

The use of eye movement corpora in vocabulary research

Marc Brysbaert¹ & Denis Drieghe²

¹ Department of Experimental Psychology, Ghent University, Belgium

² School of Psychology, University of Southampton, UK

Keywords: vocabulary, eye movements, corpus, megastudies

Ms submitted for the Special RMAL Issue: *Eye-tracking in vocabulary research*

Address: Marc Brysbaert
Department of Experimental Psychology
Ghent University
9000 Ghent
Belgium
marc.brysbaert@ugent.be

Abstract

Analysis of existing datasets of eye movements in reading is a valuable tool for vocabulary research because it allows researchers to examine word recognition in an authentic context. We argue that such secondary analysis is an important addition to new experimental studies and existing mega-studies because it examines word recognition in real text rather than in crammed conditions or in isolation. Corpora in which participants read long texts are particularly interesting because they provide rich material that can be better controlled for confounding variables, but a collection of small data sets can also be interesting because it contains more variation than is typically possible in a single study. We discuss the considerations to take into account when dealing with eye movement data in reading and urge colleagues to make their eye movement data available in the spirit of open science so that a larger database can be built more quickly.

Vocabulary and applied linguistics

Words are the building blocks of communication. This becomes evident when adults meet who do not speak a common language. Even if they have much information to share and show willingness to interact, communication is near impossible as long as there are no signs (words) with shared meanings. Sentence processing and discourse comprehension are also important for fluent interaction and follow their own rules, but are not as critical as word use. Communication is possible without grammar and inferences as long as there are shared words.

Further evidence for the importance of vocabulary knowledge for language comprehension is found in observations that vocabulary knowledge correlates more than .5 with text comprehension (Kuperman et al., 2023; Siegelman et al., 2023), and that text comprehension is severely compromised when listeners or readers do not know the meaning of 5% or even 2% of words (Laufer & Ravenhorst-Kalovski, 2010). Therefore, word acquisition and usage are important topics in applied linguistics.

New word learning is not limited to early language learners. A large-scale crowdsourcing study indicated that adult native speakers continue to learn an average of two new words per week (Brysbaert et al., 2016). Some of these words are new to describe a novel situation, but many words already existed and are seen/heard for the first time. In addition, even experienced language users encounter many word forms they have not encountered before (or very rarely). This is the case for low-frequency inflections, derivations and compound words. Such word forms are ubiquitous in languages with many inflections and derivations, or that write compound nouns as concatenated words. The proliferation of word forms is also observed in English, even though the language has a very simple inflectional system and writes new compounds as separate words. Even in English, there is no limit to the increase in the number of word types as the size of the corpus grows (Brysbaert et al., 2016).

In this paper, we consider how corpora of eye movement data can be used to investigate vocabulary acquisition and word processing in applied linguistics. First, we discuss the limitations of experiments in language research and discuss how secondary analysis of existing datasets provides a welcome addition (see also Angele et al., 2015). Then we describe what eye-movement corpora can contribute, how to use these data, and what limitations to consider. We also look at how you can collect your own corpus and contribute to the line of research.

The limits of experiments in language research

It is generally accepted in philosophy of science that experiments are the best way to test hypotheses. If one theory predicts an effect and the other does not, then setting up an experiment to decide between the two theories is a researcher's dream. As a result, applied linguistics increasingly refers to experimental evidence (e.g., McKinley & Rose, 2019). However, we must realize that two conditions must be met to set up a good experiment: (1) We are able to manipulate the independent variable, and (2) We are able to control all relevant confounding variables.

One topic for which the experimental approach works well is word priming. Suppose a researcher hypothesizes that letters of written words are converted into sounds and this conversion aids in visual word recognition. The priming effect then predicts that a target word like *floor* will be processed faster if it follows a homophonic stimulus (*flore*) than if it follows a non-homophonic stimulus with the same

number of shared letters (*floop*). Because the target words can be preceded by both related and unrelated primes, the researcher has complete control over the situation. Thus, if a phonological priming effect is observed, the researcher has good evidence that letter-sound conversions contribute to visual word processing (Perfetti & Bell, 1991).

Unfortunately, total control is often impossible in language research or has unintended consequences (Balota et al., 2013). For starters, words often cannot be manipulated at will. Suppose a researcher assumes that emotional words are easier to acquire than neutral words. Then they cannot freely assign words to conditions. Emotional value is an inherent feature of words and cannot be assigned ad libitum. All one can do is *select* words for the various conditions and try to *match* them on other relevant dimensions, such as word frequency, word length, word concreteness, and so on. One problem with the latter is that researchers currently have more than 240 variables to choose from (Gao et al., 2023). This requires many observations to separate the dimension of interest from alternative candidates.

