calibration-validation procedure was performed after the break.

2.7 Variables

In both experiments, independent variables were the frequency of occurrence for the derived word and semantic similarity between the stem and the whole derived word (e.g., *govern* and *government*). Frequency characteristics of derived words were obtained from SUBTLEX-UK (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), a 200 million token corpus of British films and subtitles. Semantic similarity was operationalized using the Latent Semantic Analysis technique (LSA; Landauer & Dumais, 1997). LSA represents each word as a vector in a multidimensional space based on co-occurrence statistics of that word with select other words (labeled "factors") in the language. The estimate of semantic similarity between any two words that those vectors represent is estimated by the cosine of the angle between the vectors, ranging from -1 to 1. The LSA scores for semantic similarity between the stem and the whole word were collected from http://meshugga.ugent.be/snaut-english/, with a default setting of 300 factors and a window of 6 words (Mandera, Keuleers, & Brysbaert, 2017). Values closer to 1 imply a greater semantic dissimilarity between the stem and the whole word, lower values represent higher estimated levels of similarity. The same SUBTLEX-UK corpus of film subtitles (van Heuven et al., 2014) was used for the LSA score calculation as for frequency estimates.

The 2 x 2 orthogonal design of both Experiments 1 and 2 was achieved by selecting stimuli that fell dichotomously into a high/low-frequency and high/low semantic similarity experimental cells. We started the stimulus selection procedure with a pool of 504 derived words. First, all words were divided into high/low frequency groups using a median split. In each of the resulting groups we identified words falling into the lower and upper quartiles of semantic similarity. Forty words were randomly chosen from each pool of words that satisfied the frequency and similarity selection criteria, to a total of 160. T-tests were conducted to show that frequency differed significantly ($t(156.79) = -15.311$, $p < .001$) between the high and low frequency groups, but semantic similarity did not ($p = .126$). Similarly, a significant difference in semantic similarity values was observed between the high and low semantic similarity groups (*t*(79.237) = -3.162, *p* = 0.002), but group means of frequency values did not differ (*p* = .189).

The critical dependent variables for the eye-tracking data analysis in Experiment 1 were measures of first pass reading: first fixation duration (the duration of the first fixation on a word) and gaze duration (the summed duration of all fixations on a word before leaving that word for the first time), all measured in milliseconds. These two measures were chosen, as they tap into early processing of a word and are most appropriate for the goal of our study is to find the earliest time point when an effect has a discernible impact on the eye-movement record. For EEG, the critical dependent variable was the mean amplitude value for each 8 ms window of the target word epoch. The whole epoch from -100 to 800 ms was included into analysis to find how early the waveform amplitudes for two conditions (formed by either frequency or semantical similarity) diverge.

2.8 Statistical Considerations

2.8.1 Quantile regression analysis of eye fixation durations. For the analysis of eye-movement data, our aim was to discern when the two effects of interest, whole word frequency and semantic similarity, have the earliest impact on fixation durations on target words during the sentence reading task (Experiment 1) or the lexical decision task (Experiment 2). The standard least squares linear regression analysis is a commonly used statistical technique that evaluates the average effect of a variable on a response. However, this method is suboptimal if the research interest lies in estimating the influence of an independent variable on the specified point or range of the response variable distribution. To answer this question, we applied *quantile regression* (Koenker & Bassett, 1978) to eye-movement data. This method can trace the earliest statistically detectable emergence of a variable's effect in response latencies. An alternative distributional analysis technique, non-parametric survival analysis (Reingold & Sheridan, 2014), was used by Schmidtke et al. (2017) and Schmidtke and Kuperman (2019) in their studies of the processing time-course of complex words. Recent criticism of survival analysis by Gómez, Breithaupt, Perea, & Rouder (2020) proposes that this technique has methodological and conceptual weaknesses, which could lead to misinterpretation. While we address this proposal elsewhere, for the present purposes we make use of quantile regression for which the weaknesses suggested for survival analysis are not problematic.

The quantile regression approach estimates the effect of the independent variable on the dependent variable at different quantiles (e.g., 10th, 20th, 30th), or τ, of the dependent variable's conditional distribution. Where, a standard linear regression focuses on the average of effect and may hide important features of the underlying relationship, quantile regression may be used to test whether effects differ across points of the distribution. A key property of quantile regression models is that they assume neither a normal parametric distribution for the response variable nor a constant variance of the response. In the present study, we use quantile regression to model fixation durations (see also Tiffin-Richards & Schroeder, 2020). Estimating the effect of lexical variables at various quantiles of the response time distribution allows us to examine in fine detail how a lexical effect unfolds in time, from the shortest fixation durations, i.e., lower quantiles, to the longest fixation durations, i.e., higher quantiles. Effects of either semantic similarity or word frequency in lower quantiles, representing the left tail of the fixation duration distribution, would indicate that lexical properties of the whole word form, including semantics, emerge early. We tested lexical effects at each decile of the response time distribution between the 10th and 90th quantile, e.g., 10th, 20th, 30th … 90th. The earliest decile for which we observed a lexical effect, should any emerge, was taken as an estimate of the onset of the lexical effect in the behavioral record. We pit these onsets against those same effects, observed for the same participants and stimuli, in the neural record.

We fitted mixed-effects quantile regression models to first fixation durations and gaze durations from the sentence reading and the lexical decision tasks. Analyses were conducted using the `lqmm()` function in the *lqmm* package for R (Geraci, 2014). The same model formula was fitted to all outcome variables. The model included semantic similarity and whole word frequency as fixed effects. The current version of the *lqmm* package permits the inclusion of one random effect, which we set to subject id, to take into account variability between participants. Statistical inference for the model parameters was performed using bootstrap resampling of data using the `summary()` based on resampling the sample data in the *lqmm* package. We set the number of bootstrap samples to 1000.

2.8.2 Generalized Additive Mixed Models (GAMMs). Eye-tracking data consists of a single value (fixation duration) for a given item (target word), whereas EEG data for an item has multiple values across many electrodes consisting of amplitude measures, the number of which is dependent on the length of a time window for analysis and the sampling rate. Thus, the difference in the nature of data obtained from two experimental methods motivated this study to explore alternative ways of EEG signal analysis.

The majority of EEG studies have utilized traditional statistical methods of analysis, such as averaging or ANOVA. In fact, there are several limitations of traditional averaging techniques and analyzing the mean differences using repeated measures ANOVA (for limitations of these methods, see Baayen, 2004). Taking all the limitations of traditional methods of EEG analysis into account, we apply a more advanced statistical technique, namely, Generalized Additive Mixed Modelling (henceforth, GAMM) (Hastie & Tibshirani, 1990; Wood, 2017). GAMM is an extension of generalized linear mixed model (GLMM) with nonparametric terms (nonlinear smoothing functions). First, one of its strongest advantages is the ability to identify non-linear effects, which is inherent to the nature of EEG data as EEG amplitude varies in a non-linear way over time. Nonlinear effects of predictors are modelled with functions called smooth terms, and nonlinear interactions between predictors are modelled with tensor products (for details, see Wood, 2017). Second, it considers the entire epoch for analysis, whereas in traditional methods, a window of analysis has to be defined. Detailed and well-illustrated examples comparing traditional averaging methods and GAMM can be found in Tremblay (2009). Third, GAMM makes a distinction between fixed and random-effect variables, and captures the dependencies between repeated measurements, within or between subjects and stimuli, thus, modelling a more complex random effect structure. Finally, GAMM can take into account the problem of autocorrelated residual errors, which is a common feature of a time series data as the amplitude value at a given time point is highly correlated with the value of amplitude at the next time point. Previously, this method has been successfully applied to EEG data in recent studies (De Cat, Klepousniotou, & Baayen, 2015; Porretta, Tremblay, & Bolger, 2017a; Tremblay & Baayen, 2010; Tremblay & Newman, 2015). The analysis was performed in R using *mgcv* (Wood, 2017) and *itsadug* (van Rij, Wieling, Baayen, & van Rijn, 2015) packages.

2.8.2.1 Modelling EEG data with GAMMs. We focused our EEG analysis on FRPs, signals time-locked to the onset of the first fixation on the target word. Each FRP epoch length was 900 ms, ranging from -100 ms to 800 ms after first fixation onset. Several regions of interest (ROI) were identified for word frequency and semantic similarity variables. The following ROIs were explored for the two effects of interest: two frontal (AF3, AFz, AF4 and Fz, F2, F4), two fronto-central (FCz, FC2, FC4 and C1, Cz, C2), two centro-parietal (CP1, CPz, CP2 and P1, Pz, P2) and two parieto-occipital (PO3, POz, PO4 and O1, Oz, O2).

The GAMM models for investigating the change in the EEG amplitude over time across factorial conditions used thin plate regression spline smooths (Wood, 2017). This is especially useful to model nonlinear dependencies with a single predictor by means of a weighted sum of smooth regular basis functions, which are better than simple powers (e.g., higher order polynomials). Treatment coding of the smooth terms for factorial predictors was used to model contrasts in the EEG amplitude across conditions with two levels, high and low. In what follows, for example, the smooth term for ordered factorial predictor, e.g., low word frequency, represented the difference with the reference level, high word frequency. Random intercepts were included in the models for items (target derived words) to allow cross-item fluctuations in baseline amplitude. Participant was added as another random-effect factor by means of including a nonlinear factor smooth for Time, which modeled the development of EEG amplitude over time. Moreover, since each participant could show a different pattern in EEG amplitude over the course of experiment for each item, we included a non-linear factor smooth for Trial as well (for similar random structure, see Baayen, van Rij, de Cat, & Wood (2016)). By including these random variables into model, we improved the model fit substantially. The autocorrelative structure in the residual error was removed by including the AR-1 autocorrelation parameter (ρ) into GAMM models, which was based on an initial estimate of the model fit without including the autocorrelation parameter (Pinheiro & Bates, 2000). The ρ value for each model is reported in the description of the results.

Due to substantial amplitude differences between ROIs, we refrained from fitting a single GAMM model of the full dataset. Instead, a total of 8 models (one for each ROI) were fitted for the FRP analyses for each of the two effects separately. One limitation of GAMMs is that it can only fit interactions of smooths with a single factor. Consequently, the inclusion of both of our factorial predictors, word frequency and semantic similarity, into one model could not provide the exploration of the differences between the two levels of a factor, e.g., high and low word frequency. *P*-values were Bonferroni-corrected as we performed multiple comparisons in our analysis (exploration of the two effects in 8 ROIs simultaneously: 8 for word frequency and 8 for semantic similarity). Consequently, the result was considered reliable only if the significance level was below 0.003 (0.05/16 = .003).

The mean amplitude value for each ROI was calculated for each time point and used as the continuous response variable in GAMM models. Critical predictors were whole word frequency or semantic similarity (with two levels: high and low).

3 Results and Discussion

3.1 Sentence Reading Experiment

3.1.1 Eye-tracking results. Three out of thirty-three participants were removed from the data analysis. One participant had most trials missing in the eye-tracking data; to match the participants across each EEG and eye-tracking methods, another two were removed due to synchronization and poor EEG acquisition. After participant removal, we had 4800 observations in total (30 participants x 160 target words). One trial was lost due to a programing error, and 8% of targets were skipped by our readers during experiment reading, leaving 4605 observations. Further, we trimmed our data so that very short (< 80 ms) and very long fixations (>1500 ms) were also excluded from analysis. A total of 4236 trials (or 88.3% of the original data pool) were included in analysis. The mean comprehension rate was 90.3% showing that participants had a good understanding of the material read during the experiment. For the eye-tracking data, we report effects for both first fixation duration and gaze duration. For the sentence reading experiment, 72% of critical words were read in one fixation. For the lexical decision experiment, 22% of trials were read in one fixation. The difference in the number of fixations on the target word between two experiments may be attributed to the nature of experimental design: similar reports were found by Kuperman, Schreuder, Bertram, & Baayen (2009), who also used a visual lexical decision experimental task. Table 1 below provides descriptive statistics for all dependent and independent variables.

Table 1: Descriptive statistics for dependent and independent variables used in Experiment 1

	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Whole word frequency	1	22	71	299.5	196	14266
Semantic similarity	0.000	0.485	0.663	0.627	0.777	0.943
Word length	5	7	9	8.88	10	14
First Fixation Duration	81	183	225	245.7	282	815
Gaze Duration	81	200	266.5	312.7	383	1457
Total Reading Time	81	244	398	477.7	632.2	1493

3.1.1.1 First Fixation Duration Analysis. Figure 1 below presents regression estimates of two effects, word frequency (top panel) and semantic similarity (bottom panel) at each decile of the first fixation duration distribution. As Figure 1 (top panel) shows, the whole word frequency effect began to significantly influence inspection times already at the second decile of the distribution (τ = 0.2, β = -5.578, *p* = .009) which corresponds to the onset of 175 ms. As expected, high frequency words were read faster than low frequency words. The word frequency effect gradually increased in magnitude towards the right tail of the first fixation duration distribution, reaching a 15 ms contrast in the longest first fixation duration. Figure 1 (bottom panel) visualizes estimated coefficients for the semantic similarity effect. The effect was not significant in any decile. See Table 1 of Supplementary materials (available online) for the detailed description of the model summary output.

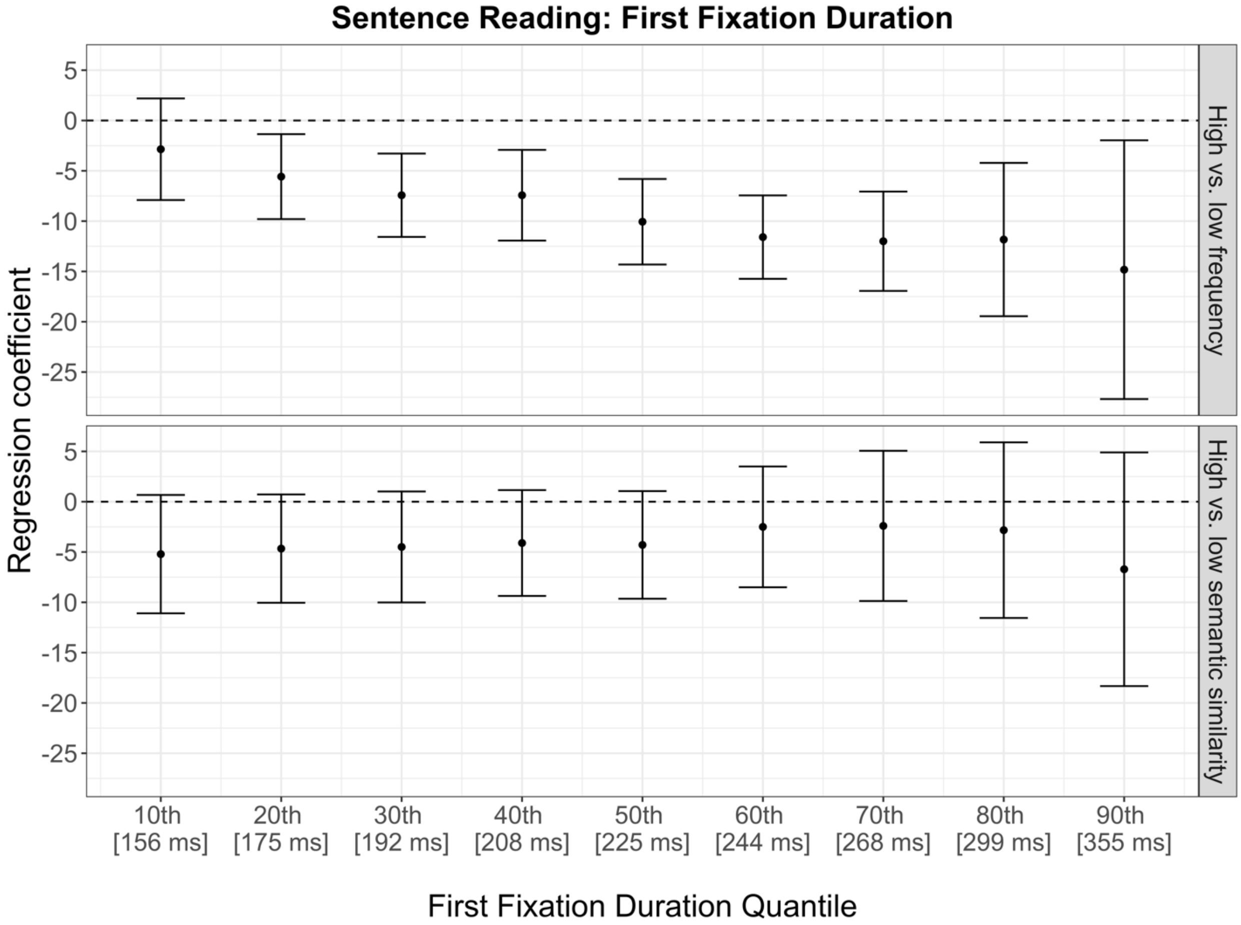

Figure 1: Estimated quantile regression effects of frequency (top) and semantic similarity (bottom) on first fixation durations in the sentence reading experiment. Error bars represent 95% confidence intervals.

3.1.1.2 Gaze Duration Analysis. We further analyzed the timeline of lexical effects using another measure of first-pass reading, namely gaze duration. Figure 2 below presents the effects of word frequency and semantic similarity on gaze duration distribution, ranging from 164 ms (10th percentile) to 524 ms (90th percentile). Similar to first fixation duration, the frequency effect – the speed advantage to high rather than low frequency words -- is reliably detected from the 20th percentile, or 189 ms ($\tau = 0.2$, $\beta = -7.144$, $p = .015$) and remains reliable until the end of the distribution. As for the semantic similarity, the effect showed marginal significance in the lower subrange of the response latency distribution, only at the second decile ($\tau = 0.2$, $\beta = -5.283$, $p = 0.064$), at 189 ms. The effect got substantially stronger and statistically significant towards the end of the distribution starting from the 70th percentile, 355 ms ($\tau = 0.7$, $\beta = -13.227$, $p = .029$), to the 90th percentile (523 ms). The magnitude of the effect reached 30 ms contrast for frequency and 45 ms for semantic similarity towards the right tail of distribution. Detailed results of the model output are found in Table 2 of Supplementary materials.

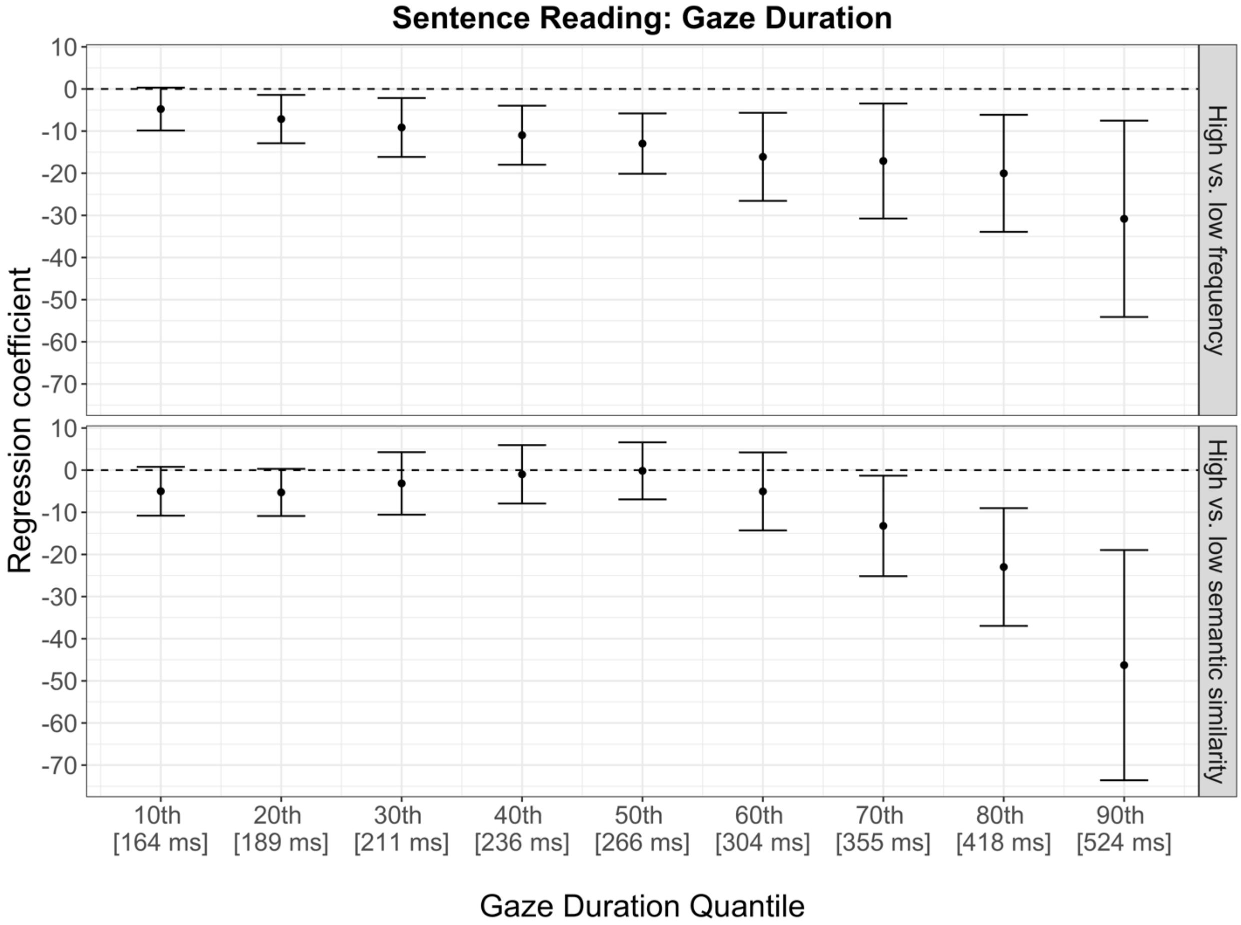

Figure 2**:** Estimated quantile regression effects of frequency (top) and semantic similarity (bottom) on gaze durations in the sentence reading experiment. Error bars represent 95% confidence intervals.

3.1.2 FRP results. We analyzed the 30 participants that had both EEG and eye-tracking data. 5.8% percent of data was lost due to missed markers for some of the target words. In total, 4520 trials were created upon epoch generation. Further, the artifact rejection procedure removed another 6.1% of data. Another 7.8% of epochs were removed due to participants' oculomotor skipping of the critical target words, leaving us with a 3845 trials (about 80%) of original data for the FRP analysis.

None of the regions showed a statistically significant difference on EEG amplitude between frequency conditions at any time of the epoch length (all *p*s > .130). Similarly, semantic similarity did not show any reliable effect on EEG amplitude in any of the ROIs and across the entire epoch (all *p*s > .189).

To summarize the results for sentence reading of derived words, we observed an onset of the frequency effect at 175 ms after first fixation landed on the word and the onset of the semantic similarity effect at 355 ms during derived word recognition. FRP analysis of EEG data, on the other hand, did not show any amplitude differences between high and low frequency and high and low semantic similarity conditions.

3.2 Lexical Decision Experiment

3.2.1 Eye-tracking results. Two participants were removed due to poor EEG recording. Trials with no response, where fixations were not received at target word area by participants (155 trials), and incorrect answers (424 observations) were excluded from the data as well. Average correct response rate across all participants was very high, 95.7%. After removing responses to pseudo-words, we were left with 4769 trials with target derived words. 147 (3.1%) out of these trials, were skipped (despite recorded lexical decision responses, targets in these trials have not received any fixations). Additionally, we excluded too short fixations (< 80 ms) for first fixation duration measure and too short (< 300 ms) for total reading time measure on a word. Exceedingly long (the top 1% of the entire fixation distributions for first fixation duration and total fixation time measures) fixations were removed as well. Removing short and long fixations excluded 831 trials in total, or near 18%. Overall, we were left with 3791 observations (or about 80% of the original data pool). Table 2 below provides descriptive statistics for all dependent and independent variables.

Table 2: Descriptive statistics for dependent and independent variables used in Experiment 2

	1st Quantile	Min	Median	Mean	3rd Quantile	Max
Whole word frequency	1	24	73	312.3	206	14266
Semantic similarity	0.000	0.479	0.591	0.624	0.777	0.943
Word length	5	7	9	8.88	10	14
Trial dwell time	307	1190	1343	1334	1487	3049
First Fixation Duration	81	373	513	535.4	725	1209
Gaze Duration	303	765	855	885	973	2016
Total reading time	307	768.5	860	892.6	979	2016

3.2.1.1 First Fixation Duration. Quantile regression analysis of first fixation durations did not show frequency effect (for the discussion, see General Discussion). Effect of semantic similarity was evident in the data only in the ninth decile, or at 861 ms (see Supplementary materials, Table 3).

3.2.1.2 Gaze Duration. More reliable traces of the effects were observed in the gaze duration measure. Figure 3 below demonstrates word frequency and semantic similarity effects on gaze duration. Estimated coefficients for the word frequency effect were sufficiently strong to reach statistical significance as early as the second decile of the gaze duration distribution, which corresponded to 741 ms (Figure 3, top panel). As expected, fixation times for the high frequency words were shorter than for the low frequency words. The processing advantage for high frequency words shown in the 20th percentile (22 ms) increased throughout the remainder of the response latency distribution and accumulated to a 63 ms advantage in the 90th percentile as seen in the mean values for regression coefficients for each decile (y-axis of Figure 3).

Figure 3 (bottom panel) shows that semantic similarity gradually increases in its effect size and reaches statistical significance at the fifth decile (855 ms). Words in which the stem and the whole derived word meanings are highly similar are faster to read than words of low semantic similarity. Detailed results with the model output can be found in Table 4 of Supplementary materials.

Comparing the above results with the sentence reading experiment, we observed that when words were read in context, the effects of whole word frequency and semantic similarity appear to influence reading times much earlier (captured by fixation measures that register early processing: 175 ms for first fixation duration and 355 ms for gaze duration, respectively) than when the words are read in isolation, in lexical decision experiment (741 ms and 855 ms both in the gaze duration measure), by at least 500 ms. This same difference was observed by Schmidtke et al. (2017) with lexical decision effects being later. This is not surprising, as in the lexical decision experiment the initial fixation point is located in the middle of the word, which specifically supresses eye-movements and removes the possibility to track the time-course of processing over multiple fixations. Also, lexical decision involves a meta-linguistic judgment which makes responses longer. Finally, the natural presentation of the entire sentence enables a parafoveal preview of upcoming words which speeds up the recognition process when the word is landed upon. The massive difference between the tasks, as witnessed in the eye-movement record, highlights the importance of using naturalistic rather than artificially constructed tasks for reading research (Liversedge, Blythe, & Drieghe, 2012; Rayner & Sereno, 1994).

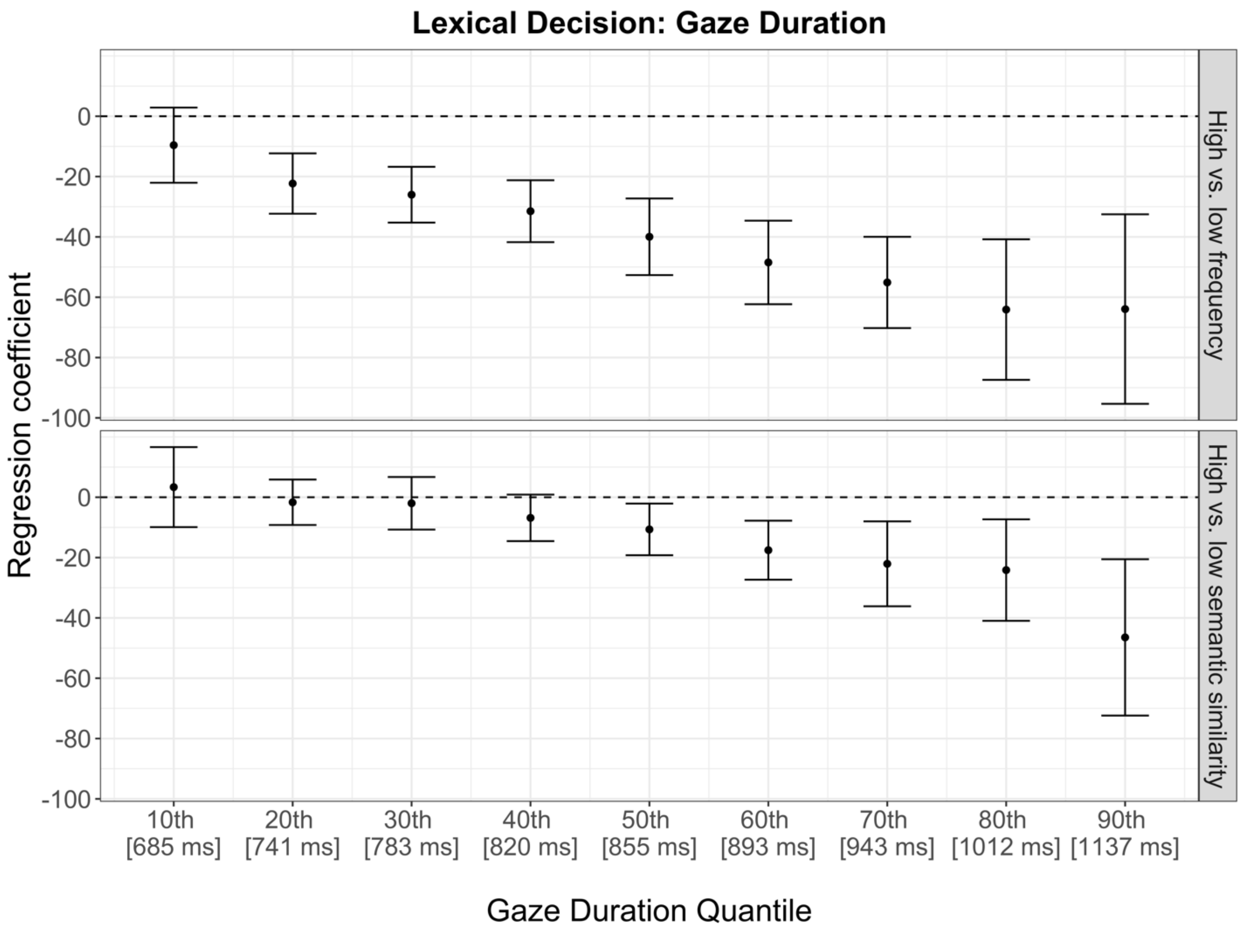

Figure 3. Estimated quantile regression effects of frequency (top) and semantic similarity (bottom) on gaze durations in the lexical decision experiment. Error bars represent 95% confidence intervals.

3.2.2 FRP results. Two participants were removed from electroencephalographical data due to poor recording and slow drifts. 9989 epochs were generated upon epoch creation losing 5.4% of trials because of missing triggers in the recording. A total of 500 trials were rejected by the artifact rejection procedure, and in another 212 trials participants did not fixate the area where the target word was located when performing a lexical decision task during the experiment. In total, we had 9277 epochs before removing non-word trials, or 87.9% of data.

3.2.2.1 Frequency effects. Out of 8 ROIs investigated, only the occipital region (O1, Oz and O2 electrodes) showed a statistically significant difference in amplitude values for high and low frequency conditions as per the fitted model. The grand average plot of this region showed a visible amplitude difference between these two types of words starting at about 400 ms post fixation (see Appendix, Figure A1). The low-frequency group showed a greater negative inflection compared to the high frequency group. However, after accounting for all random effects and correcting for autocorrelation, the difference curve evaluated by the GAMM model only shows statistical significance between conditions after 500 ms. Detailed results of the summary output of the model fitted to the EEG amplitude of averaged values at occipital region can be found in Table 3 below. Nevertheless, note that after the Bonferroni correction, this result was no longer deemed significant (with a threshold for $p < 0.003$). Despite this insignificance, these results are presented here to support our claims in the General Discussion with regard to best methods for examining the relationship between eye-movements and neurophysiology.

Table 3: Summary of generalized additive mixed model for frequency effect in occipital region in lexical decision experiment.

A. **parameter coefficients**	Estimate	SE	*t*-value	*p*-value
Intercept	0.027	0.518	1.052	0.958
Frequency (low)	-0.235	0.171	-1.371	0.170
B. **smooth terms**	Edf	Ref.df	F-value	p-value
Time	8.377	8.509	6.478	<0.0001
Time x Frequency (low)	1.010	1.019	4.645	0.0307
Item	92.932	158	1.419	<0.0001
Trial x Subject	75.152	278	0.593	<0.0001
Time x Subject	233.297	278	8.852	<0.0001

Note: N = 508990. Treatment coding was used for Frequency, the two-level factor, with Time x Frequency (low) representing the difference curve. ρ value = .95.

Parameter coefficients (Part A) indicate parametric estimates of the model. It presents the intercept showing the mean amplitude for the low frequency condition that is shifted down in the negative direction by .24 microvolts. Critically, the model did not show any significant difference in mean amplitudes for high and low frequency words. Part B of the table reports smooth terms including the thin plate regression spline smooths for the change of the amplitude over time for high frequency (first row), the nonlinear interaction of frequency by Time (second row), and random effect structure (last three rows). Edf stands for effective degrees of freedom, where smooths with higher edf tend to be wigglier, and Ref.df stands for reference degrees of freedom. The second smooth evaluates the difference curve, and shows that the waveform trends over time for frequency effect in the occipital region is significant (*F*(1.010) = 4.645, *p* = 0.031). As can be seen from the difference curve in Figure 4, the time window of significant difference for the two frequency conditions begins at 490 ms and is sustained until 769 ms. Similar to previous findings on words read in isolation, low frequency words exhibited increased negative going amplitudes around 400 ms (e.g., Hauk & Pulvermüller, 2004). However, again, after the Bonferroni correction, this result was no longer significant.

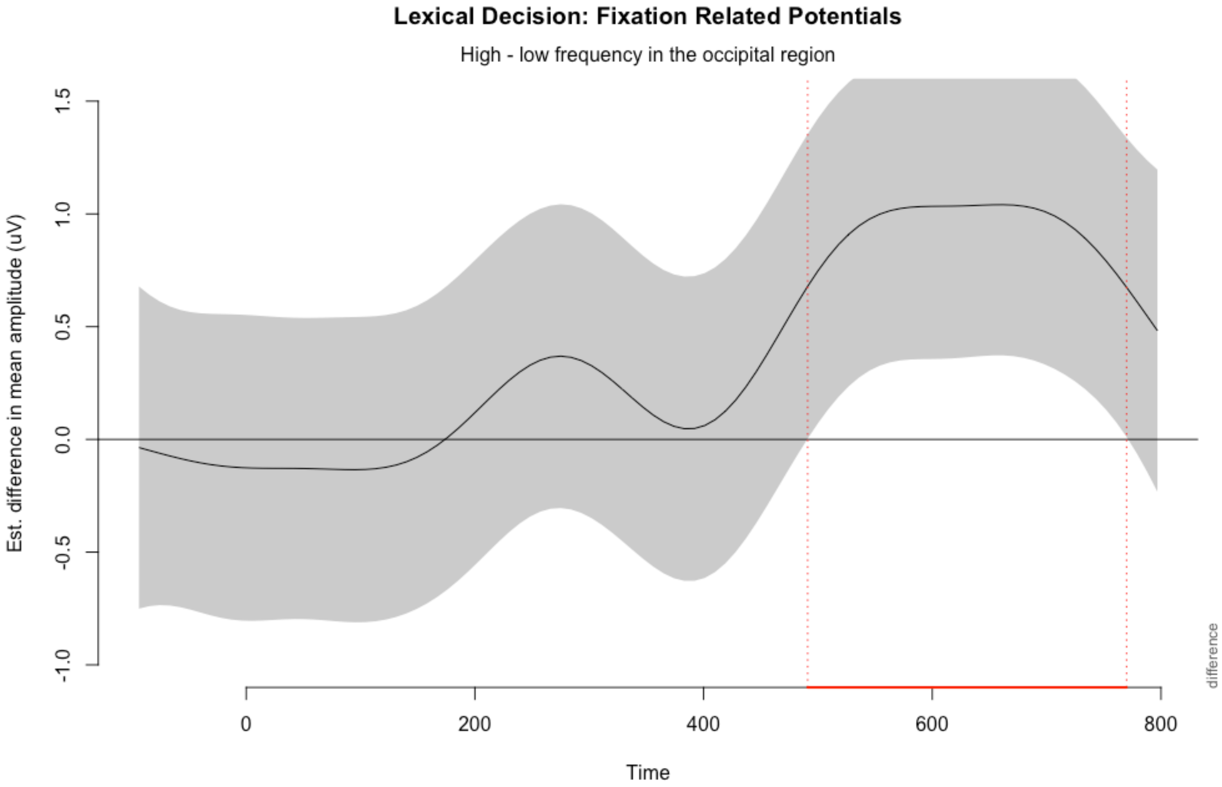

Figure 4. Difference curve plot for fixation related potentials in the occipital region based on predictions of the GAMM model. The difference between high frequency and low frequency words is plotted as a function of time. The area between the two vertical dashed lines represents the time window in which significant differences between each condition are observed.

3.2.2.2 Semantic similarity effects. GAMM models revealed that four ROIs -- centro-parietal, parietal, parieto-occipital and occipital -- showed that the waveforms for high and low semantic similarity conditions differ significantly over time (see grand averaged plots for the raw data in the Appendix, Figure A2). Due to space limitations, we only present detailed results of the model at the centro-parietal region. Summary outputs of other 3 models can be found in the Supplementary materials (available online). The summary results of the fitted model for the centro-parietal ROI can be found in Table 4 below.

Table 4: Summary of generalized additive mixed model for the semantic similarity effect in the centro-parietal region in the lexical decision experiment.

C. **parameter coefficients**	Estimate	SE	*t*-value	*p*-value
Intercept	0.241	0.716	0.337	0.736
Semantic similarity (low)	0.623	0.314	1.984	0.047
D. **smooth terms**	Edf	Ref.df	F-value	p-value
Time	8.851	8.885	39.135	<0.0001
Time x semantic similarity (low)	1.006	1.012	10.373	0.001
Item	103.387	158	1.888	<0.0001
Trial x Subject	87.262	278	0.719	<0.0001
Time x Subject	233.169	278	8.805	<0.0001

Note: N = 508990. Treatment coding was used for the two-level factor for semantic similarity with Time x Semantic Similarity (low) representing the difference curve. ρ value = .95.

The intercept in part A of the model output represents the mean amplitude for the high semantic similarity and low semantic similarity condition. The mean amplitude for the low semantic similarity condition differed significantly and shifted up by 0.62 microvolts in a positive direction. Time x Semantic similarity interaction evaluated the difference curve with respect to high semantic similarity condition, *F*(1.006) = 10.37, *p* = .001. The summary indicates that there is a significant difference between high and low semantic similarity conditions. As illustrated in Figure 5, a significant difference between each condition begins to emerge at 365 ms and is sustained throughout the remainder of the epoch.

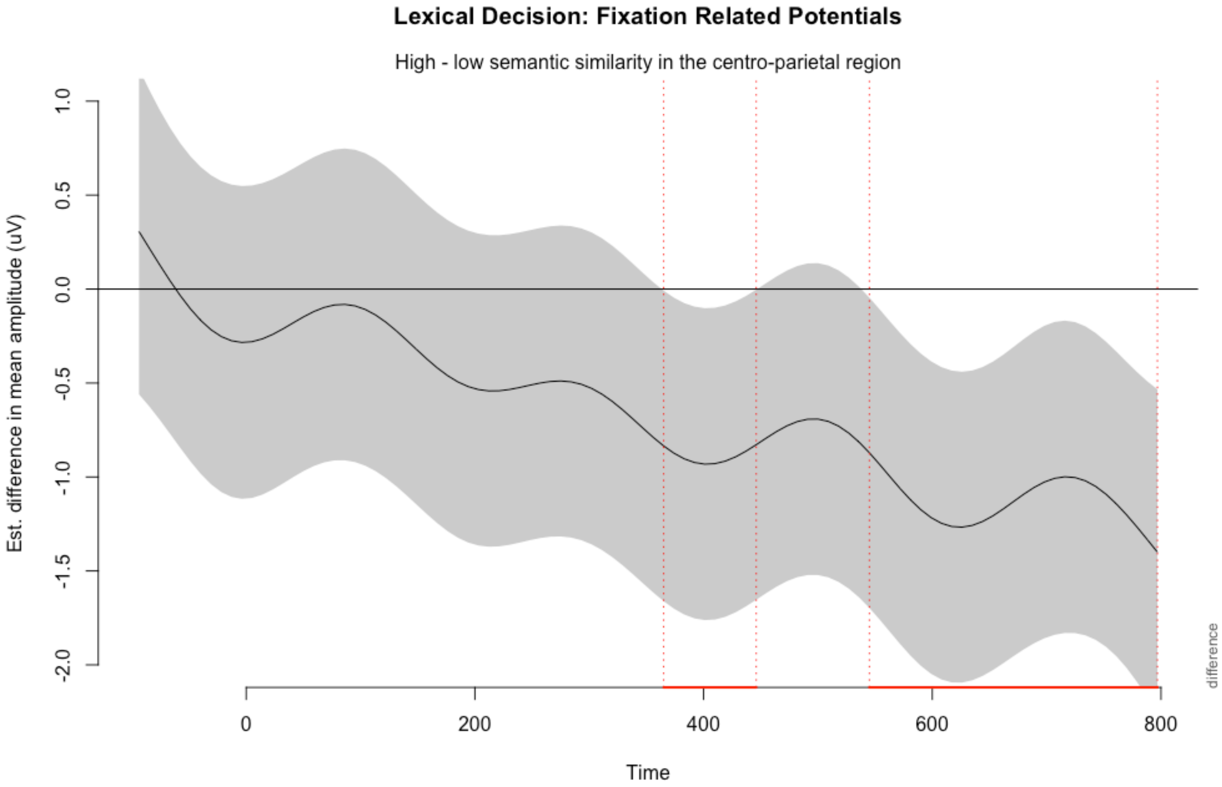

Figure 5. Difference curve plot for fixation related potentials in the centro-parietal region based on predictions of the GAMM model. The difference between words of high and low semantic similarity is plotted as a function of time. The area between the two vertical dashed lines represents the time window in which significant differences between each condition are observed.

To conclude, in the lexical decision experiment, the effect of frequency was found as early as 741 ms when analyzing the quantiles of gaze duration distributions, whereas FRP analysis did not reveal any statistically significant difference between high and low frequency conditions in any of the 8 regions explored. As for semantic similarity, FRP analysis revealed that the semantic similarity effect begins at 365 ms. Eye-tracking, on the other hand, showed the difference only at 855 ms, e.g., 490 ms later. Table 5 below summarizes the main findings of Experiments 1 (sentence reading) and 2 (lexical decision).

Table 5: Summary of significant divergence points in time for frequency and semantic similarity effects as shown by EEG and eye-tracking

Effect	Sentence reading experiment (Exp 1)		Lexical decision experiment (Exp 2)	
	EEG	Eye-tracking	EEG	Eye-tracking
Word frequency	none	175 ms	none	741 ms
Semantic similarity	none	355 ms	365 ms	855 ms

4 General Discussion

An examination of existing neurophysiological (EEG/MEG) and behavioral (reaction time, eye-tracking) studies on complex word processing reveals an apparent logical implausibility in the time-course of lexical effects during the recognition of morphologically complex words. That is, the earliest onsets of lexical effects in the behavioral record appear to predate those that are independently found in neural activity. This apparent paradox violates a core assumption of the eye-mind link hypothesis, which is that a response to a stimulus must first occur (and be registered) in the brain before an associated behavioural response is initiated. To address this paradox, this study recorded EEG and eye-movements simultaneously while participants read derived words that were presented in a sentence reading task or in a lexical decision task. We explored two well-established effects which are typically indicative of lexical access of complex words: whole word frequency and semantic similarity. Detection of onsets required the use of statistical methods based on entire temporal distributions of responses rather than the means. To this end, quantile regression and generalized additive mixed modelling were implemented to provide estimates of the earliest onsets of lexical effects in behavioral and neural activity, respectively.

Below we discuss these findings by Experiment (i.e., task) and follow up with a discussion of methodological implications of the findings. In both experiments, where observed, the effects of two critical variables – word frequency and semantic similarity – were in the expected direction. Behaviorally, more frequent derived words were processed faster, as were the words with higher values of semantic similarity (base is more similar to the whole word). These findings align with an existing body of research on complex word processing, see the Introduction. In the one instance of a reliable effect of a critical variable on the neurophysiological record, lower semantic similarity came with a greater negative deflection than its high-similarity counterpart. This result converges with earlier reports of semantic similarity effects on the EEG/MEG record, see the Introduction.

A more puzzling set of findings pertains to the focus of the paper – the temporal emergence and relative time-course of critical effects across methodologies. The results of the sentence reading task (Experiment 1) illustrates a paradoxical situation where behavioral manifestations of lexical processing do not appear to be preceded by neural activity at all. While the effect of derived word frequency on eye-movements reliably emerged at 175 ms post-onset of the fixation (20th percentile of the first fixation duration distribution), we did not observe a reliable counterpart in the EEG signal either before or after that timepoint. It is noteworthy that the temporal estimate of the effect onset is in line with prior reports obtained from sentence reading data on derived words (150 –169 ms, see Schmidtke et al., 2017). Likewise, an effect of semantic similarity was observed as reliable in eye movements at 355 ms (70th percentile in gaze duration distribution) after the beginning of the fixation (and as early as 164-189 ms post-onset, in the 10th-20th percentiles, with marginal significance, see Figure 2). Yet, no parallel effect was detected in the brain activity.

Lexical decision data in Experiment 2 somewhat qualifies the conclusions drawn above. The word frequency effect was also not found in the EEG signal while emerging reliably in the eye-movement record (746 ms, 20th percentile of the gaze duration distribution). Conversely, the semantic similarity contrast showed an effect on the brain activity relatively early (365 ms post-onset of fixation), i.e., much earlier than in the eye-movement record (860 ms, 50th percentile of the gaze duration distribution). This relative order – neural activation followed by a behavioral expression – is indeed as expected under the eye-mind hypothesis. This suggests that the EEG method has sufficient sensitivity to detect semantic effects in isolated word recognition earlier than behavioral methods do.

Several aspects of the present findings call for further discussion. One is the absence of any word frequency effect in the EEG analyses of either sentence reading or lexical decision data. The null frequency effect in sentence reading is not novel. For example, Kretzschmar et al. (2015) and Degno et al. (2019) each found null effects of word frequency in the EEG record during natural sentence reading. As an explanation for the null effect of word frequency, Kretzschmar et al. (2015) argue that the expected N400 effect arises only when bottom-up information in the input does not match the generated predictions from preceding context (e.g., "*The bill was due at the end of the <u>hour</u>*"). However, unlike the present study, earlier neurophysiological studies have reported robust frequency effects in lexical decision and other isolated word recognition tasks (Dambacher et al., 2006; Hauk et al., 2006; Norris et al., 2006; Sereno et al., 1998). There are several possibilities for the null effect of word frequency on EEG in lexical decision (Experiment 2). It is possible that the contrast between high and low frequency conditions was not strong enough to elicit differential brain activation (though it was sufficient to differentially affect fixation durations). Similarly, an increase in statistical power (e.g., a greater number of participants) might be necessary to bring out reliable effects of frequency in the neural record (even though the power of the present study was sufficient for reliable detection of effects in the eye-movement record). Additional co-registration studies are needed to corroborate these results. Another possibility discussed below is that brain responses are not phase-locked and thus do not provide a detectable temporal signature.

Taken together, the present findings suggest that neurophysiological experimental paradigms show merit in studying the time-course of processes involved in isolated word recognition, as they can detect some of critical effects (i.e., semantic similarity) approximately 500 ms earlier than one can derive from a distributional analysis of eye-movements. An obvious reason why eye-movements are not optimal for paradigms with isolated word recognition, including lexical decision, is that such paradigms specifically suppress saccade generation as a central component of oculomotor behavior. Namely, the paradigm in which a fixation point appears in the centre of the screen and is replaced by a target word appearing in the same position makes redundant saccadic movements that are typical in reading behavior. As a result, the oculomotor lexical decision task generates unusually long fixations (median first fixation duration 855 ms vs 200-250 ms typical for continuous reading) and infrequent within-word saccades. Since distributional analyses of eye movements are necessarily based on the timepoints when fixations terminate (and saccades begin), single-word presentation tasks artificially inflate the estimates of lexical onsets that such analyses are geared to produce. To illustrate, the effect of word frequency of gaze duration emerges equally early in the viewing time distribution of both sentence reading and lexical decision task (20th percentile), yet in the lexical decision data it emerges late in absolute terms (746 ms post-onset of fixation on the word). Thus, as far as isolated word recognition is concerned, eye-movements may be ill-suited to detect the earliest possible onset of critical effects due to the artificial demands of the task on the visuo-oculomotor system.

Yet isolated word recognition is not equivalent to reading, a distinction that has been highlighted in an extensive body of work. Lexical decision involves a decision component that is not present in natural reading which leads to differences in the effect sizes of lexical variables, such as word frequency (Kuperman et al., 2013; Liversedge et al., 2012; Rayner, 1998; Schilling, Rayner, & Chumbley, 1998). In contrast, the goal of sentence reading is to comprehend larger chunks of connected text. That is, sentence reading incorporates the full gamut of interacting cognitive processing operations (e.g., LaBerge & Samuels, 1974; Mézière, Yu, Reichle, von der Malsberg, McArthur, 2021; Perfetti & Stafura, 2014; Staub & Rayner, 2007), including low-level "bottom-up" perceptual processing (Balota, Pollatsek, & Rayner, 1985; Kirkby, Webster, Blythe, & Liversedge, 2008; McConkie & Rayner, 1975; Rayner & McConkie, 1976; Rayner, Sereno & Raney, 1996; Vergilino-Perez, Collins & Dore-Mazars, 2004), oculomotor programming (Drieghe, Brysbaert, Desmet, & De Baecke, 2004; Engbert, Nuthmann, Richter, Kliegl, Swift, 2005; Inhoff, Kim, & Radach, 2019; Rayner & Morrison, 1981; Vitu, McConkie, Kerr & O'Regan, 2001), and high-level processes required to integrate semantic representations into larger discourse representations (Berkum, Hagoort, & Brown, 1999; Clifton, Staub, & Rayner, 2007; Kintsch & Walter Kintsch, 1998; Molinaro, Conrad, Barber, & Carreiras, 2010). In all of these respects, the sentence reading task – though not ideal – is expected to elicit a much closer approximation of the perceptual, oculomotor and cognitive processes involved in natural reading than lexical decision. In this task, measures of eye movement control offered a plausible picture of the processing time-course (comparable with prior studies), while analyses of the EEG signal were not informative. This pattern of results replicated the paradox and the logical impossibility raised in studies of both simplex and complex word processing (Dambacher & Kliegl, 2007; Dimigen et al., 2011; Kliegl, Dambacher, Dimigen, Jacobs, & Sommer, 2012; Kretzschmar et al., 2015; Sereno & Rayner, 2003; Sereno et al., 1998; Schmidtke et al., 2017; Schmidtke & Kuperman, 2019) that robust behavioral effects appear without a preceding neural counterpart. The contribution of this study is in that it exemplifies the paradox in a more methodologically rigorous and complete way than earlier work. Thus, in both Experiments we used a co-registration within-participant paradigm, with a 2 x 2 factorial manipulation of well-attested predictors of complex word processing. We also used the same critical stimuli across sentence reading and lexical decision tasks. These design choices eliminated many potential confounds that cross-study comparisons may have suffered from (Schmidtke et al., 2017, 2019). We also relied on statistical analyses of both behavioral and neurophysiological data specifically designed to pin down the onset (rather than the peak or the central tendency) of an effect of interest.

In summary, we found eye-tracking to be a reliable technique for studying complex word processing in naturalistic reading, yet we found the EEG signal not to be an informative source of temporal data for this testing paradigm. In lexical decision, neurobehavioral data showed partial utility by demonstrating an earlier effect of semantic similarity than that detected in the eye-movement record; yet it did not reveal a word frequency effect found in oculomotor behavior. Based on our own data and earlier reports (see the Introduction), we must conclude that the present-day methodological or analytical approaches to neurophysiological data do not give rise to credible or complete estimates of the time-course of word processing as it occurs in natural reading and possibly in isolated-word recognition paradigms.

We emphasize that our criticism is solely directed at the use of neurophysiological data as a source of *temporal* information. In fact, this criticism supports a long-standing notion that neurobehavioral signals should be used to "provide insight concerning the processes, rather than a list of correlations between products" (Donchin, 1981, p. 497). On this view, behavior is comprised of multiple parallel processes, and these processes have their corresponding *products*, which can be observed as button presses, ERP components, gaze durations, etc. Even if generated by the same underlying process(es), these products may reflect different features of the processes, which may or may not be time-locked to the same degree. One possibility to entertain is that neuronal populations underlying lexical processing may produce non-phase-locked responses. In this case, time-frequency analysis could be an alternative way of looking into the problem. Another possibility could be that the areas that show the effects we are looking for may not be readily recordable using scalp EEG due to the depth, dipole direction, or that the brain coding for such an effect is dispersed enough to not have decent signal-to-noise ratio on the scalp. If this is the case, better results may be obtained by applying intracranial EEG. Clearly, more research is needed to shed light on the problem at hand. We see the role of the present paper in demonstrating that EEG amplitudes should not be used directly for temporal estimates of cognitive processes involved in word recognition during natural reading. We hope that this demonstration encourages the field of language research to explore a greater variety of features and analyses in the study of neurophysiological responses to words as stimuli.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior research methods*, *40*(1), 278-289.

Amenta, S., & Crepaldi, D. (2012). Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in psychology*, *3*, 232.

Amenta, S., Marelli, M., & Crepaldi, D. (2015). The fruitless effort of growing a fruitless tree: Early morpho-orthographic and morpho-semantic effects in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1587.

Assadollahi, R., & Pulvermüller, F. (2003). Early influences of word length and frequency: a group study using MEG. *Neuroreport*, *14*(8), 1183-1187.

Auch, L., Gagné, C. L., & Spalding, T. L. (2020). Conceptualizing semantic transparency: A systematic analysis of semantic transparency measures in English Compound words. *Methods in Psychology*, *3*, 100030.

Baayen, R. H. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental lexicon working papers*, *1*(1), 1-47.

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The mental lexicon*, *2*(3), 419-463.

Harald Baayen, R., van Rij, J., de Cat, C., & Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. *arXiv e-prints*, arXiv-1601.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive psychology*, *17*(3), 364-390.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283.

Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision research*, *19*(9), 967-983.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, *7*(6), 1129-1159.

Berkum, J. J. V., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of cognitive neuroscience*, *11*(6), 657-671.

Bertram, R., & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of memory and language*, *48*(3), 615-634.

Bertram, R. (2011). Eye movements and morphological processing in reading. *The Mental Lexicon*, *6*(1), 83-109.

Brooks, T. L., & Cid de Garcia, D. (2015). Evidence for morphological composition in compound words using MEG. *Frontiers in human neuroscience*, *9*, 215.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45-50.

Cavalli, E., Colé, P., Badier, J. M., Zielinski, C., Chanoine, V., & Ziegler, J. C. (2016). Spatiotemporal dynamics of morphological processing in visual word recognition. *Journal of Cognitive Neuroscience*, *28*(8), 1228-1242.