Researchers sometimes try to get around the problem of word features by using non-words (e.g., Driver, 2022). These are legal strings of letters that have no meaning in the language in question, making it possible to arbitrarily assign an interpretation to them. For example, some participants are taught that "to briss" means to die and "to bart" means to throw, while others are taught the opposite. It can then be examined whether the pseudowords are learned more quickly when they refer to an emotional event than when they refer to a neutral event.

While this approach solves some problems, it illustrates a second reason why experiments in language research are often suboptimal. The non-words either refer to a new language the participant needs to learn, or mislead the participants who mistakenly thinks they are getting useful information about the English language. As anyone learning a second language (L2) knows, we are not equally motivated to learn all types of words. In particular, words that we perceive as useful to ourselves are learned quickly, sometimes after only one encounter (Borovski et al., 2010). In contrast, words for which we see no utility (e.g., names of different types of fish or grass) are not picked up or easily forgotten, even if we encounter them repeatedly. The use of non-words thus raises the question of external validity if the language is unfamiliar (to what extent does learning these non-words resemble learning words in a language that interests us?) and raises ethical questions if participants are led to believe that the non-words are useful words in a language they are learning.

A third reason why experiments sometimes yield limited information about everyday language processing is that many rare words or sentence constructions must be crammed into a single study session. Participants are asked to learn dozens of new words, or they have to process a particular, infrequent syntactic construction over and over again. Although such situations are not unknown in real life (e.g. L2 classes dedicated to a particular topic), they are far removed from natural language processing (Ferreira & Yang, 2019) and their contribution to language acquisition in real life can be questioned (e.g. Ambridge et al., 2006).

A fourth limitation of experiments is that they are limited in duration. This reduces the amount of information one can present to a single participant. As a result, researchers tend to restrict the stimuli to the extremes of a continuum. For example, when studying the effect of word frequency on word recognition, they will use a set of high-frequency words and a set of low-frequency words, assuming that the relationship between word frequency and processing efficiency is linear and that the findings with extreme words generalize to words in the middle. In addition, there is a danger of experimenter bias in

stimulus selection. Especially among the low-frequency words, there are words that are commonly known (toolbar, placement, irregularly) and words that are mostly unknown (emerald, pyknic, honewort), which means that an experimenter who believes in the word-frequency effect has the ability to choose more difficult low-frequency words than an experimenter who does not believe in the word-frequency effect (Forster, 2000; Kuperman, 2015).

Correlational studies complement experiments and have promoted mega-studies

In situations where complete manipulation of conditions is not possible or has undesirable effects, research on correlations between non-manipulated variables is a useful complement to experiments. This was already argued by Bacon (1620) who distinguished between experimental history and natural history. The former referred to active manipulation to see what effects it had, the latter to observation of what was there. Bacon saw experimental history as the superior way to extract secrets from nature, but he also wrote many pages on the need to collect and organize a multitude of observations under different conditions in order to arrive at an understanding by finding patterns in the data (today called correlations between variables).

Bacon's natural history approach also applies to language research. First, such research respects the fact that many word features cannot be manipulated and can only be selected. Second, it is easier to reduce the correlation between two related variables by selecting stimuli from the entire range than by trying to create a 2x2 table of orthogonal extreme values (e.g., Liu et al., 2022). Third, correlational studies can be applied to language processing in the wild and in life-relevant contexts. Fourth, correlational studies are more consistent with the continuous nature of language features and better respect the distribution of features. Finally, one can have access to much larger data sets than are feasible in laboratory experiments, if such data have been collected and archived for other purposes.

An interesting correlational study of word stimuli was published by Spieler and Balota (2000). They collected naming times for 2,820 English words in young and old adults and investigated the effects of word frequency, word length and similarity to other words. The results revealed that all three factors predicted significant amounts of variance in word-naming latencies for both groups. However, older adults showed a larger influence of word frequency and reduced influences of word length and similarity to words compared with younger adult.

Correlation studies are particularly informative if they cover a large part of the problem space. Otherwise, there is always the possibility that the observed patterns in the data are specific to the stimulus materials tested. Therefore, Balota et al. (2007) extended the stimulus set of Spieler and Balota (2000) and collected naming times and lexical decision times for more than 40 thousand English words in what they called a megastudy. Such a study allowed the effects of different word features to be examined in unprecedented detail and also allowed to check the quality of the measures available for each feature (e.g. the various word frequency norms that have been published), as reviewed by Gao et al. (2023). Consequently, megastudies were run in several languages (see <http://crr.ugent.be/programs-data/megastudy-data-available> for a curated list).