Clifton Jr, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. *Eye movements*, 341-371.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain research*, *1084*(1), 89-103.

Dambacher, M., & Kliegl, R. (2007). Synchronizing timelines: Relations between fixation durations and N400 amplitudes during sentence reading. *Brain research*, *1155*, 147-162.

De Cat, C., Klepousniotou, E., & Baayen, R. H. (2015). Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English. *Frontiers in Psychology*, *6,* 77.

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., & Liversedge, S. P. (2019). Parafoveal previews and lexical frequency in natural reading: Evidence from eye movements and fixation-related potentials. *Journal of Experimental Psychology: General*, *148*(3), 453.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of experimental psychology: General*, *140*(4), 552.

Drieghe, D., Brysbaert, M., Desmet, T., & De Baecke, C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology*, *16*(1-2), 79-103.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

El-Bialy, R., Gagné, C. L., & Spalding, T. L. (2013). Processing of English compounds is sensitive to the constituents' semantic transparency. *The Mental Lexicon*, *8*(1), 75-95.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision research*, *42*(5), 621-636.

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

Feldman, L. B., Milin, P., Cho, K. W., Moscoso del Prado Martín, F., & O'Connor, P. A. (2015). Must analysis of meaning follow analysis of form? A time course analysis. *Frontiers in human neuroscience*, *9*, 111.

Fruchter, J., & Marantz, A. (2015). Decomposition, lookup, and recombination: MEG evidence for the full decomposition model of complex visual word recognition. *Brain and Language*, *143*, 81-96.

Gagné, C. L., Spalding, T. L., & Nisbet, K. A. (2016). Processing English compounds: Investigating semantic transparency. *SKASE Journal of Theoretical Linguistics*, *13*(2).

Geraci, M. (2014). Linear quantile mixed models: the lqmm package for Laplace quantile regression. *Journal of Statistical Software*, *57*(1), 1-29.

Gordon, P. C., Moore, M., Choi, W., Hoedemaker, R. S., & Lowder, M. W. (2020).

Individual differences in reading: Separable effects of reading experience and processing skill. *Memory & cognition*, *48*(4), 553-565.

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Language and linguistics compass*, *3*(1), 128-156.

Hastie, T., & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 1005-1016.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, *30*(4), 1383-1400.

Hyönä, J., Bertram, R., & Pollatsek, A. (2004). Are long compound words identified serially via their constituents? Evidence from an eye movement-contingent display change study. *Memory & cognition*, *32*(4), 523-532.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & psychophysics*, *40*(6), 431-439.

Inhoff, A. W., Kim, A., & Radach, R. (2019). Regressions during reading. *Vision*, *3*(3), 35.

Jared, D., Jouravlev, O., & Joanisse, M. F. (2017). The effect of semantic transparency on the processing of morphologically derived words: Evidence from decision latencies and event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(3), 422.

Juhasz, B. J. (2007). Eye movements: A window on mind and brain.

Juhasz, B. J. (2018). Experience with compound words influences their processing: An eye movement investigation with English compound words. *Quarterly Journal of Experimental Psychology*, *71*(1), 103-112.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods 42*(3), 627-633.

Kielar, A., & Joanisse, M. F. (2011). The role of semantic and phonological factors in word recognition: An ERP cross-modal priming study of derivational morphology. *Neuropsychologia*, *49*(2), 161-177.

Kintsch, W., & Walter Kintsch, C. B. E. M. A. F. R. S. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Kirkby, J. A., Webster, L. A., Blythe, H. I., & Liversedge, S. P. (2008). Binocular coordination during reading and non-reading tasks. *Psychological bulletin*, *134*(5), 742.

Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A. M., & Sommer, W. (2012). Eye movements and brain electric potentials during reading. *Psychological research*, *76*(2), 145-158.

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

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1648.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, *62*(2), 83-97.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading polymorphemic Dutch compounds: toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 876.

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

Kuperman, V. (2013). Accentuate the positive: Semantic access in English compounds. *Frontiers in psychology*, *4*, 203.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293-323.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lavric, A., Clapp, A., & Rastle, K. (2007). ERP evidence of morphological analysis from orthography: A masked priming study. *Journal of cognitive neuroscience*, *19*(5), 866-877.

Lavric, A., Rastle, K., & Clapp, A. (2011). What do fully visible primes and brain potentials reveal about morphological decomposition? *Psychophysiology*, *48*(5), 676-686.

Lavric, A., Elchlepp, H., & Rastle, K. (2012). Tracking hierarchical processing in morphological decomposition with brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 811.

Lehtonen, M., Monahan, P. J., & Poeppel, D. (2011). Evidence for early morphological decomposition: Combining masked priming with magnetoencephalography. *Journal of cognitive neuroscience*, *23*(11), 3366-3379.

Leminen, A., Smolka, E., Dunabeitia, J. A., & Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *cortex*, *116*, 4-44.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and language*, *61*(1), 30-44.

Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and language*, *84*(1), 50-64.

Liversedge, S. P., Blythe, H. I., & Drieghe, D. (2012). Beyond isolated word recognition. *Behavioral and Brain Sciences*, *35*(5), 293-294.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57-78.

Marelli, M., & Luzzatti, C. (2012). Frequency effects in the processing of Italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language*, *66*(4), 644-664.

Marelli, M., Amenta, S., Morone, E. A., & Crepaldi, D. (2013). Meaning is in the beholder's eye: Morpho-semantic effects in masked priming. *Psychonomic bulletin & review*, *20*(3), 534-541.

Matsuki, K. (2014). grt: General recognition theory. r package version 0.2.

McCarron, S. P., & Kuperman, V. (2021). Is the author recognition test a useful metric for native and non-native english speakers? An item response theory analysis. *Behavior Research Methods*, 1-12.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, *17*(6), 578-586.

Meyberg, S., Sommer, W., & Dimigen, O. (2017). How microsaccades relate to lateralized ERP components of spatial attention: a co-registration study. *Neuropsychologia*, *99*, 64-80.

Meziere, D., Yu, L., Reichle, E., von der Malsburg, T., & McArthur, G. (2021). Using Eye-Tracking Measures to Predict Reading Comprehension.

Molinaro, N., Conrad, M., Barber, H. A., & Carreiras, M. (2010). On the functional nature of the N400: Contrasting effects related to visual word recognition and contextual semantic integration. *Cognitive Neuroscience*, *1*(1), 1-7.

Morris, J., Frank, T., Grainger, J., & Holcomb, P. J. (2007). Semantic transparency and masked morphological priming: An ERP investigation. *Psychophysiology*, *44*(4), 506-521.

Morris, J., Grainger, J., & Holcomb, P. J. (2008). An electrophysiological investigation of early effects of masked morphological priming. *Language and Cognitive Processes*, *23*(7-8), 1021-1056.

Morris, J., Porter, J. H., Grainger, J., & Holcomb, P. J. (2011). Effects of lexical status and morphological complexity in masked priming: An ERP study. *Language and cognitive processes*, *26*(4-6), 558-599.

Morris, J., Grainger, J., & Holcomb, P. J. (2013). Tracking the consequences of morpho-orthographic decomposition using ERPs. *brain research*, *1529*, 92-104.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*.

Niswander, E., Pollatsek, A., & Rayner, K. (2000). The processing of derived and inflected suffixed words during reading. *Language and Cognitive Processes*, *15*(4-5), 389-420.

Niswander-Klement, E., & Pollatsek, A. (2006). The effects of root frequency, word frequency, and length on the processing of prefixed English words during reading. *Memory & Cognition*, *34*(3), 685-702.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological psychology*, *74*(3), 374-388.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading*, *18*(1), 22-37.

Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects Models in S and S-PLUS." New York: Springer. *538 p*.

Pollatsek, A., Hyönä, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 820.

Pollatsek, A., & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive processes*, *20*(1-2), 261-290.

Porretta, V., Tremblay, A., & Bolger, P. (2017). Got experience? PMN amplitudes to foreign-accented speech modulated by listener experience. *Journal of Neurolinguistics*, *44*, 54-67.

Pulvermüller, F. (2002). A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order. *Progress in neurobiology*, *67*(2), 85-111.

Pylkkänen, L., Feintuch, S., Hopkins, E., & Marantz, A. (2004). Neural correlates of the effects of morphological family frequency and family size: an MEG study. *Cognition*, *91*(3), B35-B45.

R Core Team (2014). R. (2014). *A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and cognitive processes*, *15*(4-5), 507-537.

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, *23*(7-8), 942-971.

Rastle, K., Lavric, A., Elchlepp, H., & Crepaldi, D. (2015). Processing differences across regular and irregular inflections revealed through ERPs. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 747.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements?. *Vision research*, *16*(8), 829-837.

Rayner, K., & Morrison, R. E. (1981). Eye movements and identifying words in parafoveal vision. *Bulletin of the Psychonomic Society*, *17*(3), 135-138.

Rayner, K., & Sereno, S. C. (1994). Eye movements in reading: Psycholinguistic studies.

Rayner, K., & Raney, G. E. (1996). Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, *3*(2), 245-248.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: a comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(5), 1188.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

Rayner, K., Liversedge, S. P., White, S. J., & Vergilino-Perez, D. (2003). Reading disappearing text: Cognitive control of eye movements. *Psychological science*, *14*(4), 385-388.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the EZ Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 720.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological review*, *105*(1), 125.

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. *Cognitive psychology*, *65*(2), 177-206.

Reingold, E. M., & Sheridan, H. (2014). Estimating the divergence point: A novel distributional analysis procedure for determining the onset of the influence of experimental variables. *Frontiers in Psychology*, *5*, 1432.

Sheridan, H., & Reichle, E. D. (2016). An analysis of the time course of lexical processing during reading. *Cognitive science*, *40*(3), 522-553.

Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & cognition*, *26*(6), 1270-1281.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1793.

Schmidtke, D., & Kuperman, V. (2019). A paradox of apparent brainless behavior: The time-course of compound word recognition. *Cortex*, *116*, 250-267.

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2018). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 421.

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2021). CompLex: An eye-movement database of compound word reading in English. *Behavior Research Methods*, *53*(1), 59-77.

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. *Morphological aspects of language processing*, *2*, 257-294.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport*, *9*(10), 2195-2200.

Sereno, S. C., & Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends in cognitive sciences*, *7*(11), 489-493.

Sereno, S. C., Hand, C. J., Shahid, A., Mackenzie, I. G., & Leuthold, H. (2020). Early EEG correlates of word frequency and contextual predictability in reading. *Language, Cognition and Neuroscience*, *35*(5), 625-640.

Smolka, E., Gondan, M., & Rösler, F. (2015). Take a stand on understanding: Electrophysiological evidence for stem access in German complex verbs. *Frontiers in human neuroscience*, *9*, 62.

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042-2057.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading research quarterly*, 402-433.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. *The Oxford handbook of psycholinguistics*, *327*, 342.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1280.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*(4), 263-272.

Tiffin-Richards, S. P., & Schroeder, S. (2020). Context facilitation in text reading: A study of children's eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1701.

Tremblay, A. (2009). *Processing advantages of lexical bundles: Evidence from self-paced reading, word and sentence recall, and free recall with event-related brain potential recordings* (Doctoral dissertation, University of Alberta).

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication*, *151*, 173.

Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology*, *52*(1), 124-139.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, *67*(6), 1176-1190.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & cognition*, *18*(4), 380-393.

Van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2015). itsadug: Interpreting Time Series, Autocorrelated Data Using GAMMs. R package version 1.0. 1.

Vergilino-Perez, D., Collins, T., & Doré-Mazars, K. (2004). Decision and metrics of refixations in reading isolated words. *Vision research*, *44*(17), 2009-2017.

Vitu, F., McConkie, G. W., Kerr, P., & O'Regan, J. K. (2001). Fixation location effects on fixation durations during reading: An inverted optimal viewing position effect. *Vision research*, *41*(25-26), 3513-3533.

West, R. F., Stanovich, K. E., & Mitchell, H. R. (1993). Reading in the real world and its correlates. *Reading Research Quarterly*, 35-50.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Yap, M. J., Balota, D. A., Tse, C. S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: evidence for opposing I nteractive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 495.

Zweig, E., & Pylkkänen, L. (2009). A visual M170 effect of morphological complexity. *Language and Cognitive Processes*, *24*(3), 412-439.

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. *Language and cognitive processes*, *9*(3), 341-368.

Appendix

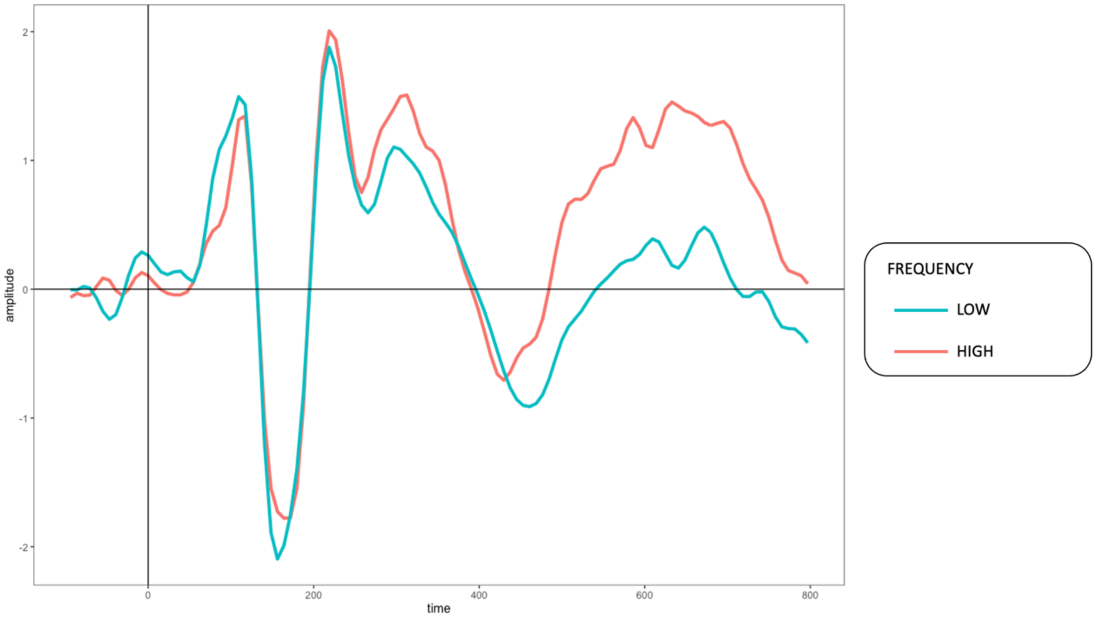

Figure A1. Grand average FRPs for the raw data for the high and low frequency words in the occipital region.

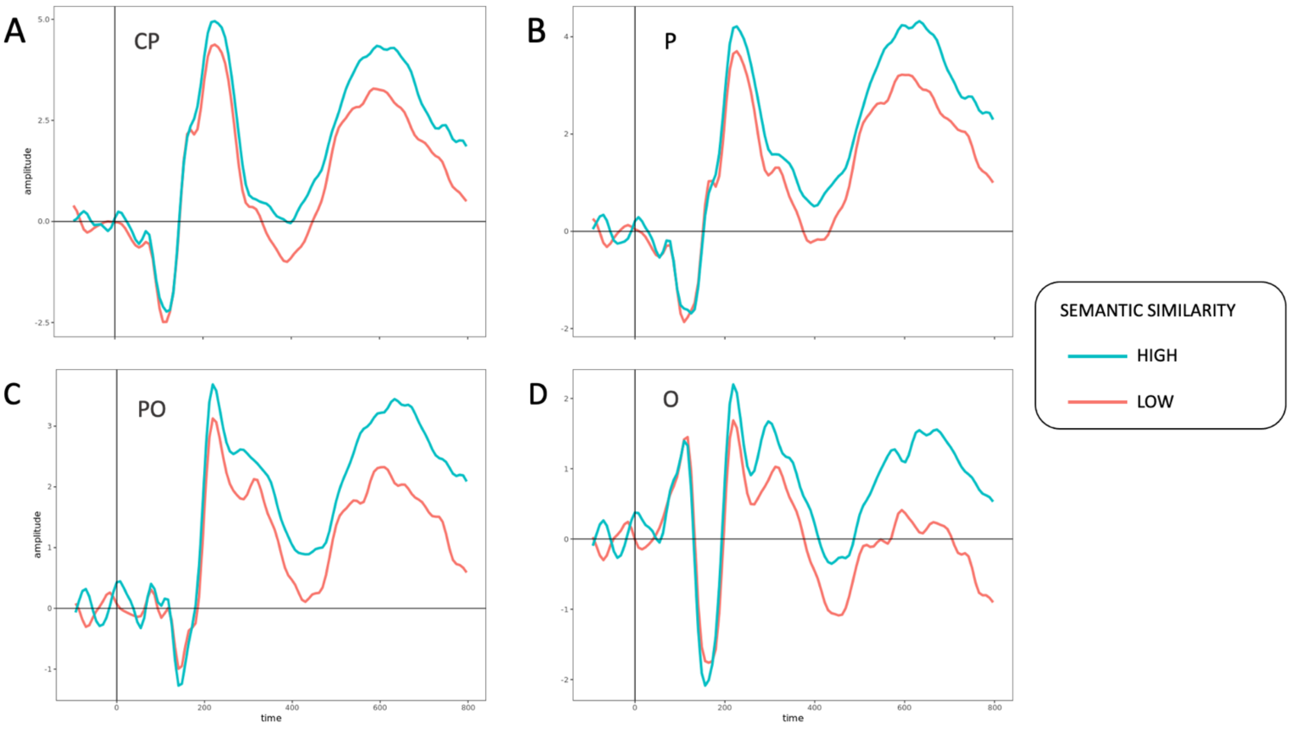

Figure A2. Grand average FRPs for the raw data for the high and low semantic similarity groups in the centro-parietal (A), parietal (B), parieto-occipital (C) and occipital (D) regions.

Chapter 3

Statistics of spelling errors affects brain responses during natural reading of Chinese: Evidence from co-registration of EEG and eye-tracking signals

This chapter has been submitted to *Journal of Experimental Psychology: Learning, Memory, and Cognition* as Oralova, G. & Boshra, R., Kyröläinen, A. J., Connolly, J.F., Kuperman, V. (submitted). Statistics of spelling errors affects brain responses during natural reading of Chinese: Evidence from co-registration of EEG and eye-tracking signals.

Abstract

Recent eye-tracking experiments found that a frequent occurrence of a spelling error negatively impacts recognition of a correctly spelled word during reading (Rahmanian and Kuperman, 2019; Kuperman et al., 2021). In this study, we investigate how and when spelling variability in language affects the brain activity in sentence reading. We examine fixation-related evoked potentials during reading of sentences in Mandarin Chinese to determine neural indices of competition between prescribed and alternative word spellings. Eye movements of thirty readers were recorded by an eye-tracker and their brain electrical activity was simultaneously registered via EEG. All targets were presented in their correct spelling and selected to represent a range of spelling entropy, a measure of competition between alternative spellings. The fixation-related potential (FRP) analysis showed average amplitude differences for words with high and low entropy values only at early, 150-300 ms, time window. These results confirm previously reported eye-tracking results and show that distributional properties of incorrect spellings influence the cognitive effort of word recognition. FRP analysis further pinned down the time-course of the effect: spelling entropy affects word processing early, during orthographic processing stages, and reflects a competition between orthographic variants. We discuss implications of our findings for theories of reading and learning.

1 Introduction

While living in a digital world, surrounded by computer devices with numerous spell checker programs, being a good speller might seem unimportant. However, proficiency in spelling is known to both directly contribute to proficiency and fluency of reading (Perfetti, 1997) and to support several other critical component skills of reading, including phonological (Ehri & Wilce, 1985) and morphological awareness (Nagy, Berninger, & Abbott, 2006). Perhaps the most direct influence that spelling has on reading emerges at the word level: better spellers are faster, more accurate, and less affected by context during word recognition (Burt & Fury, 2000; Ehri, 2000; Perfetti, 1997).

A common way of investigating spelling skills is to examine the nature and distribution of spelling errors. This is because such errors are a window to understanding how spelling is acquired (e.g., Ouellette and Sénéchal, 2008), and how properties of a specific linguistic and writing system interact with generic cognitive mechanisms of learning and memory (e.g., Burt and Fury, 2000; Rapp and Fischer-Baum, 2015). One oft-discussed link between spelling and reading is captured in the Lexical Quality Hypothesis (Perfetti & Hart, 2001). In order for a word to be fully represented in an individual's mental lexicon, all three components – orthography, phonology and semantics – need to have "crisp" representations and strong mappings amongst them. If the orthographic component is not strongly connected with others, there is a chance for a spelling error to occur. For instance, a recent study (Martin-Chang, Ouellette, & Madden, 2014) investigated whether lower-quality orthographic representations (as reflected in spelling errors) slow down the speed of single word reading. A within-participant analysis revealed that words that participants spelled consistently accurately were read faster than words which they misspelled. Moreover, the same words were read faster by individuals who always spelled them correctly, compared to those who did not (see also Dixon and Kaminska, 1997, 2007; Jacoby and Hollingshead, 1990). In sum, the literature demonstrates that spelling errors reflect deficient orthographic representations and also words that have poor-quality and unstable orthographic representations are recognized and read slower (Dixon & Kaminska, 1997; Jacoby & Hollingshead, 1990; Martin-Chang et al., 2014).

There is growing evidence that the relationship between spelling errors and orthographic representations may be reciprocal (Rahmanian & Kuperman, 2019). On one hand, spelling errors are thought to be a result of unstable orthographic representations in a writer's mind: see above and Conrad (2008), Holmes and Castles (2001). Another way of looking at spelling errors is as a cause for development of unstable, competing or deficient lexical representations. In speech production and comprehension studies, it is assumed that alternative phonetic variants of a word upon occurrence create their own lexical representations that are all stored in the mental lexicon, i.e., long-term lexical memory (e.g., Ernestus, 2014). Existence of several phonetic variants for a word arguably leads to a competition between them at comprehension. Similarly, when reading alternative orthographic forms for a certain word, multiple orthographic representations are stored in the mental lexicon and begin to compete for activation upon the next encounter of the word.

According to several theories of learning, the frequency of simultaneous exposure to forms and meanings is what determines the strength of association between each form and each meaning (Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; Ramscar, Dye, & McCauley, 2013). For example, the more you see the orthographic form *girafe* instead of *giraffe* as the name of an animal, the stronger the connection of *girafe* and the weaker the connection of *giraffe* with the meaning of the word *giraffe*. In sum, the presence and the frequency of occurrence of different spelling variants plays an important role in establishing associations between the cues and the outcomes. Consequently, frequent spelling errors may cause formation of multiple competing orthographic representations and cause uncertainty upon word identification about the correctness of the orthographic representation.

Several recent studies have examined the possibility of the reciprocal relation between spelling errors and reading behavior. First, Rahmanian and Kuperman (2019) investigated whether orthographic competition and the ensuing uncertainty of choosing between spelling variants poses difficulty in recognizing correctly spelled English words during reading. Using eye-tracking, they found that participants showed longer reading times to words that were associated with relatively frequent spelling errors (e.g., *innocent* spelled as *inocent* 31% of the time). Such words arguably came with the strongest competition between the correct and incorrect spelling, which weakened the orthographic representation of the correctly spelled words presented to readers.

While the role of spelling errors has been demonstrated in English, it is important to examine universality and specificity of reading and spelling processes across different languages and scripts. Kuperman et al. (2021) conducted a cross-linguistic analysis across different writing systems to test whether Rahmanian and Kuperman's (2019) findings in English extend to four other languages, namely, Chinese, Greek, Finnish and Hebrew. Eye-tracking data across all languages revealed that correctly spelled words with a higher relative frequency of spelling errors showed longer reading latencies.

These earlier studies clearly indicated that there is behavioural evidence that spelling errors and their frequency of occurrence affect word recognition processes across different written languages. However, this literature leaves open several questions, partly because of the limitations of eye-tracking as an experimental method. A central temporal measure of oculomotor activity is fixation duration. The eye-tracking registration provides the start and the end point of a fixation. While duration of a fixation is a highly valuable source of information about the underlying cognitive effort, it provides no insight into the dynamics of that effort *during* the fixation. Thus, inferences about the time-course of cognitive processes during word reading drawn from eye-tracking data are largely limited to whether word-related changes in cognitive effort can be statistically shown in the durations of the first or subsequent fixations or in the likelihood of having a single or more than one fixation on the word.

An alternative that registers cognitive processing both within and between fixations is electroencephalography (EEG), a recording technique that is able to monitor neural activity in real time with high temporal resolution (in milliseconds) (Duncan et al., 2009). Yet EEG and its use for the registration of event-related potentials (ERP) come with their own limitations, discussed below. The present study contributes to research of the relationship between spelling and reading by examining both behavioral and neural responses to spelling errors in Mandarin Chinese, using co-registration of eye-movements and the ERP brain activity. In the remainder of the Introduction, we review the neuropsychological literature on spelling errors and related phenomena and develop predictions as to what perceptual and cognitive mechanisms can be engaged by processing words that are often misspelled. We further discuss the co-registration technique and outline our expectations as to the time-course of the cognitive processing and its indications in the eye-movement and EEG record.

There is a rich EEG literature on how readers detect, reanalyze and repair from morphosyntactic, syntactic and orthographic violations (Hagoort, Brown, & Groothusen, 1993; Helenius, Salmelin, Service, & Connolly, 1999; Münte, Heinze, Matzke, Wieringa, & Johannes, 1998; Vissers, Chwilla, & Kolk, 2006). Perception of these violations reveals one of the important mechanisms in speech and reading comprehension: the error monitoring system. This cognitive system is argued to monitor and screen information processing for the occurrence of conflicts (e.g., Botvinick, Braver, Barch, Carter, and Cohen, 2001). While first postulated for the case of speech (Levelt, 1983), the mechanism of error monitoring in reading is assumed to mainly be triggered by a conflict when an orthographic input does not meet conceptual expectations of a reader (Newman & Connolly, 2004). Responses of the error monitoring system to such conflicts are mainly linked to an ERP (event-related potential) component called P600, which registers a reanalysis stage during word processing (Gouvea, Phillips, Kazanina, & Poeppel, 2010; Van de Meerendonk, Indefrey, Chwilla, & Kolk, 2011). P600 is a centro-parietal positive going ERP component typically found within 500-800 ms window. An increased amplitude of P600 is usually reported to accompany various types of conflicts arising, for instance, from syntactic violations, e.g., agreement violations (Hagoort et al., 1993; Kaan, Harris, Gibson, & Holcomb, 2000; Osterhout & Mobley, 1995), verb-inflection violations (Friederici, Pfeifer, & Hahne, 1993), case inflection violations (Münte et al., 1998), and phrase structure violations (Neville, Nicol, Barss, Forster, & Garrett, 1991).

Critically, this component has also been found to reflect spelling violations. Vissers et al. (2006) hypothesized that P600 occurs after orthographic violations especially when the word is highly expected in a sentence. In their experiment, they made use of pseudohomophones (words which orthographically and phonologically highly resemble the correctly spelled counterpart, e.g., *burd* vs *bird*) in sentences where the target words were highly predictable or less predictable. Vissers et al. confirmed that P600 is only observed when the target word is highly expected. These results showed that a pseudohomophone, being highly semantically and phonologically plausible but orthographically ill-formed, created a conflict, which brought reader’s brain into a state of indecision and elicited monitoring process registered by P600. Another earlier component, N270, observed by Newman and Connolly (2004) after orthographically incongruent words, was registered as well in the study. N270 was elicited only under a low predictability condition for pseudohomophones at left frontal regions. The authors argued that participants could be subject to an orthographic illusion, where they temporarily might think that the pseudohomophone is orthographically correct as other sources of information (phonology and semantics of the misspelled word) showed the sentence is correct or acceptable. This same effect was found by Assink, Bos, and Kattenberg (1996) where readers fail to recognize orthographic errors when the word is highly predictable from context as they do not completely process orthography of the word in this context.

In an EEG and fMRI study, Van de Meerendonk et al. (2011) tested whether syntactic and spelling violations (pseudohomophones) elicit the P600 component and whether these same violations would be localized similarly using fMRI. The results demonstrated that both types of violations were manifested in the P600 component, and both showed an increased activation in the left inferior frontal gyrus. However, spelling violations activated additional areas in more posterior regions, such as the left fusiform gyrus, which closely corresponds with the visual word form area responsible for computation of structural representations of words (Dehaene, Le Clec'H, Poline, Le Bihan, & Cohen, 2002). To explain these findings Van de Meerendonk et al. (2011) referred to the monitoring theory of language perception which states that competing representations (expected and observed) trigger a conflict, which in turn initiates reprocessing of errors. Furthermore, Van de Meerendonk et al. (2011) orthogonally manipulated the spelling violation and cloze probability conditions to test whether the P600 amplitude is modulated by different conflict strengths. Strongly predictive sentences were found to elicit larger P600 amplitudes to misspelled words compared to sentences in which the target word was less predictable.

Furthermore, Stowe, Rommers, Loerts, Timmerman, and Temmink (2010) conducted two ERP experiments with slow (480 ms) and fast (200 ms) presentation speeds where subjects read Dutch sentences with target words that are either misspelled or correctly spelled. A misspelled word was a real word in the language that was visually and phonologically similar to the correctly spelled word (e.g., "When Johnny fell off the slide, he had a broken *ark* (instead of *arm*), so that he had to be in a cast for three weeks"). The target word predicted from context (e.g., *arm*) was of high or low frequency. Stowe et al. (2010) aimed to address the problem of competition between candidate words during language processing. Results showed larger negativity in the N400 and larger positivity in the P600 for the low frequency words. Overall, results were similar in fast and slow presentation experiments.

Most of the above-mentioned studies emphasized later ERP components, namely N400 and P600, which could be relevant to the reanalysis stage and may reflect the consequences of competition between competing representations. However, there are other studies showing that ill-formed orthographic representations can elicit a response in earlier ERP time windows. For instance, Kim and Lai (2012) aimed to investigate the relationship between lexical semantic and sub-lexical visual word form processing during word recognition in context. They found a larger P130 amplitude for a response to pseudowords that highly resemble orthographically correct form of the words read in a sentence (e.g., cake vs ceke). However, pseudowords that were not similar to a plausible real word elicited a later component, N170. Kim and Lai (2012) suggested that the visual word recognition system is more rapidly sensitive to small deviations in the orthographic form, rather than to flagrant violations. Results also point to an interaction between lexical and sub-lexical representations (top-down and bottom-up processes), where a stored lexico-semantic representation of a word “guides” the recognition of the presented input and highlights the bottom-up mismatch. In another study, Sauseng, Bergmann, and Wimmer (2004) showed that ERP waveforms related to letter-altered words began to deviate from the correctly spelled words’ ERPs as early as 160 ms, and with a difference being still present at up to 700 ms. Sauseng et al. hypothesized that at this time a letter string input comes in contact with established memory representations of words, and any deviations from the correct representations lead to reduced peak amplitudes. In conclusion, two studies above highlighted early time windows, 130–170 ms, as a temporal locus where ERP amplitudes start to significantly diverge for correct and altered spellings. This difference in amplitudes at this time point is explained as a point of contact for top-down and bottom-up processing, where a clash between stored orthographic representations and presented input start to occur.

1.1 The present study

In this paper, we seek neurophysiological evidence for the behavioural effects found in Rahmanian and Kuperman (2019) and Kuperman et al. (2021). Specifically, we report the EEG part of an experimental study in Mandarin Chinese: the eye-tracking part of the study has been reported, along with 4 other languages, in Kuperman et al. (2021). The present experiment co-registered EEG and eye-tracking signals within participants and aimed at rectifying inherent limitations of either technique (discussed below). Reading times, gauged via eye-movements, revealed a reliable effect of spelling entropy, a measure which reflects a competition between orthographic variants of a given word. Our present goal was to determine whether the EEG signal would also reveal an effect of spelling entropy when reading sentences in Mandarin Chinese.

The use of co-registration sets our study apart from previous work that used only one of the methods (either eye-tracking or the EEG) to pin down distinct behavioural or neural mechanisms during word reading. Limitations of the eye-tracking method are mentioned above. One crucial limitation of EEG as a method is its usage of the rapid serial visual presentation (RSVP) technique when presenting sentences to participants. RSVP has been used extensively in reading studies and in ERP studies in particular. Although necessitated for technical reasons, it is undoubtedly a very artificial method to study reading. Co-registration of EEG/ERP and eye-tracking is a recent advancement that enabled connecting the natural reading behavior recorded with eye-tracking to neural signals associated with fixations (i.e., fixation related potentials; FRP) and saccades (i.e., saccade related potentials; SRP). Synchronized recordings enable direct access to FRPs and SRPs that are indicative of cognitive processing in natural reading, as opposed to rapid serial presentations of single words in a typical EEG/ERP study (Baccino & Manunta, 2005; Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011; Hutzler et al., 2007). Further, there is disagreement in the literature on whether ERP data alone serve as a valid index of temporal activity during printed language comprehension, especially when the timing of lexical effects on ERPs is compared to that in analogous eye-tracking data (Schmidtke & Kuperman, 2019; Schmidtke, Matsuki, & Kuperman, 2017). We argue that only by examining neural responses that are time-locked to eye movements can these questions be properly elucidated. In this study, we utilized this relatively new co-registration technique to reveal neurophysiological evidence for spelling entropy during natural reading and to make meaningful comparisons to the eye-tracking results from the same experiment.

Since fixation-related potentials (FRP) observed in the co-registration of the eye-tracking and EEG data rectify limitations of each respective technique, we capitalize on their ability to focus on neural activity unfolding within each fixation when encountering a word. Our primary interest is in charting a detailed time-course of processing words that have multiple (prescribed and non-normative) spellings. To reiterate, Rahmanian and Kuperman (2019) and Kuperman et al. (2021) argue that every time readers encounter a misspelled word, this occurrence strengthens an association between that orthographic form and the word meaning, while weakening an association between the correct spelling and that meaning. Statistics of encountering alternative spellings determines the quality of orthographic representations of these alternatives and influences recognition of any of the alternatives.

From the review of the previous EEG literature, we assume that this competition could create enough of a representational conflict upon reading a critical word and be reflected in early, later or both ERP component windows. If the effect of spelling entropy, as a measure of form competition, is only seen in earlier time windows (e.g., P130/N270 or 150–300 ms time windows), this would suggest that entropy modulates the strength of representational conflict between stored orthographic representation and presented input and brings the reader to a state of indecision. If only later time windows and later FRP components are affected by entropy (N400/P600 or 300–500 ms and 500–700 ms time windows), this would mean that the representational conflict is only evident when monitoring processes have launched reanalysis mechanisms. There is also a possibility of entropy affecting both early and late time windows. In this case, occurrence of a conflict might be registered in the earlier time windows – because readers' expectations about the upcoming word are violated or because there is a high uncertainty of which spelling is correct – and the conflict also initiates monitoring processes registered by later time windows (for a review of monitoring theory, see Van de Meerendonk, Kolk, Chwilla, and Vissers, 2009).

It is worth noting that all neuroimaging studies cited above focused on reading of ill-formed orthographic representations. However, all participants in the present study were shown correct orthographic forms only. This means that if a possible representational conflict has arisen it is likely due to the uncertainty in activation of possible spelling variants of a word. This experimental decision enables us to investigate whether the effect of spelling entropy (i.e., competition between alternative orthographic representations) creates enough of a conflict in order to be detectable in the electrophysiological activity of the human brain.

The language of this study is Mandarin. In Chinese orthography, word misspelling is a common phenomenon. Lists of frequently occurring character misspellings or substitutions are often published on educational websites and in newspapers, such as

“Public’s Daily”. High school students are given lists of common spelling mistakes to practice avoiding them; the lists are compiled by teachers, who collect them throughout their teaching career. Several publishing agencies have also issued books and dictionaries with hundreds of frequently misspelled words and offer spelling exercises.

Partly, this prevalence of spelling errors and the educational focus on them is due to the structure of written Chinese. The Chinese writing system (shared by speakers of Mandarin, Cantonese and other dialects) consists of basic units called characters, which map onto syllabic morphemes. One word can contain from one and up to several characters, but most words are two, three or four characters. Another prevalent feature of Chinese is that it does not have overt cues for the demarcation of word boundaries. Consequently, it has been noted that readers of Chinese do not often agree where to put word boundaries, and words themselves do not seem to be as transparent as in alphabetical languages with spaces between words (Chen, 1999; Peng & Chen, 2004). However, they were still found to be important processing units during text comprehension (Li, Gu, Liu, & Rayner, 2013; Li, Rayner, & Cave, 2009; P.-P. Liu, Li, Lin, & Li, 2013).

Characters may be classified into two categories in terms of their structural complexity: simple and compound characters. Simple characters are not divisible into distinct components, whereas compound characters, which are the majority in Chinese, are composed from two or three components, named radicals. These radicals further consist of several strokes. Radicals can provide information about meaning (semantic radicals), for example, 女 (nǚ, woman), and pronunciation (phonetic radicals) 马 (mǎ, horse), as in the character 妈 (mā, mother). This last character 妈 is related in meaning to 女, and in pronunciation to 马, thus, having both semantic and phonetic information combined in one character. According to Modern Chinese Dictionary, simple characters comprise 15% of the present day used characters, whereas the remaining 85% are phonetic compound characters (Perfetti & Tan, 1999).

In sum, due to its methodology the present co-registration study can serve as a “magnifying glass” to pin down the time-course of cognitive processes within and between fixations and test which of the predictions above are borne out. In this regard, our study of neural activity aims to expand on the available behavioral results obtained from the same participants (the Chinese eye-tracking data in Kuperman et al. (2021)) and from other written languages.

2 Methods

2.1 Participants

Thirty-seven McMaster University students (female: 28; 18–26 range, mean age: 21.43) participated in a two-hour experiment for a course credit or monetary compensation. All the participants reported to be native Mandarin speakers, with 22 subjects being at a bachelor level and 15 at a graduate level of education. Participants reported no neurological, psychological, or psychiatric problems and were not on medications that could affect their central nervous system. All participants had normal or corrected-to-normal vision. There were 36 right-handed subjects, and only one was left-handed according to the Oldfield's Edinburgh Handedness Inventory (Oldfield et al., 1971). Data from seven participants were discarded due to artifactual EEG recordings. The same set of 30 participants was included into eye-tracking and EEG analyses.

2.2 Apparatus and recording

Participants were seated in a dimly lit room in front of a 17-inch monitor at a distance of 60 cm from the chin-rest, which was used to stabilize subject's head. The monitor had a resolution of 1600 by 1200 pixels and a refresh rate of 60 Hz. Eye movements were recorded with the EyeLink 1000 eye-tracker head mount (SR Research Ltd., Kanata, Ontario, Canada) at a sampling rate of 1000 Hz. The experiment trials were presented using the Experiment Builder Version 2.1.140 software (SR Research Ltd., Kanata, Ontario, Canada). Data from 28 subjects was recorded from the right eye, and 2 from left eye due to calibration issues. Viewing was binocular. Before the recording, all subjects underwent 13-point calibration and validation. For further details see Kuperman et al. (2021).

EEG was recorded using the Biosemi ActiveTwo system from 64 channels placed with an elastic cap according to the extended 10-20 system. Two additional electrodes were used to record Electrooculographic (EOG) activity and were placed above and over the outer canthus of the left eye. Three extra electrodes were placed to record from the two mastoid processes and from the tip of the nose. All data were collected using Ag/AgCl electrodes with a sampling rate of 512 Hz, referenced online to the driven right leg circuit, and bandpass filtered at 0.01 to 100 Hz. Fixation markers were overlaid on the EEG recording utilizing TTL signals originating from the eye-tracker computer. TTL signals were sent at the beginning and the end of each trial. In addition, an invisible boundary trigger was sent when participants crossed the boundary just before the target word in the sentence.

2.3 Data synchronization

Data synchronization between continuous EEG and Eye-tracking signals was accomplished by using the TTL signals sent to the EEG-recording computer during stimulus presentation. Alignment was conducted offline via the EYE-EEG extension of the EEGLAB toolbox (Dimigen et al., 2011; http://www2.hu-berlin.de/eyetracking-eeg/index.php). The synchronization quality was confirmed by a correlation of 1 between the stimulus markers of both recordings, and deviations equal or shorter than 1 ms in absolute value.

2.4 Materials

Chinese spelling mistakes can be categorized into two groups: the first is when one of characters is written incorrectly, for example, imagine character 国 “guó” with one of the strokes missing; and the second is when one of characters is correctly spelled but does not occur in this word, as in 自已 “zìyǐ” (instead of 自己 “zìjǐ”). In this second group, spelling mistakes can be phonologically and orthographically similar, or even homophonic with the correctly spelled words. This paper focuses on the second type, i.e., character misuse or substitution, where a misused character is part of a two-character compound word.

The two most frequent causes for using an incorrect or misused character in Chinese words are phonological and/or visual similarity between the correct and incorrect characters (C.-L. Liu, Tien, Lai, Chuang, & Wu, 2009). For example, 和谐 “héxié” and 合谐 “héxié” (’harmony’), where the first two characters are visually distinct but homophonic; 青睐 “qīnglài” and 亲睐 “qīnlài”(’to favor’), where first syllables of the two words have similar pronunciations, but orthographically different; or, 安装 “ānzhuāng” and 按装 “ànzhuāng”(’to install’), where first syllables of two words are phonologically and orthographically similar. When doing an error analysis on 3208 error occurrences in Chinese words, C.-L. Liu et al. (2009) found that 76% of the errors were related to the phonological similarity between the correct and the incorrect characters; 46% were due to visual similarity, and 29% involved both factors.

In light of these language statistics, 70 frequently misspelled words were selected as targets. Collection of the target words were performed based on the Dictionary of misspelled words in Chinese (Zuo Wei, 2004). The following criteria were obeyed when selecting words: first, all words must be two-character compound words, where one of the characters was frequently misplaced with another (the misspelled character was orthographically similar and phonologically similar or identical to the correct character); second, the misspelled compound word should not form an existing word in the language. Target words included only open-class words (nouns, adjectives, and verbs).

Frequency of all target words were based on the results of web search engine Google (i.e., the number of pages a word appears in, checked on February 2018) and were further used in calculation of spelling entropy between correct and incorrect spellings of that word. Although statistics in the Web corpus is considered unreliable due to a concern for lexical and grammatical correctness (Kilgarriff & Grefenstette, 2003), this method was used as a tool for frequency calculation due to the abundance of distinct variations of word spellings existent on the web. To the best of the authors’ knowledge, there were no web-based corpora of unedited texts for Chinese, which would be suitable for frequency calculations of erroneous spellings for the target words used in the experiment.

Experimental sentences for each target word (N = 70) were extracted from The Center for Chinese Linguistics online corpus of Chinese language published by Peking University, based on a corpus of Modern Chinese of 307 million characters. Sentences were read by 3 native speakers of Chinese and verified for semantic and syntactic legality. The length of sentences did not exceed 45 characters (20-25 words), so that each occupied only one line on the screen.

After the eye-tracking and EEG portion of the test, a spelling test in Chinese was conducted to measure participants' spelling skills. The test consisted of 25 questions asking subjects to detect misspelled characters in words, phrases, and sentences. The test mimicked questions used in the Chinese national college entrance examination. Additionally, the Print Exposure Questionnaire was used to measure subjects' exposure to print by asking questions about how much time they spend per day reading printed or online materials or surfing the internet.

2.5 Procedure

The research was conducted in accordance with the 1964 Helsinki Declaration, and all subjects provided informed consent before proceeding with the experiment. Participants filled the screening form asking about their previous medical history (past head injury, current disabilities, etc.), current physiological state (current medications, sleep hours, alertness, etc.) and basic demographic information (gender, age, education, handedness, etc.).

Subjects were asked to read sentences shown one at a time, each presented on a single line. Subjects were also tasked with answering yes-no comprehension check questions after 30% of sentences, however, they were unaware of when a yes-no question would be asked. The answers were recorded using a button press on a keyboard in front of the participant. The experiment was designed to ensure the natural flow of reading behavior including regressions or leftward saccades for re-reading sentences, if necessary. Participants were instructed to read sentences naturally without suppressing eye blinks but minimizing head movements. Participants were given an opportunity to rest between trials, if requested.

Before the actual experiment, subjects read four practice sentences to ensure they got familiar with the experiment flow. Each trial began with a drift correction procedure where gaze of a participant was placed at a fixation point located at the beginning of the first word of a sentence on the left side of the screen. The experimenter validated each drift correction by ensuring participant's gaze was over the fixation point on the eye-tracker computer with a deviation of no more than one visual degree by one box around the fixation point. Participants read sentences at their own pace by freely moving their eyes over the sentences. The difference between calibration and validation measurement was kept below 0.5 degrees of the visual angle. Once they finished reading a sentence, they were asked to press a button on a keyboard to initiate the next trial.

After the computer test, subjects completed the spelling test and the Print Exposure Questionnaire.

2.5.1 Preprocessing of FRP data. Continuous EEG signals were re-referenced offline to the averaged mastoids and filtered using a 0.1-30 Hz band-pass filter. Data were then synchronized to eye movements before visual inspection to remove any blocks with muscle artifacts. Independent Component Analysis (ICA) was applied to correct for ocular artifacts following the procedure in Meyberg, Sommer, and Dimigen (2017) optimized to the simultaneous EEG and eye-tracking recording. Before applying the ICA decomposition, data were filtered further using a 15-30 Hz filter to reduce slow drifts and were epoched around eye fixations into 0.6 second segments starting at -0.1 seconds before each fixation (-0.1-0.5 s). Extended Infomax ICA algorithm (Bell & Sejnowski, 1995) was used to decompose the epochs. Independent Components that had a covariance of 1.1 with eye movements and was confirmed as artifactual by researcher were removed from the data. The decomposition was then applied to the continuous EEG and examined using the EEG-EYE extension for likely candidates for component removal (Plöchl, Ossandón, & König, 2012).

Epochs of length -100 ms to 1 s were then extracted time-locked to fixations that occurred 100 ms or less after crossing the invisible boundary. All trials confirmed by the eye-tracking data as skipped were removed from all further analyses (20% of trials). Automatic artifact rejection was also conducted where any trial containing values beyond -100-100 µV was discarded (3% of data). After all the removal of data described above, we ended up with 65% of trials that entered the analysis. All epochs were baseline corrected using the 100 ms before stimulus onset.

Due to the lack of previous literature on the spelling entropy effect in ERP/FRP, we considered a number of non-overlaping time-windows that were previously examined in misspelling studies (Kim & Lai, 2012; Newman & Connolly, 2004; Sauseng et al., 2004; Van de Meerendonk et al., 2009). Three time windows were chosen, starting at 150 ms post initial fixation on a target word and ending at 700 ms. The windows were [150-300 ms], [300-500 ms], and [500-700 ms]. Single-subject averages were calculated for every condition (high and low entropy), and average amplitude values were extracted for each window. Average amplitude defined as the mean activity within a time-window of interest. For an EEG analysis, all target words were divided into two conditions via a split at the median: low entropy condition, with entropy values ranging from 0.01 to 0.62 (35 target words), and high entropy condition, with entropy values ranging from 0.63 to 0.99 (35 target words). (For the definition of entropy see below.) Electrodes were clustered in the region of interest shown to be maximal at the selected time windows while being separated from the front of the head to further minimize ocular artifacts. The region was defined as mid-parietal: CP1, CPz, CP2, P1, P2, Pz. Electrodes within this region of interest were averaged; all others were discarded.

2.6 Variables

Similar to Rahmanian and Kuperman (2019) and Kuperman et al. (2021), we used an information-theoretic measure of entropy to quantify uncertainty regarding spelling of target words they see in the sentences (for a worked example see Milin, Kuperman, Kostic, and Baayen (2009)). Target word entropy was the main independent variable of interest here. Entropy is a measure of an average "surprise" or amount of uncertainty in a probability distribution of a random variable. It can be interpreted as an average effort selecting one of the alternatives from a set of alternatives, given their probabilities. It is calculated as a negative sum of probability of each event multiplied by a logarithm of probability of that event. In the case of spelling entropy, it is a negative sum of all probabilities for each spelling variant multiplied by logarithm of probabilities of those spelling variants.

$$H = -\sum_{i=1}^{n} p_i log(p_i),$$

where *p* is the probability of a spelling variant *i* calculated from the distribution of frequencies for all spelling variants.

A word has high spelling entropy *H* when there are many alternative competing spelling forms or when the forms are close to one another in their probability of occurrence. High spelling entropy of a word reflects a higher degree of uncertainty about which spelling to use. When a word has a low entropy, there is little to no uncertainty regarding the preferred spelling. The critical words we used had their entropies varying from 0.01 to 0.99, and they were embedded in the middle of carrier sentences.

Frequency of a correct spelling, which was presented to participants during the experiment sentence reading, was included as a predictor in the model as well. Frequency of the target word did not correlate with the entropy (*r*(68) = -.16, p = .19). Also, scores on the spelling test were added as an independent variable.

The dependent variables of the eye-tracking analysis reported in Kuperman et al. (2021) were: first fixation duration (the duration of the first fixation on a word), gaze duration (summed duration from all fixations prior moving the eyes to the next word), and total reading time (summed duration from all fixations including regressions). Table1 below summarizes descriptive statistics of all variables.

2.7 Statistical considerations

For the eye movement analysis, we used Generalized Additive Mixed Models (GAMM) with the help of *mgcv* package, Version 1.8 - 28 (Wood, 2006) in R statistical software, Version 1.1.463 (R Core Team et al., 2015). GAMM is a non-parametric regression modelling technique with a general additive structure which captures the dependencies between repeated measurements, within or between subjects and stimuli, thus, modelling a more complex random effect structure. GAMM allows researchers to model linear and non-linear relationships between dependent and independent variables. For this reason, GAMM was used in order to not impose a linear relationship between variables of interest, entropy and fixation duration. Entropy x frequency interaction was also modeled as a tensor product to capture a wiggly surface (two independent variables combined) and showed the effect of entropy at five different frequency levels equally spaced between 10th and 90th percentile.

EEG statistical analyses were conducted using repeated measures analysis of variance (ANOVA) for average amplitude in a defined time window. Degrees of freedom were corrected using the more conservative Greenhouse-Geisser estimates of epsilon when the sphericity assumption was violated (Greenhouse & Geisser, 1959). ANOVAs examined the main effects of Entropy (2 levels: high vs. low), Frequency (2 levels: high vs. low), and their interaction.

Table 1. Descriptive statistics of all dependent and independent variables.

Variable	M	SD	median	min	max
Entropy	0.57	0.32	0.63	0.04	1.00
Correct freq, log	13.76	1.11	13.61	11.70	16.91
Proficiency score	11.13	2.67	11.00	5.00	16.00
First fixation duration	271.98	107.10	254.00	83.00	941.00
Gaze duration	323.35	169.25	279.00	83.00	1243.00
Total fixation time	466.92	266.95	396.00	83.00	1293.00

3 Results

This section begins with a brief discussion of the behavioral eye-tracking results reported by Kuperman et al. (2021) for Chinese. Then, we present findings from the fixation-related potential (FRP) analysis of the EEG data, which are the focus of this paper.

3.1 Eye-tracking

After trimming the original eye movement data, a regression model was fitted to 1626 observations of eye-movements to target words from 30 participants (see Kuperman et al. (2021) for full details). A generalized additive model was fitted to total reading time on target words with entropy, frequency, spelling scores, items and trial serving as predictor variables. An interaction between frequency and entropy was modeled as a tensor product in the GAMM to allow for a non-linear relationship with the dependent variable. The model showed a significant interaction of entropy by frequency (p = 0.008). The detailed results with other predictors can be found in Table 2 below.

Table 2. Generalized additive regression model fitted to total fixation time, smooth terms.

	edf	Ref.df	*F*	*p* value
Tensor product entropy x frequency	3.509	3.646	3.532	0.008
Smooth proficiency score	1.060	1.067	1.369	0.257
Smooth word	43.270	66.000	1.896	0.000
Smooth trial order by participant	65.054	268.000	1.261	0.000

The critical interaction of spelling entropy by frequency is visualized in Figure 1. Lines represent estimated partial effects of entropy on total reading time for different levels of frequency, from the 10th (solid) to 90th (dashed line) percentile of frequency. Figure 1 shows that as frequency of a word gets higher, the effect of entropy gets more salient. High frequency words with low entropy have an average total word reading duration of about 350 ms, whereas high frequency words with higher entropy values inflated total reading times by approximately 30-40 ms. In contrast, low frequency words did not show such inhibitory effect of entropy, with average total reading time of 440-450 ms across all entropy values.

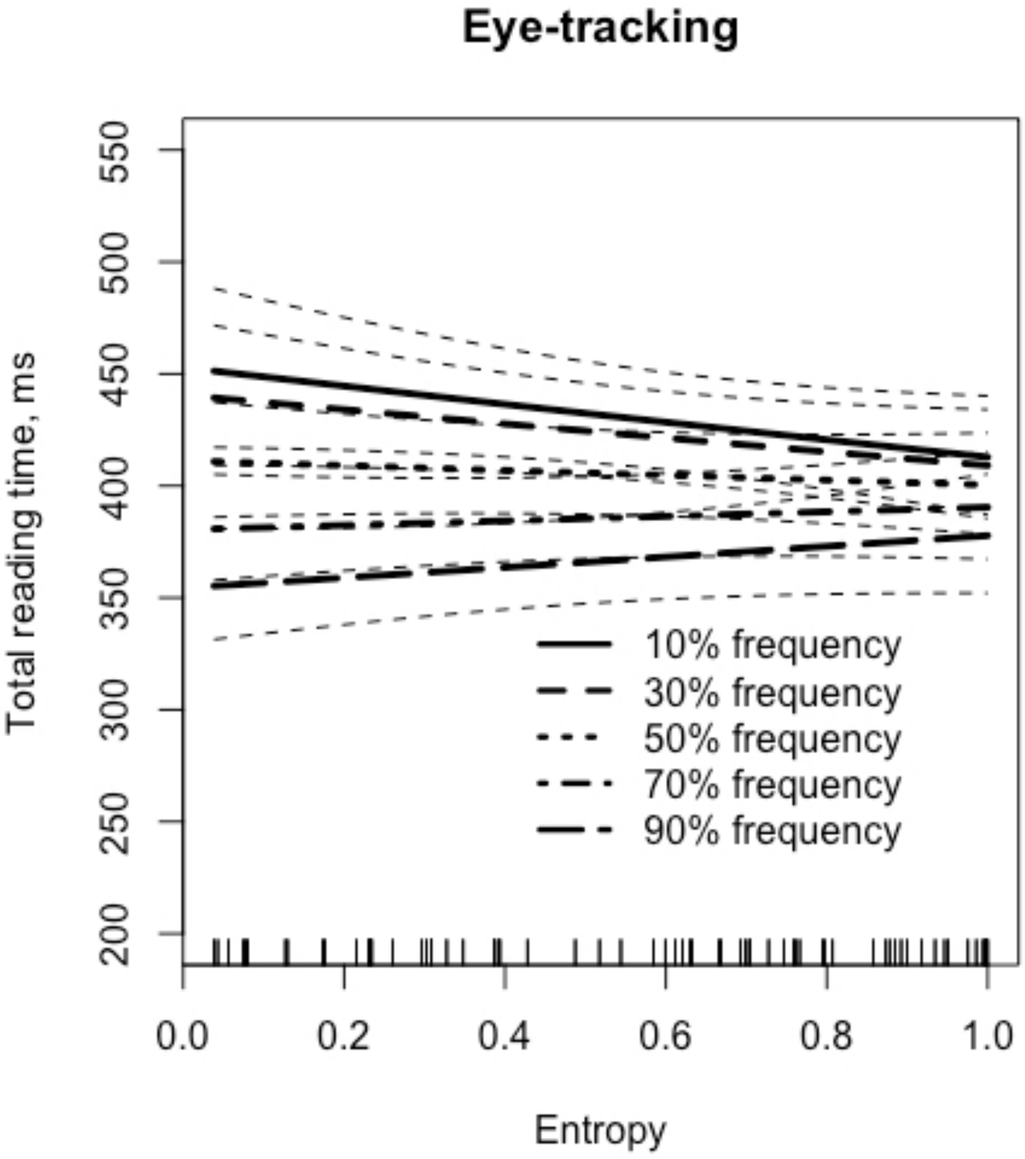

Figure 1. Frequency by entropy interaction on total reading time.