Since the publication of the Balota et al. (2007) megastudy, hundreds of researchers have reanalysed the dataset to address a plethora of new questions, not only related to word processing, but also to the processes involved in binary decision making. The size of the study is large enough to do all types of

correlational analyses, along the natural history approach proposed by Bacon (1620). The use of existing data to answer questions unrelated to the original work, is called *secondary analysis*. It works particularly well with large datasets, because the patterns tend to be robust and can consider the effects of possible confounding variables (also called the big data approach). Data can be gathered not only in the first language (L1), but also in L2 (e.g., Brysbaert et al., 2021), so that word processing in the L2 can be compared to word processing in the L1 (see also Siegelman et al., 2023).

A limitation of existing language megastudies is that they consist of single word processing, indeed mostly restricted to lexical decision. Participants are presented with sequences of letters (or sounds) and have to decide whether these form an existing word or not. Other studies investigate word naming (Tse et al., 2023) or semantic decision (Pexman et al., 2017). All studies are limited to word processing in isolation, however, rather than word processing in discourse or in text. This is where eye tracking research offers an interesting addition.

The contribution of eye movement corpora

The most natural way to study visual language processing is to track eye movements during reading. Eye movements provide a detailed picture of information intake. They have been used extensively in experimental research (Liversedge et al., 2011). In such studies researchers typically construct new stimulus materials and run a dedicated experiment to answer one or two detailed questions.

We will use an experiment of Elgort et al. (2018) to illustrate the approach and to indicate how secondary analysis can complement the findings. Elgort et al. (2018) wanted to map the acquisition of new words during reading. They asked 40 English L2 speakers to read an English text of 12 thousand words. The text introduced 14 words that were unfamiliar to the participants. Each new word was presented multiple times (up to 40 times), as is typical for newly introduced topics in text. Elgort et al. (2018) were interested in how long readers would look at the target words. They hypothesized that participants' eyes would linger quite long on a word when it was first seen, and that processing would become increasingly efficient as the word was read multiple times, indicating that some aspects of the word had been stored in memory and could be retrieved. This is indeed what the authors observed: Processing times for the new words were longest at the first encounter and decreased systematically for the next 10-12 occurrences. After that, the times flattened out to the processing times observed for low-frequency words in text (we will come back to the findings below in Figure 2).

The findings of Elgort et al. (2018) were interesting (see also Drieghe & Chan Seem, 2022; Kamienkowski et al., 2018), but required six months of solid work. Stimuli had to be created, 40 participants tested, and access to an eye tracker secured (plus the expertise to conduct the experiment). Even then, the outcome was only information about 40 participants who read 14 target words in one particular text. Ideally, this study should be repeated a few times with variations to establish the replicability and generality of the finding. This can be done by conducting new experiments or by analysing data sets already collected by other researchers for other purposes. Suppose another researcher had collected eye movements from participants reading a novel or an expository text to compare performance of different types of readers. Very likely, new words were introduced into that text (e.g., the names of the characters in the novel or a new concept explained in the expository text), allowing for secondary analysis to see whether the pattern observed by Elgort et al. is replicated. Or alternatively, if the data collected by Elgort et al. are

made available to other researchers, they may be able to answer other questions about vocabulary processing, using Bacon's natural history approach.

Although data from any text reading study are potentially useful for secondary analysis (as long as the data are made available), in practice secondary analysis is most informative when the dataset is rich (just as we saw for the megastudy approach). Large collections of eye movement data are usually called eye movement corpora rather than eye movement megastudies (in line with the name used for other data repositories in language research).

Pynte and Kennedy (2006) published one of the first eye movement corpora, which they called the Dundee corpus. It consisted of British and French young adults reading newspaper articles presented on a computer screen, five lines at a time. This provided data for more than 52 thousand word tokens per language (about 10 thousand word types). The corpus allowed the authors to investigate topics such as the influence of word frequency and word length on different eye movement parameters (see below), the degree to which the upcoming (parafoveal) word is processed, and differences between French and English. Other researchers then used the data to test further hypotheses they had.

Insert Table 1 here

Table 1 contains a selection of eye movement corpora that are large enough to be of interest for secondary analysis. A complete, updated list can be found at <http://crr.ugent.be/programs-data/megastudy-data-available>. An example of a widely used database is the Ghent Corpus (GECO; Cop et al., 2017). It contains the eye movements of English- and Dutch-speaking participants reading an entire novel (55 thousand word tokens). The Dutch participants not only read part of the novel in L1, but also part in English as L2, allowing interesting language comparisons. Recently, Chinese-English bilinguals were added to the participant sample, who read the same novel in L1 and L2 (Sui et al., 2023).