Interestingly, there was a significant frequency by entropy interaction on the number of regressions, or looking back at the word after leaving it to the right, (F = 15.333, edf = 5.457, p = 0.018), with higher entropy words initiating greater number of regressions. Moreover, this effect was pronounced more on high frequency target words. This once again proves that high entropy words are cognitively more effortful for processing, and that is again contingent on frequency of the word. No other measures of reading times on the targets showed significant entropy by frequency interaction.

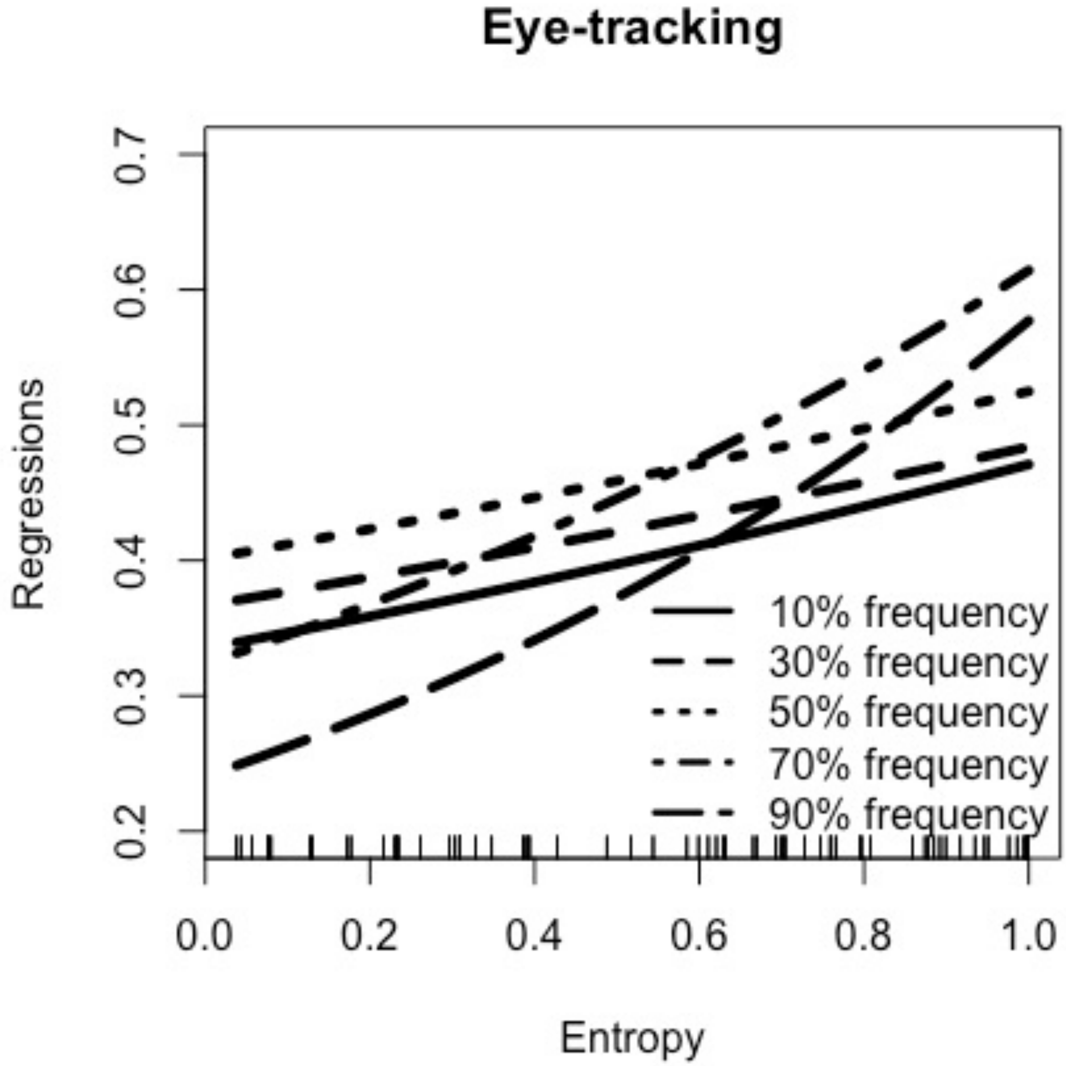

Figure 2. Frequency by entropy interaction on regressions.

These findings proved the hypothesis put forward by Rahmanian and Kuperman (2019) that the distribution of correct and incorrect spellings in our mental lexicon has an apparent effect on our reading behavior. Higher spelling entropy leads to a greater processing effort even when reading correctly spelled words. Interestingly, entropy influenced reading times to target words in a late measure, namely, total reading time, and was not reliable in earlier eye-movement measures like first fixation or gaze duration (not shown). This suggests that spelling entropy has a relatively late behavioral effect during word recognition, and likely still manifests during the re-analysis stage. This is consistent with recent ERP studies on orthographic processing (e.g., Van de Meerendonk et al.,2011; Vissers et al., 2006).

3.2 Fixation-related potentials

We present effects of critical predictors on the brain activity by time windows.

3.2.1 Entropy.

150-300 ms. The P200 component peaked at around 200 ms post-fixation and was observable in the mid-parietal region (see Figure 2) with the low entropy condition eliciting a more positive waveform than the high entropy one. The mean amplitude of 1.11 μV was observed for the low entropy condition, and 0.36 μV for the high entropy condition at 150-300 ms window. This difference in amplitudes was statistically significant (F(1,29) = 5.93, p = 0.021).

300-500 ms. There was a marginally significant effect of spelling entropy at 300-500 ms window (F(1,29) = 3.27, p = 0.081) at the region of interest where a canonical N400 effect tends to occur (see Figure 2). A more negative waveform was observed for the high entropy condition after the fixation on a target word with mean amplitudes of -0.05 μV for high and 0.34 μV for the low entropy condition.

500-700 ms. Finally, we observed a marginally significant effect of entropy in a later time window, at 500-700 ms (F(1,29) = 3.23, p = 0.083). High entropy target words yielded more negative amplitudes than words with low entropy values, with mean amplitudes of 1.19 μV and 1.35 μV, respectively.

To summarize, Figure 2 demonstrates that two entropy conditions show consistent amplitude differences. The high entropy waveform generally showed a sustained negative trend over the whole epoch. The FRP analysis using repeated measures ANOVA showed the effect of entropy at a relatively early time window. Average amplitude values started to show a statistically significant difference already at a 150-300 ms window. Later time windows, 300-500 ms and 500-700 ms, showed a negative deflection for the high entropy condition as well, however, the results were shown to be not statistically significant.

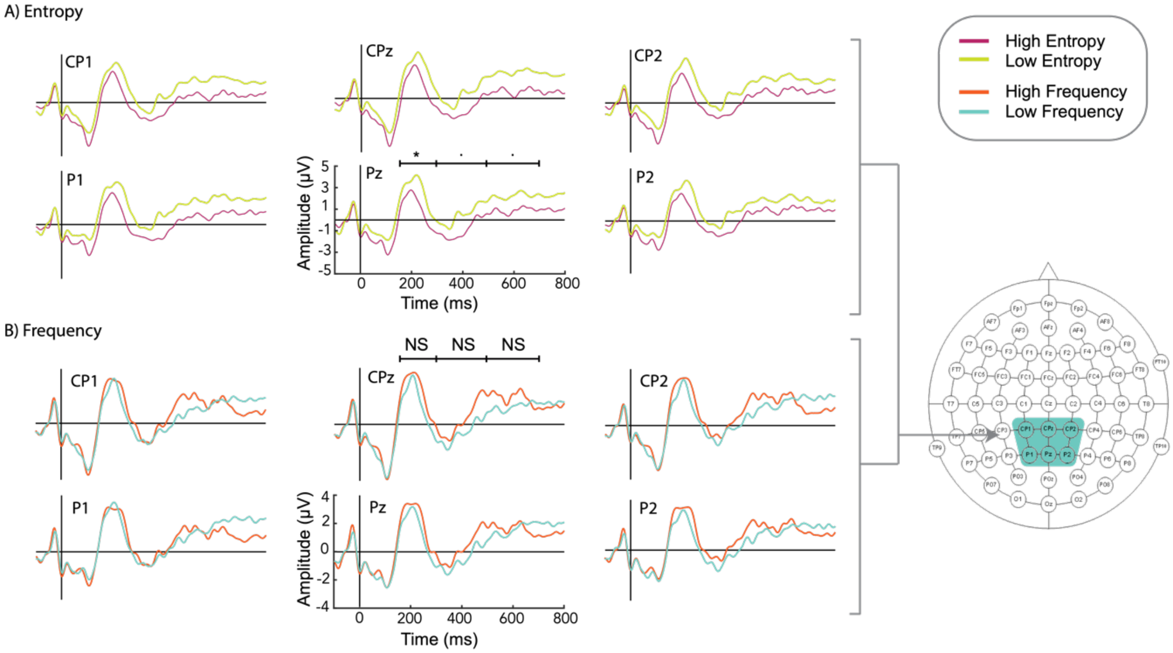

Figure 3. Grand averaged fixation-related potentials (FRPs) time locked to first fixation on the target word for high and low entropy conditions. A peak preceding fixation onset at 0 ms is a myogenic spike potential at saccade onset (see Dimigen et al., 2011).

3.2.2 Frequency and Frequency x Entropy interaction. There was no observed effect of word frequency on any of the time windows we analysed: 150-300 ms: (F(1,29) = 0.60, p = 0.444), 300-500 ms: (F(1,29) = 0.35, p = 0.559), and 500-700 ms: (F(1,29) = 0.032, p = 0.857). Contrary to the eye-tracking results, there was no statistically significant interaction of entropy by frequency either (all ps > 0.167).

Taken together, the present results, along with the eye-tracking findings, support the hypothesis that the distribution of spelling variants in written language has an effect on word recognition processes. Moreover, this effect is relatively early. This study is thefirst to show spelling entropy affecting FRPs during natural sentence reading.

4 General Discussion

Studies of spelling errors and their behavioral implications provide insight into both orthographic learning and component skills important for proficient reading. Historically, spelling errors are viewed as a consequence of deficient orthographic representations. The present paper contributes to a recent line of research that asks whether the co-existence of alternative orthographic forms for a given word in a language may also cause deficiencies in orthographic representations of that word. Eye-tracking studies across languages (Rahmanian & Kuperman, 2019 and Kuperman et al., (2021)) report that the probability distribution of spelling variants (*giraffe* vs *girafe*) – and more specifically, entropy of this distribution – influences reading behaviour, with higher-entropy words being more difficult to process than lower-entropy ones. Since entropy is a measure of the average difficulty of discriminating one alternative from an available set, higher-entropy words are the ones in which zooming in on one spelling variants is relatively effortful: in such words, spelling variants are either more numerous or similar to one another in their probability of occurrence or both. In sum, behavioral data suggest increased cognitive effort due to competing orthographic representations that emerged in the reader's mental lexicon as a result of exposure to spelling variants.

Pure behavioural methods of investigation have their limitations when it comes to the dynamics and exact timing of certain cognitive processes in the readers brain. Analysis of fixation related potentials (FRP) registered via electroencephalography is an alternative technique, which can help to explore questions related to the time-course of particular effects during word recognition at the exact point of time during a fixation.

The objective of the study was to investigate neurophysiological support for a behavioral effect of spelling entropy shown by Kuperman et al. (2021). The target stimuli were two-character words in Mandarin Chinese embedded and presented in sentences for natural reading. The words were spelled correctly but represented a broad range of spelling entropy. By time-locking eye-tracking and EEG, we were able to observe neural processes of the reading behaviour under natural reading conditions. Based on behavioral data, we expected to observe indices of more effortful cognitive processing in higher-entropy than in lower-entropy words. Furthemore, based on results from spelling violation studies (see the Introduction), we hypothesized that the differences could emerge either in earlier, 150–300 ms, and/or in later, 300–500 ms and 500–700 ms, time windows. This time-course would be indicative of the processing mechanisms triggered by mental activation of competing orthographic representations even though the readers were presented with correct orthographic forms.

Results from the FRP analysis showed significant amplitude differences for low and high entropy words only in the early (150–300 ms) time window. We interpret this finding to be in line with the prior EEG literature on spelling errors. Namely, early components (e.g., P130) arguably reveal a state of indecision at the orthographic processing stage resulted from a representational conflict. Moreover, we argue that amplitude differences in the early time window reflect a competition between orthographic variants of the correct and incorrect spelling.

A logical possibility existed that the effect of entropy on the neural activity would be reflected in later time windows. Amplitude differences in later time windows are associated with general stage of re-analysis once the representational conflict occurred. However, the statistical test showed only a marginally significant outcome. As demonstrated in previous research, the P600 component during this late time window shows an increased amplitude only in the case of a ’strong enough’ mismatch between the reader’s expectation of the upcoming input and the actual input. We suggest that the representational conflict as created by the effect of spelling entropy is not strong enough to initiate the re-analysis mechanisms in the sequence of processes during word recognition. It is also possible that the absence of P600 is explained by the fact that the correct orthographic form was presented to participants.

The present findings support the finding in Kuperman et al. (2021) that spelling entropy is a reliable measure of orthographic competition between possible representations and plays an important role in word recognition processes. One point of discrepancy between neurophysiological and behavioral data recorded within participants is that only the main effect of entropy was found in the EEG data rather than an interaction of entropy by frequency seen in the eye-tracking data. This discrepancy is perhaps not entirely surprising given that neurophysiological paradigms do not often find a reliable word frequency effect in natural sentence-reading experiments. Typically, word recognition/reading experiments using EEG are designed using rapid serial visual presentation technique (RSVP), where sentences are shown one word at a time (see Dambacher, Kliegl, Hofmann, and Jacobs, 2006; Hauk and Pulvermüller, 2004; Sereno, Rayner, and Posner, 1998). In these experiments, frequency effects were shown at a very early time window, 100-120 ms (e.g., Hauk, Davis, Ford, Pulvermüller, and Marslen-Wilson, 2006. However, ERP research looking for word frequency effects under natural reading conditions is quite scarce (e.g., Dimigen et al., 2011; Kornrumpf, Niefind, Sommer, and Dimigen, 2016) and, to the best of our knowledge, it consistently reports null effects of word frequency (see Kretzschmar, Schlesewsky, and Staub, 2015). In our experiment, there was no main effect of frequency on FRP amplitudes either. Thus, in the absence of the word frequency effect in the EEG record the absence of a much subtler entropy x frequency interaction is perhaps expected.

Considering the timelines of the observed effects, EEG analysis demonstrated that the effect of spelling entropy took place relatively early in time during visual word recognition: the time window that showed significance was 150-300 ms, see discussion above. In the eye-tracking results, on the contrary, the effect of frequency by spelling entropy interaction was captured by a late measure of eye fixation duration, total reading time, with values ranging from 380-470 ms. Taken together, these findings illustrate the paramount utility of the simultaneous consideration of the two time-sensitive paradigms of studying cognitive processing. Specifically, that the effect of spelling entropy has first appeared in the brain activity and then registered in the behavior of the visual system guided by the brain accords with a natural progression of cognitive processing. In this regard, our findings stand out from a wealth of observations that demonstrate a "paradox of brainless behavior", i.e, a situation when effects of a variable on neural activity lag behind the effects of the same variable on one's behavior (see Schmidtke et al., 2018; Schmidtke and Kuperman, 2019). The paradox has so far been observed widely but only in studies that report either only the eye-movement data or only the EEG data, separately. Moreover, in line with the common practice of neurophysiological research, all EEG studies that demonstrated the said paradox registered the brain activity either during lexical decision on individual words or the RSVP presentation of sentence. Beyond pursuing a question about orthographic representation in one's mind, the present co-registration study fills an important methodological lacuna enabling us to deepen our understanding of processes involved in visual word recognition. We believe that by co-registering behavioural and neurophysiological methods, researchers are able to not only reveal an existence of a certain linguistic effect, but able also to chart the timing of that particular effect and its progression through neural and behavioral activity.

In sum, this study reveals for the first time a neuropsychological signature of the competition between orthographic representations of correct spellings (presented in the experiment) and alternative spellings (learned through exposure to the natural written language). In this regard, the study uncovers an effect that both pre-dates and underpins a similar effect in reading behavior of Mandarin Chinese sentences observed within the same participants (Kuperman et al., (2021)). Moreover, it points to the timeline of the effect (within 150–300 ms post-onset) and the mechanism that is likely responsible for the emergence of the spelling entropy effect in the first place: a conflict between an expected input and the actual one. The stronger the conflict, the more competition there is between spelling variants of the word, i.e., the higher the spelling entropy of that word is. A final methodological contribution of this paper is that it utilized a co-registration of EEG and eye-tracking: we believe that this paradigm will help resolve existing controversies in the field of psycholinguistics and create an opportunity for exploring new research questions.

References

Assink, E., Bos, A. A., & Kattenberg, G. (1996). Reading Ability and the Use of Context in Orthographic Information Processing. *The Journal of Genetic Psychology*, *157*(4), 381–396. doi:10.1080/00221325.1996.9914873

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, *118* (3), 438.

Baccino, T., & Manunta, Y. (2005). Eye-fixation-related potentials: Insight into parafoveal processing. *Journal of Psychophysiology*, *19*(3), 204–215.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, *7* (6), 1129–1159.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108* (3), 624.

Burt, J. S., & Fury, M. B. (2000). Spelling in adults: The role of reading skills and experience. *Reading and Writing*, *13*(1-2), 1–30.

Chen, H.-C. (1999). How do readers of chinese process words during reading for comprehension. *Reading Chinese script: A cognitive analysis*, 257–278.

Conrad, N. J. (2008). From reading to spelling and spelling to reading: Transfer goes both ways. *Journal of Educational Psychology*, *100* (4), 869.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain research*, *1084*(1), 89–103.

Dehaene, S., Le Clec'H, G., Poline, J.-B., Le Bihan, D., & Cohen, L. (2002). The visual word form area: A prelexical representation of visual words in the fusiform gyrus. *Neuroreport*, *13*(3), 321–325.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and eeg in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, *140*(4), 552.

Dixon, M., & Kaminska, Z. (1997). Is it misspelled or is it mispelled? the influence of fresh orthographic information on spelling. *Reading and Writing*, *9*(5-6), 483–498.

Dixon, M., & Kaminska, Z. (2007). Does exposure to orthography affect children's spelling accuracy? *Journal of Research in Reading*, *30* (2), 184–197.

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., . . . Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, p300, and n400. *Clinical Neurophysiology*, *120*(11), 1883–1908.

Ehri, L. C. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in Language Disorders*.

Ehri, L. C., & Wilce, L. S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic? *Reading Research Quarterly*, 163–179.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, *142*, 27–41.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive brain research*, *1*(3), 183–192.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and cognitive processes*, *25* (2), 149–188.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes*, *8*(4), 439–483.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, *30*(4), 1383–1400.

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*(5), 1090–1103.

Helenius, P., Salmelin, R., Service, E., & Connolly, J. F. (1999). Semantic cortical activation in dyslexic readers. *Journal of cognitive neuroscience*, *11* (5), 535–550.

Holmes, V. M., & Castles, A. E. (2001). Unexpectedly poor spelling in university students. *Scientific Studies of Reading*, *5*(4), 319–350.

Hutzler, F., Braun, M., Võ, M. L.-H., Engl, V., Hofmann, M., Dambacher, M., . . . Jacobs, A. M. (2007). Welcome to the real world: Validating fixation-related brain potentials for ecologically valid settings. *Brain Research*, *1172*, 124–129.

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *44* (3), 345.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and cognitive processes*, *15* (2), 159–201.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, *29*(3), 333–347.

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of cognitive neuroscience*, *24*(5), 1104–1112.

Kornrumpf, B., Niefind, F., Sommer, W., & Dimigen, O. (2016). Neural correlates of word recognition: A systematic comparison of natural reading and rapid serial visual presentation. *Journal of Cognitive Neuroscience*, *28* (9), 1374–1391.

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from co-registration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41* (6), 1648.

Kuperman, V., Bar-on, A., Bertram, R., Boshra, R., Deutsch, A., Kyröläinen, A.-J., . . . Protopapas, A. (2021). Exposure to spelling errors affects reading behaviour across languages. *Journal of Experimental Psychology: General*.

Li, X., Gu, J., Liu, P., & Rayner, K. (2013). The advantage of word-based processing in Chinese reading: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 879.

Li, X., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive Psychology*, *58*(4), 525–552.

Liu, C.-L., Tien, K.-W., Lai, M.-H., Chuang, Y.-H., & Wu, S.-H. (2009). Phonological and logographic influences on errors in written Chinese words. In *Proceedings of the 7th workshop on Asian language resources* (pp. 84–91). Association for Computational Linguistics.

Liu, P.-P., Li, W.-j., Lin, N., & Li, X.-S. (2013). Do Chinese readers follow the national standard rules for word segmentation during reading? *PloS one*, *8* (2).

Martin-Chang, S., Ouellette, G., & Madden, M. (2014). Does poor spelling equate to slow reading? the relationship between reading, spelling, and orthographic quality. *Reading and Writing*, *27*(8), 1485–1505.

Meyberg, S., Sommer, W., & Dimigen, O. (2017). How microsaccades relate to lateralized ERP components of spatial attention: A co-registration study. *Neuropsychologia*, *99*, 64–80.

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in grammar: Form and acquisition*, 214–252.

Münte, T. F., Heinze, H.-J., Matzke, M., Wieringa, B. M., & Johannes, S. (1998). Brain potentials and syntactic violations revisited: No evidence for specificity of the syntactic positive shift. *Neuropsychologia*, *36*(3), 217–226.

Nagy, W., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of educational psychology*, *98*(1), 134.

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of cognitive Neuroscience*, *3*(2), 151–165.

Newman, R. L., & Connolly, J. F. (2004). Determining the role of phonology in silent reading using event-related brain potentials. *Cognitive Brain Research*, *21*(1), 94–105. doi:10.1016/J.COGBRAINRES.2004.05.006

Oldfield, R. C. et al. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and language*, *34* (6), 739–773.

Ouellette, G., & Sénéchal, M. (2008). Pathways to literacy: A study of invented spelling and its role in learning to read. *Child development*, *79* (4), 899–913.

Peng, R., & Chen, J. (2004). Even words are right, odd ones are odd: Explaining word segmentation inconsistency among Chinese readers. *Chinese Journal of Psychology*, *46*(1), 49–55.

Perfetti, C. A. (1997). The psycholinguistics of spelling and reading.

Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill.

Perfetti, C. A., & Tan, L. H. (1999). The constituency model of Chinese word identification. In *Reading Chinese script* (pp. 127–146). Psychology Press.

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: Identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in human neuroscience*, *6*, 278.

R Core Team, D. et al. (2015). R: A language and environment for statistical computing. version 3.1. 2. *R Foundation for Statistical Computing*.

Rahmanian, S., & Kuperman, V. (2019). Spelling errors impede recognition of correctly spelled word forms. *Scientific Studies of Reading*, *23*(1), 24–36.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of" mouses" in adult speech. *Language*, 760–793.

Rapp, B., & Fischer-Baum, S. (2015). Uncovering the cognitive architecture of spelling. In the *Handbook of adult language disorders* (pp. 75–102). Psychology Press.

Sauseng, P., Bergmann, J., & Wimmer, H. (2004). When does the brain register deviances from standard word spellings?—An ERP study. *Cognitive Brain Research*, *20*(3), 529–532. doi:10.1016/J.COGBRAINRES.2004.04.008

Schmidtke, D., & Kuperman, V. (2019). A paradox of apparent brainless behavior: The time-course of compound word recognition. *Cortex*, *116*, 250–267.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43* (11), 1793.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *Neuroreport*, *9*(10), 2195–2200.

Stowe, L. A., Rommers, J., Loerts, H., Timmerman, J., & Temmink, E. (2010). Word frequency and misspelling: Effects of presentation speed. In *4th international conference on cognitive science*.

Van de Meerendonk, N., Indefrey, P., Chwilla, D. J., & Kolk, H. H. J. (2011). Monitoring in language perception: Electrophysiological and hemodynamic responses to spelling violations. *NeuroImage*, *54*(3), 2350–2363. doi:10.1016/j.neuroimage.2010.10.022

Van de Meerendonk, N., Kolk, H. H., Chwilla, D. J., & Vissers, C. T. W. (2009). Monitoring in language perception. *Language and Linguistics Compass*, *3*(5), 1211–1224.

Vissers, C. T. W., Chwilla, D. J., & Kolk, H. H. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research*, *1106*(1), 150–163.

Wood, S. N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, *48*(4), 445–464.

Chapter 4

Effects of spacing on sentence reading in Chinese

This chapter has been published as Oralova, G., & Kuperman, V. Effects of Spacing on Sentence Reading in Chinese. *Frontiers in Psychology*, 5027.
doi:10.3389/fpsyg.2021.765335

Abstract

Given that Chinese writing conventions lack inter-word spacing, understanding whether and how readers of Chinese segment regular unspaced Chinese writing into words is an important question for theories of reading. This study examined the processing outcomes of introducing spaces to written Chinese sentences in varying positions based on native speaker consensus. The measure of consensus for every character transition in our stimuli sentences was the percent of raters who placed a word boundary in that position. The eye movements of native readers of Chinese were recorded while they silently read original unspaced sentences and their experimentally manipulated counterparts for comprehension. We introduced two types of spaced sentences: one with spaces inserted at every probable word boundary (heavily spaced), and another with spaces placed only at highly probable word boundaries (lightly spaced). Linear mixed-effects regression models showed that heavily spaced sentences took identical time to read as unspaced ones despite the shortened fixation times on individual words (Experiment 1). On the other hand, reading times for lightly spaced sentences and words were shorter than those for unspaced ones (Experiment 2). Thus, spaces proved to be advantageous but only when introduced at highly probable word boundaries. We discuss methodological and theoretical implications of these findings.

1 Introduction

One of the differences between Chinese and many other languages is that the Chinese writing system does not have inter-word spacing, thus offering no overt visual cues for identifying word boundaries. This fact gave rise to a large-scale ongoing inquiry into how Chinese readers segment print information into chunks for processing and what guides this segmentation process. The present paper contributes to this inquiry by studying the effects of introducing space symbols at specific character transitions on word and sentence recognition. We begin with a brief review of the relevant literature and orthographic system of written Chinese and follow up with an outline of the present study.

The central unit of the Chinese writing system is a box-like character which normally corresponds to a monosyllabic morpheme. A single word can consist of one or several characters (or *zi*). Similarly, one character, or *zi*, can constitute a single word or can be part of another multi-character word. A Chinese word, or *ci*, as defined by the traditional grammar, is a linguistic unit which denotes a meaning and a pronunciation, may stand alone to constitute a sentence and can be a grammatical unit on its own (Hoosain, 1991, 1992).

Chinese readers sometimes disagree on what constitutes a word or where a word's boundaries are in a given sentence. This observation is the central finding of a study by Liu, Li, Lin, and Li (2013) in which they asked Chinese readers to identify word boundaries by inserting slashes between words in a natural unspaced text. Liu et al. (2013) found that judgments about where the boundaries should be placed varied widely across participants. The average inter-rater agreement on segmentation judgments was 64%. Liu et al. (2013) also observed that Chinese raters tended to group characters into larger informational units and that their segmentation was influenced by syntactic categories. For example, they combined consecutive nouns to form a single chunk and combined function words with content words to form a single unit.

In an earlier study, Hoosain (1992) instructed Chinese speakers to segment sentences into words and also found that they had a substantial degree of disagreement on what constitutes a word boundary. Interestingly, when asked to explain their word boundary decisions, participants indicated that they aimed to separate "one thing" or "one idea" with boundaries. Hoosain (1992) explains that reading for meaning is a cause of divergent segmentation decisions, because units of meaning may go beyond character and word units. Ultimately, what a reader considers "one thing" or "one idea" could vary depending on their focus at the time of processing. Despite the abundance of evidence on word boundary disagreement, other research shows that certain word properties influence a range of reading measures in Chinese, signifying that words are psychologically real in Chinese minds (see Li, Bicknell, Liu, Wei, & Rayner, 2014).

1.1 Effects of Spacing on Reading Behavior

In order to better understand what constitutes a unit of processing in Chinese and other languages without overt segmentation cues, researchers have often introduced spaces into a normally unspaced text in experimental studies of Chinese word segmentation. This manipulation examines if spaces benefit readers of unspaced languages by facilitating the segmentation process and, importantly, increase reading speed or improve comprehension. The present study makes use of this manipulation as well.

The role of spacing is well documented in alphabetic languages with conventional inter-word spaces. For instance, when inter-word spacing is eliminated, the reading rate of English readers is slowed down by 30-50% (Rayner & Pollatsek, 1996). This is because spacing guides saccadic movements of the eye and helps word recognition in general (Epelboim, Booth, & Steinman, 1994; Pollatsek, Rayner, & Henderson, 1990). Nevertheless, this facilitatory effect of spacing is not universal across all languages, and certainly not in the languages that do not use spaces. In an eye-tracking study where Japanese speakers read spaced and unspaced texts in pure Hiragana and mixed Kanji-Hiragana scripts, Sainio, Hyönä, Bingushi and Bertram (2007) found that spaces did not facilitate text reading rate (measured in words per minute) either in the syllabic Hiragana script nor in the mixed script condition. Facilitatory effects were found only at the word-level analysis (word fixation duration measures) and only for the mixed Kanji-Hiragana script. The proposed explanation for facilitation was that in Japanese, characters frequently appear at the beginning of words, and as a result, in the mixed Kanji-Hiragana script, their occurrence serves as a segmentation cue for word boundaries (Sainio et al., 2007). In a study with another non-spaced language, Winskel, Radach, and Luksaneeyanawin (2009) tested English-Thai bilinguals when they were presented with spaced and non-spaced Thai texts and found that sentence reading times were 5% longer in the spaced condition than the non-spaced condition. The authors suggest the lack of facilitatory effects from spacing in Thai was due to the visual salience of words in the segmented text. This resulted in the words attracting more fixations, which led to an increase in sentence reading times. More importantly, Thai has certain language-specific word segmentation cues, such as letter clusters (vowels occurring before the consonants at syllable beginnings, e.g. โรค written as /o:rk/ *disease*) or tone markers (placed above syllables or lexemes, e.g. หน้าต่าง /na:2ta:ŋ1/ window), which is redundant with additional segmentation information in the form of spacing. Thus, in both Japanese and Thai, there are other visual characteristics of the printed text serving as word boundary cues that affect segmentation decisions. In these circumstances, the addition of spaces brings about null or inhibitory effects.

Similar inhibitory effects of spacing were found in some studies on Chinese reading. Bai, Yan, Liversedge, Zang, & Rayner (2008) investigated whether the introduction of spaces into naturally unspaced Chinese helps reading. They used 4 types of sentences in their spacing conditions: (i) unspaced sentences, (ii) sentences where spaces were between words, (iii) sentences where spaces were placed in positions such that non-words were created, and (iv) sentences with spaces between each character. The researchers found that readers made shorter fixations on words in condition (ii), and the longest fixations on words in conditions (iii) and (iv). Although fixation times were shorter on words demarcated by spaces, these benefits were short-lived, and no differences were found in sentence reading times whether the sentences were fully unspaced or spaced just at the word level. Another eye-tracking study by Inhoff, Liu, Wang, and Fu (1997) presented Chinese sentences in 3 conditions: normal non-spaced, word-spaced with a space between every word, and non-word spaced where spaces were inserted such that character combinations formed non-words. Results did not show any differences between conditions, neither in total sentence reading times, nor in word fixation times. Interestingly, Bassetti (2009) compared sentence reading times and comprehension rates of native and non-native Chinese readers when they read texts with inter-word spacing and unspaced texts in Chinese. Their results likewise did not indicate any signs of facilitated reading for Chinese texts with inter-word spacing in either of the groups.

On the contrary, some studies show beneficial effects of spaces in Chinese. For example, Hsu & Huang, (2000a, 2000b) found that although spacing between words did not facilitate reading, however, when a space was inserted to guide segmentation decisions in reading of overlapping ambiguous strings, sentence reading time was reduced. Interestingly, some other studies also showed the beneficial effects of spacing, but they were observed only at the word-level, with sentence reading times being identical in spaced and unspaced conditions. For instance, Cui, Drieghe, Bai, Yan and Liversedge (2014) hypothesized that spacing between words would allow for a more focused allocation of attention, which would enhance the parafoveal preview benefit compared to the control unspaced condition. In their study, using a gaze boundary paradigm with a correct and incorrect preview character, they showed that there was a bigger preview effect in the spaced condition, but only for one-character words. Cui et al. (2014) concluded that overt boundary cues enhance allocation of attention and lead to more efficient parafoveal processing in Chinese reading. In another study, Zang, Liang, Bai, Yan and Liversedge (2013) examined children and adults' eye movement behavior when they read spaced and unspaced texts in Chinese. Zang et al. (2013) showed that inter-word spacing decreased first pass reading times (first fixation duration, single fixation duration and gaze duration) in both groups, indicating that inter-word spacing facilitates the word identification process. This word-level advantage in Cui et al. and Zang et al. runs counter to a logically possible hypothesis that introduction of spaces causes upcoming words to be located further away from the current fixation and thus might decrease the efficiency of parafoveal preview[1]. However, sentence reading times in Zang et al.'s data were similar for spaced and unspaced conditions. They concluded that introducing spaces between words may help early segmentation, but the unusual visual presentation of the spaced text may cause a disruption to online global text comprehension. A trade-off between disruption and facilitation results in a statistically unreliable difference in total sentence reading times.

Cumulatively, the studies presented above indicated that, despite the short-lived advantage in reading speed at the word-level, by and large, spaces fail to significantly facilitate sentence reading times in Chinese, but do not appear to disrupt processing either.

[1] We thank the anonymous reviewer for raising this point.

1.2 Statistical Cues During Word Segmentation

Earlier investigations of whether spaces inserted at character transitions help or hinder Chinese word segmentation have led to mixed results. Importantly, to our knowledge, not all of the studies mentioned above used a range of segmentation probabilities to guide the experimental decision of where to place spaces to demarcate word boundaries in Chinese texts. Consequently, it is possible that in previous studies spaces were put in places which some readers may have found counterintuitive. Yet, the statistical probabilities of character transitions either co-occurring within a word or straddling a word boundary are known to serve as efficient cues to reading in Chinese (see Inhoff & Wu, 2005, and discussion above). For instance, Zang et al. (2016) assessed whether Chinese readers segment words according to how likely a character was to appear as a single-character word or as a part of another two-character word. Results showed that the preview benefit from the second character was reduced when the first character was more likely to be a single character word. Zang et al. (2016) proposed that the first character acted like an "anchor" to signify that there is a word boundary, and hence, any additional characters to the right of fixation were not processed to the same degree prior to fixation. In another study, Yen, Radach, Tzeng, and Tsai (2012) embedded two-character words in sentences and manipulated the contrast between the probabilities of the ending character (C2) of the target word (C12) being used as a word beginning or ending in all words containing it. They found that the probability of within-word positions affected character-to-word assignment and translated into longer reading times in lower-probability combinations of characters. In sum, Zang et al.'s (2016) and Yen et al.'s (2012) findings provide evidence that the segmentation probability of characters between and within words plays a crucial role in word segmentation and eye-movement control in Chinese.

We acknowledge that spacing is only one method of drawing readers' attention to segmentation cues, which interferes with the common visual layout of Chinese and may introduce artificial oculomotor and attentional demands on reading. Other artificial, less disruptive segmentation cues have been fruitfully used in the field, such as color grouping of words indicating a word boundary. Color marking of word boundaries consistently showed a beneficial effect on eye movement parameters (e.g., Perea & Wang, 2017; Zhou, Wang, Shu, Kliegl & Yan, 2018). We opted for the use of spacing for comparability of the present results with a broader existing literature in the field, and also for its practicality. If one of the manipulations of spacing were to lead to sizable consistent benefits in reading times at the word or sentence levels, spacing can be typographically implemented in Chinese texts for language learners or proficient readers with greater ease than, say, font coloring.

The literature above motivates the present study, which takes into account segmentation probabilities in an eye-tracking study of natural unspaced Chinese sentences and their spaced counterparts (see Zang et al., 2016). In the remainder of the Introduction, we introduce the critical experimental manipulation and predictions of our study.

1.3 The Present Study

It is logical to assume that a segmentation cue like a space is the most beneficial when it is applied in an appropriate position in a sentence, for instance, at a transition between characters that is undoubtedly a word boundary. Conversely, inserting a space between characters that undoubtedly belong to the same word is likely disruptive to reading. Yet, all too often Chinese readers disagree on where the word boundaries are (Liu et al., 2013; Wang, Huang, Yao, & Chan, 2015). That is, only a few character transitions are clearly fit or unfit for space insertion in Chinese. To our knowledge, no experimental study so far has exploited naturally occurring differences in segmentation probabilities to systematically examine the range of efficiency that spaces may offer as potential segmentation cues and the variable impact that such cues may have on word and sentence reading in Chinese.

We made use of segmentation judgments for word boundaries reported in Liu et al. (2013) and Wang et al. (2015) to create three experimental conditions based on their stimulus sentences: a natural unspaced condition; a heavily spaced condition, where spaces were inserted between a large number of character transitions (the transitions where at least 5% of raters agreed on a word boundary); and a lightly spaced condition, where spaces were inserted only in highly probable transitions. We defined a highly probable transition as a location where at least 90% of raters agreed to place a word boundary. All conditions used the same sentences and only differed in the amount of spacing. The rationale behind this setup was to explore whether insertion of spaces at transitions between characters in a sentence that varied in their suitability as word boundaries would have a detrimental or beneficial effect on reading times both in experimentally manipulated sentences and in natural unspaced sentences in written Chinese. The three conditions were distributed between two experiments conducted with two different groups of participants from the same participant pool. The first experiment included the heavily spaced and the unspaced conditions, whereas the second experiment included the lightly spaced and the unspaced conditions. The unspaced conditions were identical between the two experiments. We provide the motivation for our two experiments and their details in the Methods section below.

With a relatively large stimulus set (220 sentences), this study aims at exploring the existing uncertainty regarding the role of spacing in Chinese sentence reading and serves as a high-power extension of previous studies (see the literature review above). This study is novel in that it examines the role of probabilities of spaces as segmentation cues at a larger scale throughout entire sentences rather than in one or two specific positions in the sentence. The chosen experimental design enables us to examine the following question of interest. We ask whether spacing has an effect on Chinese reading of individual words and at the level of sentences in lightly and heavily spaced conditions. We expect to see longer sentence reading times for the heavily spaced condition compared to the unspaced one, as spaces at less probable word boundaries will be unexpected, and thus potentially disruptive for at least some readers. Additionally, adding spaces may either decrease parafoveal pre-processing efficiency and subsequently prolong reading times on individual words or helpfully guide the reader's attention to segmentation cues and thus shorten reading times (e.g., Cui et al., 2014). It is also possible that spaces at highly probable word boundaries (the lightly spaced condition) may facilitate segmentation of characters into larger meaningful units (words or phrases) and may thus facilitate reading. Conversely, the lightly spaced condition may still present a disruption to the normal reading of unspaced Chinese sentences due to the unusual presentation of the text. If this is the case, we might observe a slow-down (even if a mild one) in the lightly spaced condition as compared to the unspaced one.

Another feature of our study is that we consider all words in all sentences, rather than specifically selected lexical fragments of sentences. This enabled us to link comparisons of sentence reading times across experimental conditions to comparisons of word reading times across the same conditions. As demonstrated below, such links allow for a greater precision in achieving our goal of identifying sources of similarities and differences between different types of unspaced and spaced Chinese texts.

2 Methods

2.1 Participants

Eighty-two undergraduate students (mean age: 19.7) from McMaster University participated in the study. Forty-one participants took part in Experiment 1 (mean age: 19.5), and the remainder participated in Experiment 2 (mean age: 19.9). They were all native speakers of Chinese with Mandarin (72), Cantonese (9) and Wu (1) being their home dialects. Although there are differences in accent, lexis and minor differences in grammar between dialects of People's Republic of China, thanks to the use of a logographic script and a unified writing system, written Chinese is said to transcend dialectal differences (Li, 2006). Moreover, all our participants reported they were fluent speakers and readers of Mandarin. The mean time spent in Canada was 4.4 years, with a range of 0.5 to 16 years. All subjects had normal or corrected to normal vision. All participants received a course credit or a monetary compensation of 20 CAD for their participation.

2.2 Apparatus

Participants' eye-movements were monitored using the SR Research EyeLink 1000 system (Kanata, Ontario, Canada) at a sampling rate of 1000 Hz. The participant's head was stabilized with a chin and forehead rest. Eye movements were recorded from the right eye only. The stimuli were presented using Experiment Builder software on a white background in NSimSun fixed-size font on the monitor with a 1,024x768-pixel resolution. The distance between the monitor and participant's head was 60 cm, and characters were the size of 28x28 pixels and the size of a space (in spaced conditions) between words was equal to one-character size. One degree of visual angle included about 1.5 characters.

2.3 Materials and Design

Stimuli. We used all 100 sentences from Liu et al. (2013) and 120 sentences from Wang et al. (2015) where every transition between Chinese characters is associated with the percentage of raters who placed a word boundary in that position. Liu et al. (2013) collected their segmentation judgements from 142 undergraduate and graduate students in Beijing, whereas Wang et al. (2015) used a crowdsourcing method on the CrowdFlower platform from more than 120 raters who were all native speakers of Chinese. We operationalized this percentage as a word's segmentation probability. What probability threshold to choose for the insertion of spaces is a design decision that can influence the reading strategy in both the spaced and unspaced conditions, as well as the role of spacing and that of segmentation probabilities. No single choice of a probability threshold is optimal. For instance, limiting insertion of spaces to only high-probability transitions would reflect a very small fraction of segmentation preferences among readers whose natural consensus on word segmentation is around 64% (Liu et al., 2013). Moreover, that would only cover a small subset of cases in which readers have to make a segmentation choice. On the other hand, allowing spaces at most transitions, including ones that are viewed as valid word segmentation cues by only a small fraction of readers (i.e., low-probability transitions) will offer a greater sample of segmentation choices, but will make reading of spaced texts less naturalistic. A full investigation of the interplay between segmentation probability and spacing requires a series of studies in which the probability threshold for space insertion is systematically manipulated along a range. In this study, we implemented two extremes of segmentation probability as realized in our two conditions: a heavily spaced condition, where we inserted a space between characters if the segmentation probability of that transition was 0.05 or higher (i.e., if 5% or more of raters put a word boundary at that transition in the rating task); and a lightly spaced condition, where a space was inserted between characters if the segmentation probability was 0.90 or higher (i.e., if 90% or more raters identified it as a word boundary).

Effectively, in the heavily spaced condition spaces were only missing in the between-character transitions that were not considered a suitable word boundary by virtually any rater. We opted for a low-probability threshold for space insertion to make sure that our spaced condition is a true counterpart to the unspaced condition, where readers' decisions on how to segment characters into words are made based on both the low- and high-probability character strings. Also, as our literature survey above demonstrates, the case of spacing in high-probability character transitions is better studied, while the full inter-word spacing option in written Chinese is only used in a handful of studies (e.g., Inhoff et al., 1997). Even with our lax inclusion criteria of 5% in the heavily spaced condition, the median segmentation probability of spaced transitions in this condition was 93% (Table 3). Thus, most of the target transitions in the heavily spaced condition were supported by the consensus and the number of spaces was no more than a half of the number of characters in every sentence (see example in Table 1). The number of inserted spaces was obviously smaller in the lightly spaced condition, which had a 90% threshold of the raters' consensus as a spacing threshold, see Table 1. We reasoned that if spacing is beneficial for reading of Chinese, such a condition will create the best environment for the benefit to materialize.

In the unspaced condition, sentences were presented in their conventional form, without spaces. A spaced counterpart (lightly and heavily spaced) was created for every original unspaced sentence. Examples of stimuli can be found in Table 1 below. Experiment 1 presented one group of readers with the unspaced and heavily spaced sentences, while Experiment 2 presented another group of readers with the (same) unspaced and lightly spaced sentences. In each experiment, two counterbalanced lists presented a mixture of unspaced and (heavily in Experiment 1 or lightly in Experiment 2) spaced sentences, such that every participant was presented with one of the lists and saw each sentence in only one format. Each list contained 110 spaced and 110 unspaced sentences. Sentences appeared on a single line, with a minimum of 19 characters and a maximum of 42 characters.

Table 1. Example Sentence with Two Spacing Conditions and Segmentation Probabilities between Words

Condition	Sentence
Normal unspaced	中国拥有巨大的市场，在游戏产业中无疑应当成为主导力量。
Heavily spaced	中国 拥有 巨大 的 市场，在 游戏 产业 中 无疑 应当 成为 主导力量。
Lightly spaced	中国 拥有 巨大的 市场，在游戏产业中 无疑 应当 成为 主导力量。
Segmentation probabilities between words	中国 1.0 拥有 1.0 巨大 0.31 的 1.0 市场，在 0.88 游戏 0.62 产业 0.60 中 0.95 无疑 0.95 应当 0.90 成为 1.0 主导 0.52 力量。
Translation	China has a huge market, and it should undoubtedly become a dominant force in gaming industry.

Procedure. Upon arrival, participants signed a consent form and were instructed to read sentences silently for comprehension. Yes/no comprehension questions appeared after roughly 30% of sentences. Participants were asked to answer 'yes' by pressing '1' on the keyboard in front of them, and 'no' by pressing the '0' button. After setting up the eye-tracker, a 9-point calibration was conducted. We required calibration accuracy to be below 0.5 degrees of the visual angle to proceed with testing. If the validation procedure was not successful, the participant was removed from the study. Then, participants read six practice trials prior to presentation of the critical stimuli. Each trial started with a drift correction procedure, which was initiated with a dot placed at the location of the first character of a sentence. After finishing reading each sentence, participants were asked to fixate on a grey box in the lower right corner of the screen. Once the box was fixated for 200 ms, the screen was changed to display the next sentence. After the reading task, all subjects completed the LEAP-Q (Language Experience and Proficiency) questionnaire for every language they were fluent in (Marian, Blumenfeld, & Kaushanskaya, 2007). Since we tested readers of Chinese outside of China, this information was important to assess their proficiency in reading Mandarin, as well as their degree of exposure to other languages. The whole experiment lasted about 60 minutes. Both experiments had identical procedures.

2.4 Variables

We considered effects of spacing and control covariates at the level of word and sentence. The unit of analysis at the word level was the interest area contained between two spaces in the spaced condition of each Experiment. We contrasted these interest areas with respective fragments of written sentences in the unspaced conditions of Experiments 1 and 2. In the heavily spaced Experiment 1 those interest areas were obviously shorter than in the lightly spaced Experiment 2. For simplicity, we label these interest areas 'words' in all conditions. The word level used the following dependent variables: first fixation duration, gaze duration (summed duration of all fixations made on a word in the first pass, prior to a saccade to another word), and total fixation time (summed duration of all fixations on a word). First fixation duration and gaze duration are early measures of lexical access, while total fixation time is considered a cumulative measures of word processing. Joint consideration of these measures can point to the time-course of the spacing effect on word reading.

The sentence level analysis has sentence as a unit and recruited the following dependent variables: sentence reading times (i.e., the total time spent reading a sentence) and comprehension rate (rate of correct responses to comprehension questions). Sentence reading time taps into the amount of cognitive effort that subjects experience when reading spaced or unspaced sentences, while comprehension rate taps into the effect of spacing on comprehension. We also considered total saccade duration (summed duration of all saccades in the sentence) and total number of saccades per sentence, see rationale and analysis below.

2.4.1 Independent Variables. The critical variable was the experimental condition of spacing with three levels: unspaced (identical in Experiments 1 and 2), heavily spaced (Experiment 1) or lightly spaced (Experiment 2). For the word level analysis, word length in characters, word position in a sentence and position of a sentence in the experiment were included as controls. Because all our texts are identical – with a sole exception of spacing – we do not consider the many lexical predictors known to affect eye-movements, e.g., word frequency, predictability, and spatial density: these predictors are kept constant across conditions.

For the sentence level analysis, sentence length in characters (including spaces) was taken into account as a control. We also considered the position of a sentence in the experiment as a potential predictor of reading times at the sentence level. If found, such effects may indicate habituation to the unusual presentation in the spaced condition and perhaps development of a strategy towards using spaces as segmentation cues. With regard to individual differences, we considered years of education, reading comprehension in Mandarin and years spent in Canada to predict sentence reading times. We further tested interactions of these participant variables with spacing.

2.5 Statistical Considerations

Durational dependent variables (measured in ms) showed skewed distributions and were log-transformed, as indicated by the Box-Cox test, in order to obtain a more symmetrical distribution and conform with the requirements of regression modeling. This is in line with recommendations from Baayen and Milin (2010). The comprehension rate scale was a distribution of 0 and 1, where 0 stands for an incorrect answer and 1 stands for a correct response. Logistic regression was fitted to explore the effect of spacing on comprehension rate.

We used library lme4 version 1.1-19 (Bates et al., 2018) in the statistical software platform R 3.4.3 (R Core Team, 2017) to fit linear mixed-effects models to calculate the effect of multiple predictors on each dependent variable mentioned above. The model utilized sentences and subjects as random intercepts, which allowed us to examine systematic effects considering the variability across participants and testing items. We further modelled by-participant contrasts of spacing condition as random slopes. Since this step led to consistent failure-to-converge errors in regression models, we removed this random effect (Barr, Levy, Scheepers, & Tily, 2013). The fixed effects in our models are described in the Independent Variables section above.

Our further model selection process involved fitting fully defined models (with independent variables as described above) and then back-fitting the model to retain significant fixed effects and obtain a final, best-fit model. Specifically, we used the likelihood ratio method for model comparison to identify whether removal of a predictor has led to a significant decrease in the model performance. Predictors that did not lead to such a decrease were removed with the exception of the critical predictor of experimental condition. At each step, no more than one predictor was removed, and the model was refitted; the process was iterated until removal of any predictor in the model (except for that indicating experimental condition) led to a significant loss in the model performance.

Justification of this practice is outlined in Baayen, Davidson, and Bates (2008). In consideration of space, we do not publish all regression models involved in the back-fitting process: these models are available upon request from the authors.

When fitting each model, in order to eliminate the influence of outliers, we also removed residuals that exceed 2.5 standard deviations, see Baayen and Milin (2010). The models in which critical predictors and interactions reached statistical significance are reported in the Supplementary materials S1.

We chose to confirm our critical conclusions by estimating the amount of support for the null or alternative hypothesis by calculating the Bayes Factor. The Bayes Factor quantifies the ratio between the likelihood of the data under the alternative hypothesis and the likelihood of the data under the null. To estimate the Bayes Factor, we followed the procedure outlined in Masson (2011): we extracted the Bayesian Information Criterion (BIC) value for the target model and compared it to the BIC value of a model without predictors of interest (the 'null model'). The Bayes Factor can be approximated as the natural exponent raised to the power of half the difference between the BICs of two models (see Masson, 2011). Following Jeffreys (1961), a Bayes Factor (BF) value below 1/3 indicates moderate support for a null hypothesis (and above 3 for the alternative) and those below 1/10 indicate a strong support for the null hypothesis (and above 10 for the alternative). We have also estimated the size of all critical effects by means of Cohen's *d* for a comparison of the two groups formed by experimental conditions.

3 Results

In total, seven participants were excluded from Experiment 1. Data from two participants were discarded due to poor calibration (2042 observations, 1.4%). Another three participants were excluded from the analysis due to excessive skipping rates (8416 data points, 5.8%), one participant was excluded due to zero answers recorded when answering comprehension check questions and one more was removed due to an at-chance comprehension rate (7588 observations, 5.2%). Similarly, analysis of Experiment 2 excluded two participants as none of their answers to comprehension questions were recorded and one more participant was removed due to an at-chance comprehension rate (6396 observations, 7.3%), Thus, this left us with a pool of 72 participants for two experiments.

For word level analysis, after removal of a total of ten participants and sentence-initial and -final interest areas, we had a pool of 121,494 interest areas (83.03%) for Experiment 1 and 65,661 (75.5%) interest areas for Experiment 2. We further trimmed the bottom and top 1% of fixations from the distribution of total fixation time (Experiment 1: 1211 observations (0.8%); Experiment 2: 1948 observations (2.2%)). At this point, this trimming resulted in 120,283 data points for Experiment 1 and 63,713 for Experiment 2. This full dataset was used to calculate the skipping rate, which was around 52% for Experiment 1 and 25% for Experiment 2. According to Chen et al. (2003) the probability of skipping tends to be much higher in Chinese readers than in English readers (42% vs 20%). After removing skipped words, we had a total of 61,779 observations for both conditions (heavily spaced and unspaced) in Experiment 1 and a total of 46,656 data points for lightly spaced and unspaced conditions in Experiment 2 for word-level fixation time analysis. For sentence level analysis, after removing ten participants, we had a pool of 7,330 sentences for Experiment 1 and 8322 sentences for Experiment 2. We further trimmed the bottom and top 2% of the sentence reading time distribution (629 trials, 4%). In total, Experiment 1 had 7,036 sentences and Experiment 2 had 7,987 sentences that entered sentence level analysis.

Below we begin with reporting the main effect of spacing on eye-movements at the word and then sentence level for each of the two experiments separately. Table 2 reports descriptive statistics for independent and dependent variables (see the Variables section).

The mean comprehension rate of two experiments was 87.5%, which indicates that participants generally had a good comprehension of experimental sentences. No difference in accuracy was observed between the heavily spaced and unspaced conditions and lightly spaced and unspaced conditions, (all *p*s > 0.626). Also, there was no statistically significant difference in comprehension scores between spaced and unspaced conditions in both experiments when participants' individual measures (e.g., years of education) were added as co-variates, all *p*s > .11 (models not shown, available upon request).

Table 2. Descriptive Statistics of Independent and Dependent Variables across Three Conditions

Variable	Exp	Condition	Range	Mean	Median	SD	Range of log values
N of trials			1:220				
Segmentation probabilities			0.06:1	0.75	0.93	0.30	
N of space-separated items in a sentence	1	HS	11:31	11.18	11	6.70	
	2	LS	5:22	7.24	7	4.50	
N of characters in a sentence			19:42	31.35	32	4.90	
Sentence Reading Time, ms	1	HS	1030:8554	3158	2795	1473.90	6.94:9.05
		US	1035:8547	3160	2819	1494.42	6.94:9:05
	2	LS	1214:9109	3442	3091	1556.27	7.10:9.12
		US	1216:9110	3511	3178	1589.36	7.10:9.12
First Fixation Duration, ms	1	HS	50:947	221.37	202	85.60	3.91:6.85
		US	50:976	231.93	213	91.51	3.91:6.88
	2	LS	51:995	246.76	223	105.46	3.93:6.90
		US	51:984	246.50	224	104.00	3.93:6.89
Gaze Duration, ms	1	HS	50:976	236.40	208	106.61	3.91:6.88
		US	50:980	246.46	218	113.88	3.91:6.89
	2	LS	51:1746	312.13	253	195.63	3.93:7.47
		US	51:1794	312.24	254	195.47	3.93:7.49
Total Fixation Time, ms	1	HS	50:980	294.27	240	164.64	3.91:6.89
		US	30:980	310.67	254	173.86	3.91:6.89
	2	LS	51:1793	438.23	344	303.64	3.93:7.49
		US	51:1794	431.74	336	299.03	3.93:7.49
Total Saccade Duration, ms	1	HS	106:4803	787.67	681	452.52	4.67:8.48
		US	106:4580	730.00	625	432.41	4.66:8.43
	2	LS	68:6482	743.59	562	652.28	4.22:8.78
		US	52:6147	749.55	579	629.84	3.95:8.72
Total Saccade Number	1	HS	7:45	18.33	17	7.71	1.95:3.81
		US	7:45	17.30	16	7.33	1.95:3.81
	2	LS	7:60	17.74	16	8.72	1.95:4.09
		US	7:60	18.11	16	8.60	1.95:4.09
Skipping rate	1	HS	0:1	0.47	0	0.50	
		US	0:1	0.51	1	0.50	
	2	LS	0:1	0.26	0	0.44	
		UN	0:1	0.23	0	0.42	

Years of education			10:23	14.02	13	2.16
Reading comprehension score			4:10	8.67	9	1.43
Years in Canada			0.5:16	4.42	3	3.81
Mean accuracy for comprehension questions	1	HS	0:1	0.88		
		US	0:1	0.89		
	2	LS	0:1	0.89		
		US	0:1	0.89		

Note: HS: Heavily spaced; LS: Lightly spaced; US: Unspaced.

3.1 Experiment 1: Heavily Spaced vs Unspaced Conditions

Word-level analysis. We explored the effect of spacing on a word by fitting separate regression models to first fixation duration, gaze duration and total reading time on a word with spacing, ordinal trial number, number of characters in a word, and word position in a sentence as predictors. All measures of the word-level analysis showed a significant effect of spacing condition, where words surrounded by spaces were read faster compared to non-spaced counterparts (first fixation duration, $\beta = -.058$, SE = .003, $p < .001$, gaze duration, $\beta = -.066$, SE = .004, $p < .001$, total reading time, $\beta = -.097$, SE = .005, $p < .001$). Thus, word level analysis showed a beneficial effect of heavy spacing. Detailed results of all three regression models can be found in Supplementary Materials available online (Tables S1 a, b; S2 a, b; and S3 a, b).

Sentence-level analysis. We fitted a linear mixed-effects model to log-transformed sentence reading time as a dependent variable and spacing condition as a critical predictor. Sentence length (in characters) and an ordinal trial number served as controls. We observed a significant positive main effect of sentence length on sentence reading times, $\beta = .084$, SE = .011, $p < .001$: unsurprisingly, it took more time to read longer sentences. Total reading times for sentences appeared to be numerically almost identical across experimental conditions, 3158 vs 3159 ms, ($d = .001$). This effect was not significant when controlling for other predictors, $\beta = .007$, SE = .008, $p = .388$. The Bayes Factor analysis indicated extremely strong evidence in favor of the null effect of spacing, BF < .001. Detailed results of both regression models can be found in Supplementary Materials (Tables S4a, b and S5a, b).

The results of the sentence analysis are consistent with previous studies, which mainly showed that reading times are statistically identical for spaced and unspaced sentences (e.g., Bai et al., 2008). Yet they may appear unexpected given that heavy spacing granted readers a small but significant advantage in speed at the word-level in Experiment 1 (see Figure 1 below). This word-level advantage was apparently cancelled out by other factors when accumulated over a sentence.

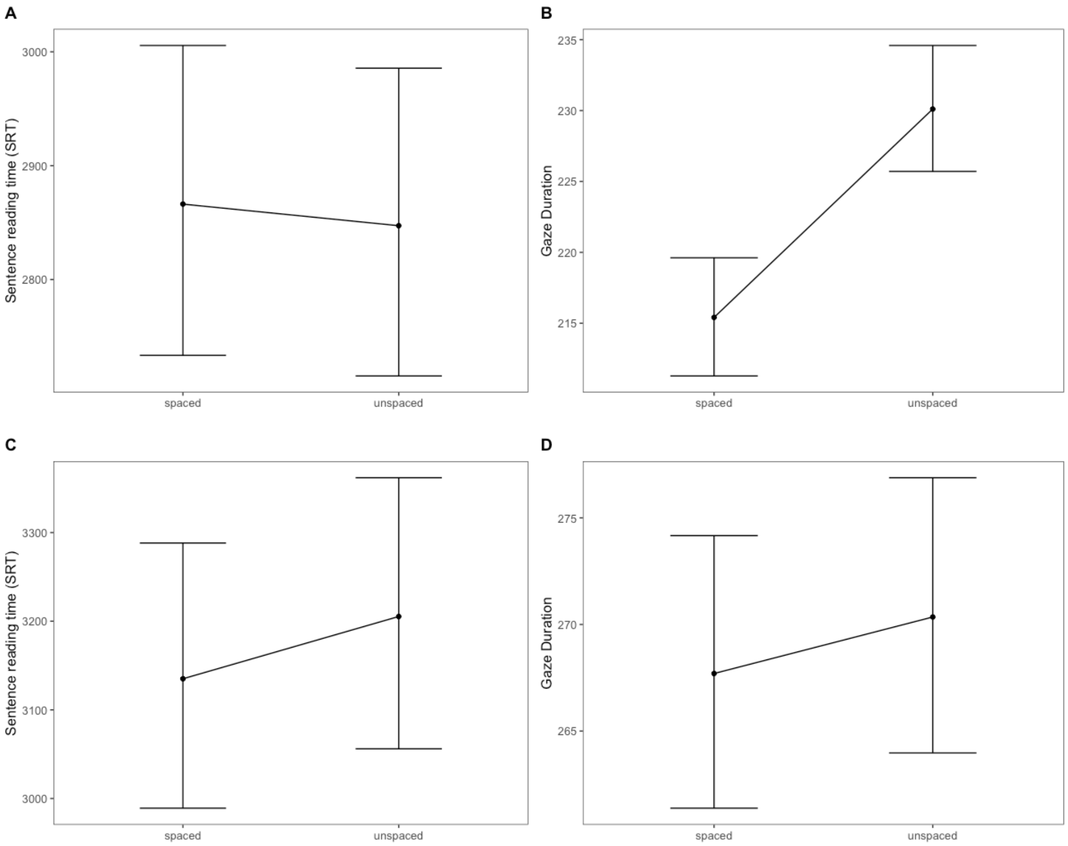

Figure 1. Top left (A): Partial effects of spacing (heavily spaced vs unspaced) on sentence reading times, Experiment 1. Top right (B): Partial effects of spacing (heavily spaced vs unspaced) on gaze duration, Experiment 1. Bottom left (C): Partial effects of spacing (lightly spaced vs unspaced) on sentence reading times, Experiment 2. Bottom right (D): Partial effects of spacing (lightly spaced vs unspaced) on gaze duration, Experiment 2. Error bars stand for 95% confidence intervals.

We examined potential sources of this discrepancy. First, participants skipped more words in the unspaced condition than in the spaced one (51% and 47%, a small but a highly reliable 4% difference, $\chi 2$ = 234.12, df = 1, p-value < 0.001). Thus, even though each individual word was processed faster, more words contributed to reading times of the spaced sentences. A more drastic discrepancy, which may explain the null effect of spacing at the sentence level, was found in the measure of total saccade duration, or the sum of all saccade durations in the sentence. We found that total saccade duration in spaced sentences was longer than in unspaced sentences by an average of 58 ms (788 ms vs 730 ms). This difference was confirmed as reliable in the mixed-effects regression model fitted to total saccade duration per sentence with sentence length and ordinal trial number in the experiment as a predictor (β = .077, SE = .017, *p* < .001). Spacing did not interact with sentence length (see Tables S6 a, b in Supplementary materials).

Saccade durations are rarely considered in studies of word reading. This is because the influence of inter-word saccades on word reading times is negligible as compared to fixation durations, and durations of intra-word saccades do not contribute to word reading times at all. However, in an experiment with sentences that have a median of 19 words, the number of saccades is considerable and saccade durations add up to a substantial proportion of sentence reading time (788 ms out of 3158 ms or 25% in the spaced condition, and 730 ms out of 3159 ms or 23% in the unspaced one). The accumulated total saccade duration is a factor that, along with other factors, appears to override the word-level advantage of spacing and lead to statistically identical reading times for spaced vs unspaced sentences.

We further investigated whether the spacing-driven difference in total saccade durations is due to an inflation in the duration of individual saccades in the spaced condition or an increase in the average number of saccades (or fixations) in this condition, or both. The former option may arise because spaces introduce an extra character at every potential word transition, which also adds a disruption to the benefit of the parafoveal preview. Thus, spaced sentences might elicit intra-word saccades that need to be longer in amplitude and in duration. The latter option might stem from a smaller number of skips (and hence a larger number of fixated words and of inter-word saccades) in the spaced condition. The follow-up analyses revealed that average saccade duration was nearly identical in the two conditions (42.08 vs 42.16 ms). Notwithstanding, the spaced condition came with a significantly higher total number of saccades per sentence than the unspaced condition (18.33 vs 17.30). This contrast was confirmed as statistically reliable (β = 1.016, SE = .280, *p* < .001) in the regression model with sentence length as a control: spacing and sentence length did not interact (see models in Tables S7 a, b). In sum, the processing advantage seen in the spaced condition at the word-level is cancelled at the sentence level, because spaced sentences elicit a larger number of saccades and fixations and—once the durations of those saccades are accounted for—come with the same total processing effort compared to their unspaced counterparts.

3.2 Experiment 2: Lightly Spaced vs Unspaced Conditions

Word-level analysis. This analysis used the same dependent and independent variables as in Experiment 1. First fixation duration and total reading time analysis did not show any significant effect of light spacing, β = -.004, SE = .004, *p* = .305, β = .001, SE = .008, *p* = 0.951, respectively. Interestingly, another eye-tracking measure, gaze duration, which captures time spent on a word during first pass reading, showed a beneficial effect of space at the significance threshold, β = -.010, SE = .005, *p* = .041. Detailed results of the regression model can be found in Supplementary Materials (Tables 8 a, b).

Sentence-level analysis. As in Experiment 1, mixed-effects regression model was fitted to log transformed sentence reading times including spacing condition, sentence length, and trial number as predictors. As expected, we observed a significant main effect of sentence length, β = .102, SE = .011, *p* < .001, meaning that longer sentences took longer to read on average. Contrary to the results of the first experiment, the difference in sentence reading times was significant and indicated a speed advantage for lightly spaced sentences, β = -.022, SE = .007, *p* < .001. Sentences were read faster if spaces were inserted in highly probable word transitions (respective means 3442 ms vs 3511 ms, with a relative difference of 2%; *d* = .05). The advantage of light spacing was then confirmed at both the word-level (gaze duration) and sentence-level: in both cases the effects were significant but small in size (see Figure 1). Detailed results of the regression model can be found in Supplementary Materials (Tables S8 a, b and S9 a, b).

By comparing the results from Experiments 1 and 2, we observe that the amount of spacing presented to our readers modulated their reading times. Heavily spaced sentences were read in the same amount of time as unspaced ones (Experiment 1), whereas Experiment 2 showed that spaces can bring advantage in reading speed when they are placed only at highly probable word transitions. In other words, spaces are only advantageous when introduced in positions where the majority of readers agree to place a word boundary.

Similar to Experiment 1, we further explored how the word-level findings link to sentence-levels ones, looking at skipping rates, and total saccade number and their duration in Experiment 2.

Skipping rate for the unspaced condition was 23.3%, while for the lightly spaced condition it was 26.0%. A chi-square test showed that this difference is statistically significant (χ2 = 60.32, df = 1, p-value < 0.001): there were more words skipped in the lightly spaced condition. Thus, more words contributed to the reading times of the unspaced sentences, which partially explains the inflated sentence reading times in that condition. Additionally, we explored if the number of saccades and their duration contributed to shortened sentence reading times in the lightly spaced sentences. Total saccade duration per sentence did not show any significant difference between lightly spaced and unspaced sentences (β = -.023, SE = .015, *p* = .130). Although the number of saccades was numerically smaller for spaced sentences (mean: 17.74 saccades for spaced, and 18.11 for unspaced), regression analysis revealed only a marginally significant difference (β = -.447, SE = .247, *p* = .071). To conclude, it was mainly the shortened fixation durations on words and a higher skipping rate in the lightly spaced condition that brought the advantage in sentence reading times to this condition over the unspaced counterpart.

We further examined the potential role of individual differences in the participants' education level, subjective assessment of Mandarin reading comprehension or duration of stay outside of China. None of these measures turned out to be predictive. Years spent in Canada or years of education did not show any effect on word or sentence reading times in Experiment 1 or 2. Higher subjective evaluation of reading comprehension was associated with shorter sentence times ($p = 0.041$). Critically, none of the measures modulated the effect of spacing on either word or sentence reading times.

4 Discussion

Chinese does not have overt visual markers to separate words in a sentence, and the very notion of a word in this language is debated. There is no definitive consensus between Chinese readers on word boundaries, and their decisions on how to segment words in a sentence are contingent on a number of syntactic and semantic factors (Liu et al., 2013). This has led researchers to the question of how readers of Chinese segment a continuous sequence of characters into processing units and whether word units have a psychological reality in Chinese. A common approach to this question, which we also followed, is to artificially introduce spaces into naturally unspaced sentences. Previous research on the effects of spacing in Chinese sentences gave rise to mixed conclusions. Reports vary in whether these effects are facilitatory or inhibitory at the word level, and whether they exist at the sentence level (see the Introduction). Furthermore, while statistical probabilities of transitions between characters have long been recognized as a factor influencing mental segmentation of Chinese sentences, these probabilities have only been manipulated in a handful of studies (Yen et al., 2012; Zang et al., 2016) and, to our knowledge, not in conjunction with spacing manipulations.

Our study offered an examination of the effects of spacing on reading Chinese sentences by comparing natural unspaced sentences with counterparts that were either lightly spaced (spaces only at high probability transitions) or heavily spaced (spaces at every probable transition). The main goal of our study was to add to the currently incongruous body of evidence about the role of visual cues to lexical segmentation in Chinese reading by investigating the role of segmentation probabilities in reading artificially spaced text. We pursued this question by recording eye-movements in a sentence reading study in Chinese where participants read either conventional unspaced sentences or their spaced counterparts for comprehension. We also aimed at pinning down the specific sources of similarities and differences between spaced and unspaced texts by linking word-reading and sentence-reading times.

4.1 Effects of Spacing in the Heavily Spaced Condition

The central result of Experiment 1 was that heavily spaced sentences and sentences without spaces took identical time to read. This is surprising, since heavily spaced sentences were spatially longer than their unspaced counterparts and should take a longer time to read. Nevertheless, this result is consistent with previous studies, which mainly showed statistically identical reading times for spaced and unspaced sentences (e.g., Bai et al., 2008). The word-level analysis showed that eye fixation durations became shorter when words were demarcated with spaces. All measures of early and late processing (first fixation duration, gaze duration, total fixation time) showed a small, but significant facilitatory advantage of having spaces as visual cues. This pervasive effect conflicts with Bai et al.'s (2008) argument that the segmentation into words appears not to happen at early stages of processing. We believe that the small effects on early eye-movement measures emerged as reliable in our study due to the higher statistical power of our dataset (see Brysbaert & Stevens, 2018, for recommended sample sizes).

Although we observed *shorter* eye fixation durations on individual words in the spaced condition, sentence-level analysis showed that this advantage was completely over-ridden at the sentence level: sentences with spaces and without spaces took identical time to read. We argue that partial explanations for this reversal come from the processing costs of spacing that are not noticeable in individual words but accumulate and become noticeable in sentence reading times, and especially the inflated total duration of saccades in the sentence. Total saccade duration, defined as a summed duration of all saccades in the sentence, accounted for about 25% of total sentence reading time in both conditions and was 58 ms longer on average in the spaced rather than unspaced sentences. We highlight the utility of total saccade duration as a measure that is largely overlooked in the studies of sentence or passage reading.

To our knowledge, this is the first study that attempts to explain contradicting results in the previous literature, which show word-level advantage of spacing but fail to do so at the sentence level. To reiterate, we found that a larger number of saccades and other factors, including a reduced skipping rate, in spaced sentences appear to inflate sentence reading times to an extent that cancels out the slight word-level advantage. In sum, when overt visual cues for word segmentation are inserted at almost every transition where segmentation is possible (though not always very probable), spacing is not a cue that increases reading efficiency in Chinese. It also does not lead to improved reading comprehension.

4.2 Effects of Spacing in Lightly Spaced Condition

Experiment 2 draws a different picture. In sentences where spaces were placed only at highly probable word transitions, a beneficial effect of spacing on sentence reading times was demonstrated. Additionally, word-level analysis showed a beneficial effect of spacing through shortened word reading times (gaze duration) and increased skipping rates in the spaced sentences. That is, both sentence- and word-level analyses showed that spaces inserted only where the majority of readers expect a word boundary is demonstrably advantageous for reading Chinese. The observed difference between heavily and lightly spaced conditions in our Experiments 1 and 2 may partly explain discrepant findings in the earlier literature. The magnitude and direction of the spacing effect is contingent on the prevalence of spacing and, even more so by the probability of the character transition interrupted by a space as a word boundary. Since these probabilities were not systematically controlled in most earlier studies using spacing, divergence in results across studies is expected. In sum, the prevalence of spacing and its allocation in a sentence does indeed modulate sentence reading times: spaces at highly probable word boundaries lead to a small (around 2% of relative difference) but reliable advantage.

The present study contributes to the existing body of knowledge on effects of spacing in the following ways. First, results from both experiments indicate that the effects of spacing are selective and contingent on the prevalence and exact positioning of spacing in the Chinese text. Contrary to some previous research, this study shows spacing to be a cue beneficial to both the Chinese word segmentation process and, in one of conditions, for sentence level processing. However, spacing only becomes beneficial when readers find spaces at suitable word boundaries, and even then, the processing advantage is minute. These findings demonstrate that segmentation probabilities are an important yet relatively under-studied factor to consider in research of Chinese reading.

Second, our joint analyses of reading times at the word and sentence level enabled us to uncover reasons for similar or discrepant processing times across experimental conditions that much earlier research left unexplained. For instance, it highlighted the role of saccades, which increase in number and duration with the prevalence of spaces and can cancel advantages conferred by spacing as a segmentation cue at the word level. We advocate the use of a largely neglected saccade analysis in eye-tracking reading studies as a useful tool for studying reading behavior.

Finally, our investigation of two extremes of segmentation probability as criteria for the placement of spaces across all sentences suggests that spacing is not an effective segmentation cue in Chinese reading. In either the heavily or lightly spaced conditions, the advantages that spacing confers at the word level, if any, are small in size and either completely cancelled out at the sentence level or diminished to the effect size of no practical importance. Most likely, further investigation of spacing at less extreme points of the probability scale will lead to a similar result. It is plausible that other methods of guiding attention through Chinese sentences (e.g., coloring or highlighting segmentation boundaries) will not lead to the presently observed increased difficulty of saccadic planning and enable the word-level advantage in processing effort to propagate to the sentence-level. An investigation that combines the use of less invasive segmentation cues with probabilistic characteristics of character transitions is a promising avenue for future research.

Funding

Gaisha Oralova’s contribution was supported by Cognitive Science of Language Graduate Scholarship, McMaster University, The Elder Family Graduate Award (2017-2018) and by the Social Sciences and Humanities Research Council of Canada Doctoral Fellowship 752-2019-1264 (2019-2021). Victor Kuperman’s contribution was partially supported by the Ontario Early Researcher award (Kuperman, PI), the Canada Research Chair (Tier 2; Kuperman, PI), the Social Sciences and Humanities Research Council of Canada Partnership Training Grant 895-2016-1008 (Libben, PI), and the Canada Foundation for Innovation Leaders Opportunity Fund (Kuperman, PI).

Acknowledgments

We thank Paige Cater for her help with data collection and Kaitlyn Battershill for proofreading our paper.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

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

Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1277.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bassetti, B. (2009). Effects of adding interword spacing on Chinese reading: A comparison of Chinese native readers and English readers of Chinese as a second language. *Applied Psycholinguistics*, *30*(4), 757-775.

Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen-Christensen, R. H., Singmann, H., ... & Grothendieck, G. (2018). Package lme4 v1. 1-19.

Cui, L., Drieghe, D., Bai, X., Yan, G., & Liversedge, S. P. (2014). Parafoveal preview benefit in unspaced and spaced Chinese reading. *The Quarterly Journal of Experimental Psychology*, *67*(11), 2172-2188.

Epelboim, J., Booth, J. R., & Steinman, R. M. (1994). Reading unspaced text: Implications for theories of reading eye movements. *Vision Research*, *34*(13), 1735-1766.

Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity: a case study of Chinese*. Lawrence Erlbaum Associates, Inc. Hillsdale, New Jersey.

Hoosain, R. (1992). Psychological reality of the word in Chinese. In *Advances in psychology* (Vol. 90, pp. 111-130). North-Holland.

Hsu, S.-H., & Huang, K.-C. (2000a). Effects of word spacing on reading Chinese text from a video display terminal. Perceptual & Motor Skills, 90, 81–92.

Hsu, S.-H., & Huang, K.-C. (2000b). Interword spacing in Chinese text layout. Perceptual & Motor Skills, 91, 355–365.

Inhoff, A. W., Liu, W., Wang, J., & Fu, D. J. (1997). Use of spatial information during the reading of Chinese text. In D. L. Peng, H. Shu, & H. C. Chen (Eds.), *Cognitive research on Chinese language* (pp. 296 –329). Jinan, China: Shan Dong Educational Publishing.

Inhoff, A. W., & Wu, C. (2005). Eye movements and the identification of spatially ambiguous words during Chinese sentence reading. *Memory & Cognition*, *33*(8), 1345-1356.

Jeffreys, H. (1961). *Theory of Probability*. Oxford, UK: Oxford University Press.

Liu, P. P., Li, W. J., Lin, N., & Li, X. S. (2013). Do Chinese readers follow the national standard rules for word segmentation during reading? *PloS one*, *8*(2).

Li, D. C. (2006). Chinese as a lingua franca in Greater China. *Annual Review of Applied Linguistics*, *26*, 149-176.

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

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940-967.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null hypothesis significance testing. *Behavior Research Methods, 43*(3), 679–690.

Perea, M., & Wang, X. (2017). Do alternating-color words facilitate reading aloud text in Chinese? Evidence with developing and adult readers. *Memory & Cognition, 45,* 1160–1170. https://doi. org/10.3758/s13421-017-0717-0.

Pollatsek, A., Rayner, K., & Henderson, J. M. (1990). Role of spatial location in integration of pictorial information across saccades. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 199-210.

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. Retrieved from URL http://www. R-project. org/., page R Foundation for Statistical Computing*.

Rayner, K., & Pollatsek, A. (1996). Reading unspaced text is not easy: Comments on the implications of Epelboim et al.'s (1994) study for models of eye movement control in reading. *Vision Research*, *36*(3), 461-465.

Sainio, M., Hyönä, J., Bingushi, K., & Bertram, R. (2007). The role of interword spacing in reading Japanese: An eye movement study. *Vision Research*, *47*(20), 2575-2584.

Wang, S., Huang, C. R., Yao, Y., & Chan, A. (2015). Create a Manual Chinese Word Segmentation Dataset Using Crowdsourcing Method. In *Proceedings of ACL-IJCNLP 2015* (pp. 7–14). Retrieved from https://aclweb.org/anthology/W/W15/W15-3102.pdf

Winskel, H., Radach, R., & Luksaneeyanawin, S. (2009). Eye movements when reading spaced and unspaced Thai and English: A comparison of Thai–English bilinguals and English monolinguals. *Journal of Memory and Language*, *61*(3), 339-351.

Yen, M. H., Radach, R., Tzeng, O. J., & Tsai, J. L. (2012). Usage of statistical cues for word boundary in reading Chinese sentences. *Reading and Writing: An Interdisciplinary Journal*, *25*(5), 1007-1029.

Zang, C., Liang, F., Bai, X., Yan, G., & Liversedge, S. P. (2013). Interword spacing and landing position effects during Chinese reading in children and adults. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 720-734.

Zang, C., Wang, Y., Bai, X., Yan, G., Drieghe, D., & Liversedge, S. P. (2016). The use of probabilistic lexicality cues for word segmentation in Chinese reading. *The Quarterly Journal of Experimental Psychology*, *69*(3), 548-560.

Zhou, W., Wang, A., Shu, H., Kliegl, R., & Yan, M. (2018). Word segmentation by alternating colors facilitates eye guidance in Chinese reading. *Memory & Cognition, 46,* 729–740. https://doi.org/10.3758/s13421-018-0797-5.

Chapter 5

Summary and conclusions

The main goal of this thesis was to contribute to the *what, how* and *when* questions about visual word recognition through the use of natural reading paradigms. Specifically, Chapter 2 made an endeavour to resolve a long-standing debate about the *when* question: (i) how early the well-established word properties, such as word frequency and semantic similarity, make a discernible impact on recognition performance of English derived words. Furthermore, by using the co-registration of EEG and eye-tracking technique that allows free eye movement behavior, Chapter 2 aimed at solving the paradox evident in previous literature where behavioral signatures of lexical effects during word recognition precede those in the brain. Chapters 3 and 4, in their turn, mainly concerned the *what* and *how* of visual word processing in Chinese, and again using eye-tracking or co-registration of EEG and eye-tracking methodologies investigated (ii) whether segmentation probabilities affect word segmentation and identification in Chinese; and (iii) whether statistics of spelling errors influence the recognition of correctly spelled words in Chinese.

Through a series of reading experiments, this thesis aimed at filling specific gaps in the previous literature on visual word recognition. Based on the findings obtained, it can be concluded that: first, there are many factors that influence word recognition performance, which are still not considered in the current models of word recognition. As Chapter 4 shows, some of the factors can come from the investigation of languages other than English, on which most word recognition models have been based. This chapter explains how studies in Chapters 3 and 4 enrich the existing knowledge on word recognition and points to certain word properties that influence word recognition during reading of connected texts that current recognition models do not consider. Second, we conclude that statistical word properties influencing word recognition are better explored utilizing experimental paradigms that emphasize natural reading techniques, such as eye-tracking or a co-registration of EEG & eye-tracking (Chapter 2). Although isolated word reading experiments have provided useful results to the field, an emerging literature utilizing natural reading of words in context has revealed that additional factors, contextual or behavioral, have an immediate impact on word reading and change the dynamics of word processing itself. To understand how readers process words naturally, the application of experimental techniques that do not interfere with the natural course of reading is necessary.

Below, this chapter elaborates on how research findings in this thesis contribute to the existing body of knowledge on visual word recognition. Additionally, this chapter outlines future directions and recommendations based on the results obtained from three of the studies in this dissertation.

1 Summary of findings and their broader significance

1.1 The when of frequency and semantic similarity effects in derived word recognition

The paradox is still present in the co-registered EEG and eye-tracking data. Previous research has put a tremendous effort into establishing the time-course of word recognition (Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006; Pylkkänen & Marantz; Grainger & Holcomb, 2009; Rastle, Davis, Marslen-Wilson, & Tyler, 2000; Sheridan, & Reichle, 2016). Due to their high temporal resolution, this line of research was mainly conducted with two classes of experimental paradigms: behavioral, such as eye-movement studies, and neurophysiological, such as EEG and MEG. However, upon comparison of the results obtained by both techniques separately, or by their combination (e.g., co-registration of EEG and eye-tracking), several studies pointed to an apparent paradoxical observation: lexical and semantic effects observed in the behavioral record tend to predate those in the brain activity registered using neuroimaging techniques (Schmidtke & Kuperman, 2019; Schmidtke, Matsuki, & Kuperman, 2017) or did not show up at all (e.g., Kretzschmar, Schlesewsky, & Staub, 2015; Degno, Loberg, Zang, Zhang, Donnelly, & Liversedge, 2019). This observation runs counter to the foundational assumption in the neuroscientific literature which assumes that human behavior takes its source from underlying neural activity (Just & Carpenter, 1976; Reichle, 2006). Failure to align the time stamps of the same cognitive processes across two methodologies will question the validity of the results obtained in the prior literature and bring doubt to the analytical procedures employed in the analysis of neurophysiological or behavioral data.

One potential reason for the contradictory results in earlier studies is the use of various experimental tasks, different sets of stimuli and different groups of participants. One way to address this problem is by combining two or more methodologies simultaneously to eliminate all the above-mentioned sources of variance. Chapter 2 of this thesis did exactly that by looking into the onsets of the well-established lexical effects in derived word recognition, whole word frequency and semantic similarity, by co-registering EEG and eye-tracking in two experimental tasks, sentence reading and lexical decision. Additionally, prior research on the time course of lexical effects mainly utilized analytic techniques that focused on the analysis of central tendency. Chapter 2 specifically sought after the methods that could point to the onset of an effect or when an effect starts to impact the response variable: Chapter 2 used mixed-effects quantile regression for the analysis of eye-tracking data and generalized additive mixed modelling for the EEG data analysis. Results showed that the earliest time point when the whole word frequency effect had a discernible impact on derived word recognition is evidenced in the eye-tracking data as early as 175 ms. The semantic similarity effect was well observed later, at 355 ms, again in the eye-tracking data, however, a weak significance already started to appear at 164-189 ms. The frequency effect was not revealed by EEG data in either of the two experiments, while semantic similarity was observed at 365 ms only in the lexical decision experiment. In sum, the earliest time point for the effects was found in the eye-tracking data in the sentence reading experiment when words were read in context. The main contribution of these findings is to highlight the paradox about the time-course of lexical effects in the word recognition research in a more methodologically rigorous way by eliminating potential confounds and sources of variance that earlier studies did not consider and applying analytical techniques beyond the analysis of central tendency.

Yet, even with the elimination of variance resulting from the separate use of behavioral and neurophysiological experimental techniques, the paradox still can be observed, where signatures of lexical effects seen in eye fixation durations are absent in the brain activity (or appear at the same time). In light of these findings, it is also possible that new methodological and analytical techniques may need to be tested or developed when analyzing the neurophysiological data in order to uncover the existence or the time-course of lexical effects in word recognition under natural reading conditions. In this regard, this study calls for the investigation of other dimensions of neurophysiological activity, which include but are not limited to time-frequency analysis, coherence measures between frequencies, and power spectrum (Hald, Bastiaansen, & Hagoort, 2006; Khader & Rösler, 2004; Weiss & Mueller, 2003).

Another potential reason for prior contradictory results could be related to the nature of experiment design and stimuli presentation. Previous neurophysiological research mainly used a rapid serial visual presentation (RSVP) technique in combination with a manual lexical decision task, aimed at eliminating eye movements and using long intervals between words to prevent the overlapping of brainwaves evoked by the successive presentation of words (Sereno, Rayner, & Posner, 1998). In natural reading paradigms, participants have freedom in moving their eyes whenever and wherever they need. Two problems arise in this regard. First, there is a problem of overlapping neural responses from preceding or subsequent events, as in natural reading readers take up the information from the words seen in fovea and parafovea simultaneously. Second, there are complex influences of task specific variables that cannot be eliminated or controlled, such as saccadic movements. Consequently, there are two possible outcomes that could result from these problems. First, it might be the overlapping neural responses from the surrounding stimuli or from the eye-movement activity imposed by the experimental task that may override certain effects, such as the absent whole word frequency effect in the neurophysiological data. It also might be that lexical effects are delayed because of several cognitive processes being involved at the same time. Analytical techniques for analyzing ERP/FRP data that could take into account the problem of overlapping components are needed for the exploration of these questions. There are recent developments in this direction, however, the application of this technique to the present data is out of the scope of the studies included in this thesis (Woldorff, M.G., 1993; Ehinger & Dimigen, 2019).

Taken together, results from Chapter 2 suggest that eye-tracking coupled with natural reading paradigms and distributional analysis techniques turned out to be an effective technique for studying the time-course of word recognition processes. In contrast, EEG is very effective in studying isolated word recognition processes. Present-day analytical techniques of EEG data need to be improved for the analysis of data obtained from natural reading paradigms. Additionally, through an investigation of the time-course of lexical effects, Chapter 2 demonstrates an example where results obtained from isolated word reading and sentence reading paradigms differ substantially. Specifically, the sentence reading experiment elicited visible differences between high and low frequency words, and the difference between words of high and low semantic similarity was observed earlier in that experiment. These findings advocate for the use of natural reading paradigms in exploring the cognitive processes underlying visual word recognition.

1.2 The what, how and when of spelling entropy

The frequent encountering of spelling errors leads to a difficulty in the processing of correctly spelled words in Chinese. Another statistical property that was recently demonstrated to influence recognition of words is the distribution of spelling errors for a certain word in the language. Recent studies by Rahmanian and Kuperman (2019) and Kuperman et al. (2021) found that the frequent encountering of spelling errors impedes the recognition of words in their correct orthographic form in a number of languages, such as English, Greek, Hebrew and Finnish. They hypothesized that alternative spelling variants of a word create their own orthographic representations and are stored in the mental lexicon along with the correct orthographic form. It was suggested that the underlying cause of the inhibitory effect observed in the above two studies is the competition for activation between spelling variants of a word. The orthographic competition, or in other words, the uncertainty in choosing spelling variants was measured with an information-theoretic measure of entropy. Subsequently, this phenomenon was coined as "the effect of spelling entropy", where a word with a high spelling entropy is a word in which available orthographic variants are of similar probabilities, and a word with low spelling entropy is a word which has only one dominant spelling variant. Consequently, upon their encounter, words with high spelling entropy had a high uncertainty when choosing between alternative spelling variants for activation. This was revealed by a fixation duration measure of a total reading time on a word. Interestingly, the effect was more profound on high frequency words.

However, can we evidence the same for Chinese, a language that has a distinct writing system when compared to alphabetical languages? In Chapter 3, we aimed to investigate this question. The results showed that this is the case: words with high spelling entropy had a slower reading time than words with low spelling entropy. Similarly, higher frequency words were more affected by entropy than lower frequency words. These findings suggest that the newly uncovered measure of spelling entropy has a universal effect that can be seen across many languages and across various writing systems.

The results obtained from this line of research on effects of spelling entropy are compatible within the framework of naïve discriminative learning (NDR) (Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011) and Lexical Quality Hypothesis (LQH) (Perfetti, 1985; 2007). According to NDR, learning is a discriminative process that minimizes the uncertainty between a set of cues and a predicted outcome with every encounter of a cue with the mapped outcome. In the case of the learning of spelling, cues in the NDR can be seen as spelling variants, and the outcome can be seen as the meaning of a word. Within this framework, every encounter of an incorrect spelling of a word (cue) leaves episodic memory traces of that spelling and diminishes an opportunity to strengthen the connection of the correct spelling with the meaning of that word, leading the reader to "unlearn" the correct orthographic form. For instance, the more you see the word accident written as *acident*, the weaker the connection of the correct spelling *accident* with the meaning of that word. As a result, these weaker connections result in weaker orthographic representations of a word.

According to LQH, a high-quality lexical representation of a word is comprised of strong connections between a word's orthography, phonology, and semantic representations. Under this hypothesis, spelling error is a reflection of a low-quality lexical representation. The existence of an additional spelling variant learned from frequent exposure creates a weak connection of the correct orthographic representation with phonology and semantics. Thus, NDR and LHQ both suggest that learning of an incorrect form of a word creates a weak orthographic representation that results in a spelling error. Consequently, this points to a reciprocal relationship between spelling variations and the quality of orthographic representations: spelling variations not only reflect weak orthographic representations but also cause them. Our findings along with the findings from Kuperman and Rahmanian (2019) and Kuperman et al. (2021) all point to this nature of relationship.

Orthographic competition takes place early in word recognition. One of the important pieces of information that is usually missing from models of word recognition is the time course of when various information types become available and are used to recognize a word. Time course is of particular importance because from a modelling perspective it is crucial to know if one or more processes occur simultaneously or sequentially one after another. Unfortunately, to date, there are no models that specify their temporal dimension in detail; a situation attributable to a lack of information or consensus as to when certain information becomes available during word recognition. Chapter 3 extended the results obtained from alphabetical languages to Chinese on the existence of the spelling entropy effect. Another contribution of this chapter is that by co-registering EEG and eye-tracking, it also found the neural correlates of this effect in the brain and shed light on the timing when spelling entropy influences word recognition.

Chapter 3 hypothesized that processing difficulty of high spelling entropy words results from a competition between spelling variants stored in the mental lexicon. According to the Interactive Activation Model (McClelland & Rumelhart, 1981), low-level visual and orthographic information flows bottom-up to a high-level lexical representation of words and flows back down to the low-level again. Upon reading a word, orthographic information obtained from visual input travels to the lexical representation level and tries to match a previously stored orthographic representation. However, if a word's lexical representation has several spelling variants or another (incorrect) spelling variant of similar probability to the correct spelling, upon matching, an uncertainty arises about which stored orthographic representation needs to be activated, thus resulting in a competition between orthographic representations. EEG results in Chapter 3 revealed that this competition has an early impact on word processing and is found in the 150 – 300 ms window, when as previously shown, orthographic processing takes place (Grainger, Kiyonaga, & Holcomb, 2006; Hauk et al., 2006; Mariol, Jacques, Schelstraete, & Rossion, 2008).

1.3 The what and how of Chinese word segmentation

Segmentation probabilities play an important role in Chinese word segmentation and identification. While most alphabetical languages have an effective cue for word segmentation in the form of spacing, logographic languages, such as Chinese, lack any visual cues to guide segmentation decisions. Subsequently, theories about Chinese word segmentation and identification have been proposed in the field. There have been three main hypotheses about how a continuous Chinese text is segmented into words. One hypothesis by Perfetti and Tan (1999) advocated a serial approach. Namely, it proposes that the default segmentation strategy of Chinese readers is to segment characters into two-character words as most words in Chinese are made of two characters. Perfetti and Tan (1999) found evidence that three-character targets ABC embedded in lexical garden-path sentences were read slower when A-B and B-C character pairs could form existing words as opposed to the control condition where B-C formed a word but A-B did not. They suggested that Chinese readers process characters and segment them into words sequentially and serially in the direction of reading, from left to right.

A competing proposal by Inhoff and Wu (2005) is a multiple activation hypothesis, which assumes that multiple written units are active in parallel in one's perceptual span and simultaneously available for analysis. Inhoff and Wu (2005) monitored eye-movements while participants read sentences with 4-character ambiguous strings as targets. Each sentence contained a critical area (a sequence of 4 characters, C1234), where C12 and C34 formed separate words (control condition), or an area where C12, C34 and C23 formed different words (ambiguous condition). Participants had longer reading times in the ambiguous condition. Inhoff and Wu (2005) proposed that the lexical form of C23 must have been active during critical area viewing. With an increased number of words in the perceptual span, the time required to make decisions about word segmentation increases as well.

Another study by Li, Rayner, and Cave (2009) argues for a mixture of parallel and serial processes in Chinese reading. They propose a model of Chinese word segmentation and identification that assumes the parallel processing of characters in the perceptual span combined with a serial recognition of words one at a time. In this model, the activation of characters takes place at the character level, where several character recognizers operate in parallel. Next, the information gathered at this level feeds forward and activates the mental representation at the word level. Compatible with the activated characters lexical information about this candidate word feeds back to the character level to validate if these characters are part of an activated word. With enough iterations, activation of the word reaches a threshold, so the word is identified and segmented simultaneously. Li and Pollatsek (2020) also assumed that characters and words in the perceptual span are both activated in parallel and word segmentation and identification is a unified process. Li et al.'s word segmentation and identification model could process 4-character strings only and reported simulations for a whole-report task. Li and Pollatsek made a step further and constructed a model that simulated reading of whole sentences, where perception span was designed to move across sentences. Moreover, in addition to the word processing module this model integrated an eye-movement control module that determines when and where to move the eyes.

Although evidence exists to support all, the serial, the parallel and the mixed-mechanism processing hypotheses, the field of reading does not yet have a definitive answer to how boundaries between these units are found. All the hypotheses outlined above lack the criteria upon which the segmentation decision is made. No matter how many characters or words are active in the perceptual span, and no matter if they are processed sequentially or in parallel, it is unclear how characters are combined to form a word. Chapter 4 of this thesis utilizes segmentation probabilities, which are segmentation judgments for word boundaries experimentally obtained from Chinese readers. One of the goals of Chapter 4 was to propose that segmentation probabilities could be a potential source of guidance for segmentation decisions. By measuring eye fixation durations with the help of eye-tracking, Chapter 4 has found that spaces only in highly probable word boundary positions, where a segmentation probability is considered high, are beneficial for processing. It was evidenced by shortened reading times on individual words and sentences, which points to a psychological reality of segmentation probabilities in Chinese minds and suggests that they serve as potential cues during the word segmentation process. These findings inform the current Chinese word segmentation models by providing a more detailed look at what could be the driving source for assigning a set of characters available in the perceptual span to words.

Spacing is not an effective segmentation cue for Chinese word segmentation. With spacing being the most prominent visual cue during word segmentation in alphabetical languages (Perea & Acha, 2009), a great amount of work has been done on whether spacing is as beneficial for word identification in unspaced languages, such as Thai and Chinese. This question was mainly investigated by artificially introducing spaces into natural unspaced text to see if it is advantageous for reading at the word and sentence levels if compared to their unspaced counterparts. In those studies, shorter reading times on words or sentences were associated with benefits in processing. Previous research on beneficial effects of spacing in Chinese has brought contradictory results. Despite the shorter fixation times on spaced words, spaced sentences took identical or longer times to read as unspaced sentences. Another goal of Chapter 4 was to further explore if spacing is in fact beneficial for processing by experimentally manipulating the prevalence of spacing based on segmentation probabilities. Two experiments were created, where in one experiment we inserted a space into natural unspaced text at every probable word boundary position, and in another experiment, spaces were inserted only at highly probable word boundaries. Results demonstrated that the beneficial effect of spacing is selective and contingent on the prevalence of spacing: shorter fixation times on words (although marginal) and sentences were only observed in the experiment where a space was located at high segmentation probability places.

Nevertheless, despite the advantage of spacing at the sentence level, the effect size it has on the word level is rather small. Additionally, considering the fact that Chinese readers often do not agree on where to put a word boundary in a continuous string of characters, there is no clear way that it could fit every reader's expectation. This makes a space not as efficient as it is in alphabetical languages.

Famous models of eye-movement control, such as SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005) and E-Z Reader (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Rayner, & Pollatsek, 2003) are based on alphabetical language reading, and all assume that inter-word spaces play a crucial role in word segmentation. However, in Chinese and in several other non-spaced languages, such as Japanese Kanji and Thai, there are no spaces to mark word boundaries. Remarkably, as it is the case in Chinese, spaces turned out not to be effective for word segmentation and reading, so it is very difficult to extend the existing models to the reading of unspaced Chinese. Although Rayner, Li and Pollatsek (2007) attempted to extend the E-Z Reader model to Chinese reading, the model completely lacked an explanation of how words in this language are segmented. As it is argued by Chapter 4, segmentation probabilities may be used as a source of information for word segmentation during reading in Chinese. It is self-evident that more research is needed for the exploration of the word segmentation issue in Chinese and other non-spaced languages. Also, more modelling work is required to make existing models universal in accommodating the characteristics of every language.

2 Limitations of the findings and future directions

2.1 The word frequency effect and confounding factors

To review, Chapter 2 explored the timing of whole word frequency and semantic similarity effects by co-registering EEG and eye-tracking techniques in two experiments, sentence reading and visual lexical decision. In both experiments, the whole word frequency effect was not observed as revealed by fixation-related potential analysis. Due to the abundant evidence for the word frequency effect that was revealed during isolated word reading in the lexical decision task in EEG studies, it was surprising not to find it in our lexical decision data. One possibility that caused null effects in the experiment is the existence of potential confounding factors, such as word length. Previous literature that investigated the word frequency effect found it already within 150 ms after stimulus onset (e.g., Hauk et al., 2006). However, some studies found that this effect can be dependent on word length. For instance, Assadollahi and Pulvermüller (2003) showed that low frequency words elicited stronger brain responses at 120-170 ms for short words only, and at 225-250 ms for long words exclusively. The stimuli for both experiments in Chapter 2 included derived words with word length ranging from 6 to 14 letters, however, it was not systematically controlled in the analysis of FRP data. It is possible that it was the confounding factor of word length that resulted in null effects in the lexical decision experiment in Chapter 2. Future studies should better control for potential confounds in their experiments when exploring the time-course of psycholinguistic factors that influence word recognition processes.

2.2 Spelling entropy and individual differences in Chinese

Most studies on word recognition performance focused on group-level data that based their assumptions on the average across participants. Consequently, computational models of visual word recognition were constructed in a way that rarely considered individual differences between poor and skilled readers (although there are exceptions, see Zevin & Seidenberg, 2006). Recently, there is more and more research showing that word processing efficiency during reading is modulated by reading skill (e.g., Yap, Balota, Sibley, & Ratcliff, 2015). The results from the study in Chapter 3 demonstrated that words that have a high spelling entropy value are difficult to recognize as demonstrated by increased reading times on these words. Hence, readers with poor reading and spelling skills should have greater difficulty in recognizing them during reading. The study in Chapter 3 asked participants to complete a spelling test that required them to identify spelling errors in words and sentences. When the effect of spelling entropy was modelled as a function of word reading time, the resulting score from this test was included in the model to account for individual differences in our participants' spelling skills. Spelling test scores did not influence reading latencies in our model, nor was there a significant interaction with entropy values.

One of the limitations of Chapter 3 is that it only included spelling test scores as a measure of spelling proficiency, however, there are also reading proficiency measures that potentially may interact with spelling entropy. One example is the Author Recognition Task, which measures the reader's exposure to print. It may be possible that people with a greater exposure to reading materials in the past are better and faster at recognizing words that are misspelled frequently in the language. Similarly, Rahmanian and Kuperman (2019) did not find an effect of spelling score, however, they did find a significant negative correlation of ART with word total reading time.

Also, the effect of spelling or reading proficiency was not investigated in the EEG analysis of Chapter 3. Does spelling or reading proficiency modulate the amplitude and latency of FRP components? Although the main effect of spelling score was not found in the eye-tracking record, it is very possible to observe this effect in the EEG data. We leave these questions for investigation in future research.

2.3 The time-course of word segmentation in Chinese

One of the limitations of Chapter 4 is that it did not explore the time course of the word segmentation process during Chinese reading. Although Chapter 4 proposed segmentation probabilities to be considered in the existing models of word recognition, it did not provide the timing of when the effect of segmentation probabilities occurs during recognition. A starting point could be to run an EEG and eye-tracking co-registration study to simultaneously explore the effects of segmentation probabilities and shed light on the timing of the segmentation process.

One of the interesting questions that could be explored using co-registration of EEG and eye-tracking is whether word segmentation processes are concurrent with word identification or whether one precedes the other. Previous models of Chinese word segmentation all assumed that word segmentation is achieved simultaneously when a word is identified (e.g., Li et al., 2009). For instance, Chinese reading model (CRM) by Li and Pollatsek (2020) hypothesized that all words that are supported by the characters activated in the perceptual span compete for activation, and once a word wins the competition, it is automatically identified and segmented at the same time. By finding an effect of segmentation probabilities on word reading performance, we could possibly find neural signatures of this effect and shed light on the timing when information about word segmentation starts to kick in. The results of this research would be the first step towards investigating the chicken and egg problem in Chinese word recognition: whether it is the word recognition that comes before word segmentation or vice versa.

We can pose the same question to alphabetical languages that use spacing information for guiding their word segmentation decisions. Does word segmentation precede word recognition? When is the spacing information visible in the parafovea during word recognition processed, and a word segmented and identified? Current models of eye movement control (e.g., E-Z Reader, SWIFT, Glenmore) all assume that inter-word spacing is critical for word segmentation, however, no model explains how word segmentation is achieved and when this process takes place during the timeline of word recognition. While there are numerous eye movement studies that explore the spacing effect on reading behavior, only one study explored neural signatures of this effect during sentence reading. Degno, Loberg, Zang, Zhang, Donnelly, and Liversedge (2019) compared word reading in two conditions: normally spaced and unspaced, where spaces between words were filled with random letters. They replicated the results of previous eye movement studies, where in the unspaced condition readers fixated words longer than in the normal spaced condition and confirmed that the unspaced condition disrupts eye guidance of the next eye movement and the identification of the currently fixated word. EEG results demonstrated that the spacing effect is present during very early stages of processing, at around 120-300 ms, during the activation of the orthographic representation of the fixated word. However, again, even if we assume that the time when spacing takes an effect is when word segmentation is achieved, does this indicate that this is the time when a word is identified? Clearly, more research is needed to find answers for this question.

3 Conclusions

There are many psycholinguistic properties of words that influence word recognition performance during reading. Many of them have been discovered already using isolated word reading paradigms that drastically differ from how we read naturally in everyday life. Moreover, most existent models of word recognition are heavily based on reading in English. This thesis aimed at filling specific gaps in the current understanding of word recognition processes by employing natural reading experimental paradigms and using English and Chinese as languages of investigation. Specifically, the study on English derived word recognition (Chapter 2) found the earliest time when word frequency and semantic similarity effects occur during word identification and pointed to the effectiveness of eye-tracking methodology coupled with distributional analysis techniques in finding the time-course of word recognition during natural reading. Research on Chinese misspelled word recognition (Chapter 3) shows that the 'spelling entropy' effect is existent in Mandarin Chinese and is available early during the orthographic processing stage. Research from Chinese word segmentation (Chapter 4) demonstrates that the "taken-for-granted" visual word segmentation process in alphabetical languages is not a trivial problem and requires extension of existing models of word recognition to unspaced languages. All in all, this thesis advocates for the use of natural reading paradigms and the investigation of languages other than English to obtain a multifaceted understanding of word reading processes.

References

Assadollahi, R., & Pulvermüller, F. (2003). Early influences of word length and frequency: a group study using MEG. *Neuroreport*, *14*(8), 1183-1187.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, *118*(3), 438.

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., & Liversedge, S. P. (2019). Parafoveal previews and lexical frequency in natural reading: Evidence from eye movements and fixation-related potentials. *Journal of Experimental Psychology: General*, *148*(3), 453.

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., & Liversedge, S. P. (2019). A co-registration investigation of inter-word spacing and parafoveal preview: Eye movements and fixation-related potentials. *PloS one*, *14*(12), e0225819.

Ehinger, B. V., & Dimigen, O. (2019). Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ*, *7*, e7838.

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

Grainger, J., Kiyonaga, K., & Holcomb, P. J. (2006). The time course of orthographic and phonological code activation. *Psychological science*, *17*(12), 1021-1026.

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Language and linguistics compass*, *3*(1), 128-156.

Hald, L. A., Bastiaansen, M. C., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain and language*, *96*(1), 90-105.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, *30*(4), 1383-1400.

Inhoff, A. W., & Wu, C. (2005). Eye movements and the identification of spatially ambiguous words during Chinese sentence reading. *Memory & cognition*, *33*(8), 1345-1356.

Just, M. A., & Carpenter, P. A. (1976). The role of eye-fixation research in cognitive psychology. *Behavior Research Methods & Instrumentation*, *8*(2), 139-143.

Khader, P., & Rösler, F. (2004). EEG power and coherence analysis of visually presented nouns and verbs reveals left frontal processing differences. *Neuroscience Letters*, *354*(2), 111-114.

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1648.

Kuperman, V., Bar-On, A., Bertram, R., Boshra, R., Deutsch, A., Kyröläinen, A.J., Mathiopoulou, B., Oralova, G., & Protopapas, A. (2021). Prevalence of spelling errors affects reading behavior across languages. *Journal of Experimental Psychology: General*.

Li, X., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive psychology*, *58*(4), 525-552.

Li, X., & Pollatsek, A. (2020). An integrated model of word processing and eye-movement control during Chinese reading. *Psychological Review*, *127*(6), 1139.

Mariol, M., Jacques, C., Schelstraete, M. A., & Rossion, B. (2008). The speed of orthographic processing during lexical decision: electrophysiological evidence for independent coding of letter identity and letter position in visual word recognition. *Journal of Cognitive Neuroscience*, *20*(7), 1283-1299.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, *88*(5), 375.

Perea, M., & Acha, J. (2009). Space information is important for reading. *Vision Research*, *49*(15), 1994-2000.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific studies of reading*, *11*(4), 357-383.

Perfetti, C. A., & Tan, L. H. (1999). The constituency model of Chinese word identification. In *Reading chinese script* (pp. 127-146). Psychology Press.

Perfetti, C. A. (1985). *Reading ability*. oxford university Press.

Pylkkänen, L., & Marantz, A. (2003). Tracking the time course of word recognition with

MEG. *Trends in cognitive sciences*, *7*(5), 187-189.

Rahmanian, S., & Kuperman, V. (2019). Spelling errors impede recognition of correctly spelled word forms. *Scientific Studies of Reading*, *23*(1), 24-36.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and cognitive processes*, *15*(4-5), 507-537.

Rayner, K., Li, X., & Pollatsek, A. (2007). Extending the E-Z reader model of eye movement control to Chinese readers. *Cognitive science*, *31*(6), 1021-1033.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, *26*(4), 445-476.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological review*, *105*(1), 125.

Reichle, E. D. (2006). Computational models of eye-movement control during reading: Theories of the" eye-mind" link.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1793.

Schmidtke, D., & Kuperman, V. (2019). A paradox of apparent brainless behavior: The time-course of compound word recognition. *Cortex*, *116*, 250-267.

Sheridan, H., & Reichle, E. D. (2016). An analysis of the time course of lexical processing during reading. *Cognitive science*, *40*(3), 522-553.

Woldorff, M. G. (1993). Distortion of ERP averages due to overlap from temporally adjacent ERPs: Analysis and correction. *Psychophysiology*, *30*(1), 98-119.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 597.

Zevin, J. D., & Seidenberg, M. S. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language*, *54*(2), 145-160.