Coskun et al. (2023) carried out a secondary analysis of the GECO corpus to investigate the extent to which inflected word forms prime each other: Are we faster at reading the word *tree* if we read the word *trees* several minutes earlier? Such long-lag morphological priming has been observed in lexical decision and Coskun et al. (2023) wondered if it would also be observed in normal reading. They listed all nouns, verbs and adjectives that occurred twice in the GECO and analysed whether reading times were faster in the second encounter than in the first. Repetition priming was observed for identical words that were repeated (*tree* after *tree* or *trees* after *trees*), but not for morphologically related words (*trees* after *tree* or *tree* after *trees*). Needless to say, this finding, if repeated, will have a huge impact on our theories of lexical organisation and on the best way to conceive of representations in the mental lexicon (word forms, lemmas, flemmas, or word families?).

Currently, eye movement corpora are still limited in number and size, compared to the megastudies available for various languages. The hope is that this will soon improve (see below for some recommendations), especially when it becomes possible to include non-invasive eye cameras in e-readers, enabling eye tracking under the most natural and comfortable reading conditions.

Secondary analysis of existing eye movement corpora serves two purposes. The first goal is to understand what guides eye movements during reading. What variables determine what and how long we look at during reading? Are eye movements entirely controlled by ongoing language processing? How

much information do we extract from parts of the text that we do not look at directly? The second goal is to understand language processing itself. What processes are involved when we read words, assemble them into sentences and integrate them into a coherent message? What stimulus variables are important? How can we understand differences between participants? Differences between languages? Between L1 and L2 reading? The second goal is probably more interesting for applied linguists, but both are closely related.

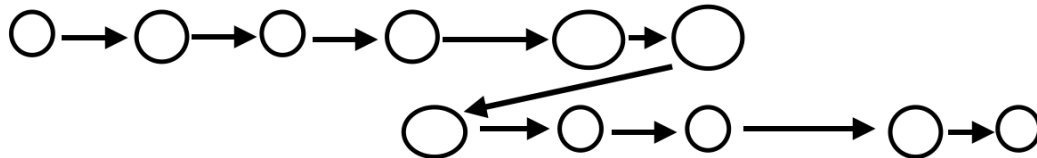
As we saw in the example of Coskun et al. (2023), analysis of eye-movement corpora is likely to provide new data of interest, because eye movements tap into processes not shared by isolated word processing tasks. Kuperman et al. (2013) correlated reading times for words (gaze durations; see below) with lexical decision times for the same words. When the reading times were based on a limited set of words presented in short neutral sentences that were read carefully, the correlation between gaze duration and lexical decision times was $r = .72$ in one study (based on $N = 47$ words) and $r = .44$ in another study ($N = 80$ words). These are quite good correlations, as can be expected if both variables measure the same underlying processes. However, when lexical decision times were compared with gaze times from the Dundee corpus (Pynte & Kennedy, 2006), the correlation was only $r = .24$ ($N = 6817$ words), meaning that less than 6% of the variance in processing times was shared. Dirix et al. (2019) replicated the latter finding when they calculated the correlation between lexical decision times obtained in mega-studies and the GECO word reading times ($r = .29$; $N = 2982$). Thus, reading words in connected text taps into partly different processes than deciding whether a letter sequence constitutes a known word or not, which makes it interesting to explore the differences and commonalities of the tasks.

Eye movement variables

When studying eye movement corpora, it is important to know which variables to examine. As Figure 1 shows, eye movements during reading consist of short periods of standstill (fixations), followed by rapid movement to a new part of the text (saccade). During fixations, language information is extracted from the text. Saccades are necessary because we can only extract language information from a limited window (roughly the currently fixated word and the next word), so we have to move our eyes when we want to retrieve new information (Rayner, 1975; Simons & Levin, 1997). Usually, saccades take us to the next word(s) in the text. Occasionally, however, we make a backward saccade (regression) if we have not fully processed the information so far.

Figure 1: Eye movements while reading. The circles represent fixations, the arrows saccades. Most saccades are forward. Occasionally, however, a backward saccade is made (regression), usually indicating text integration problems that require reanalyzing what was read before.

The man who hunts ducks out on weekends.



Several variables have been described to summarise eye movements when reading. These include:

- Fixation duration: (average) fixation time, measured in milliseconds.
- First fixation duration: duration of the first fixation on a word conditional on the word being fixated.
- Single fixation duration: duration of fixation on a word conditional on whether the word is fixated exactly once.
- Refixation probability: percentage of words that are fixated more than once on the first encounter.
- Regression probability: probability that eyes are directed to a previously read part of the text.
- Gaze duration: sum of fixations on a word before the word is exited (by a forward saccade or by a regression); conditional on whether the word is fixated.
- Go past time: time between when a word is first fixated and when the eyes leave the word to the right.
- Total reading time: sum of fixations on a word, independent of whether the word is first encountered or after regressions to it.
- Forward saccade length: length of the saccade to the next word(s).
- Skipping rate: percentage of words not looked at during the first read.
- Landing position: place in the word where the eyes land.

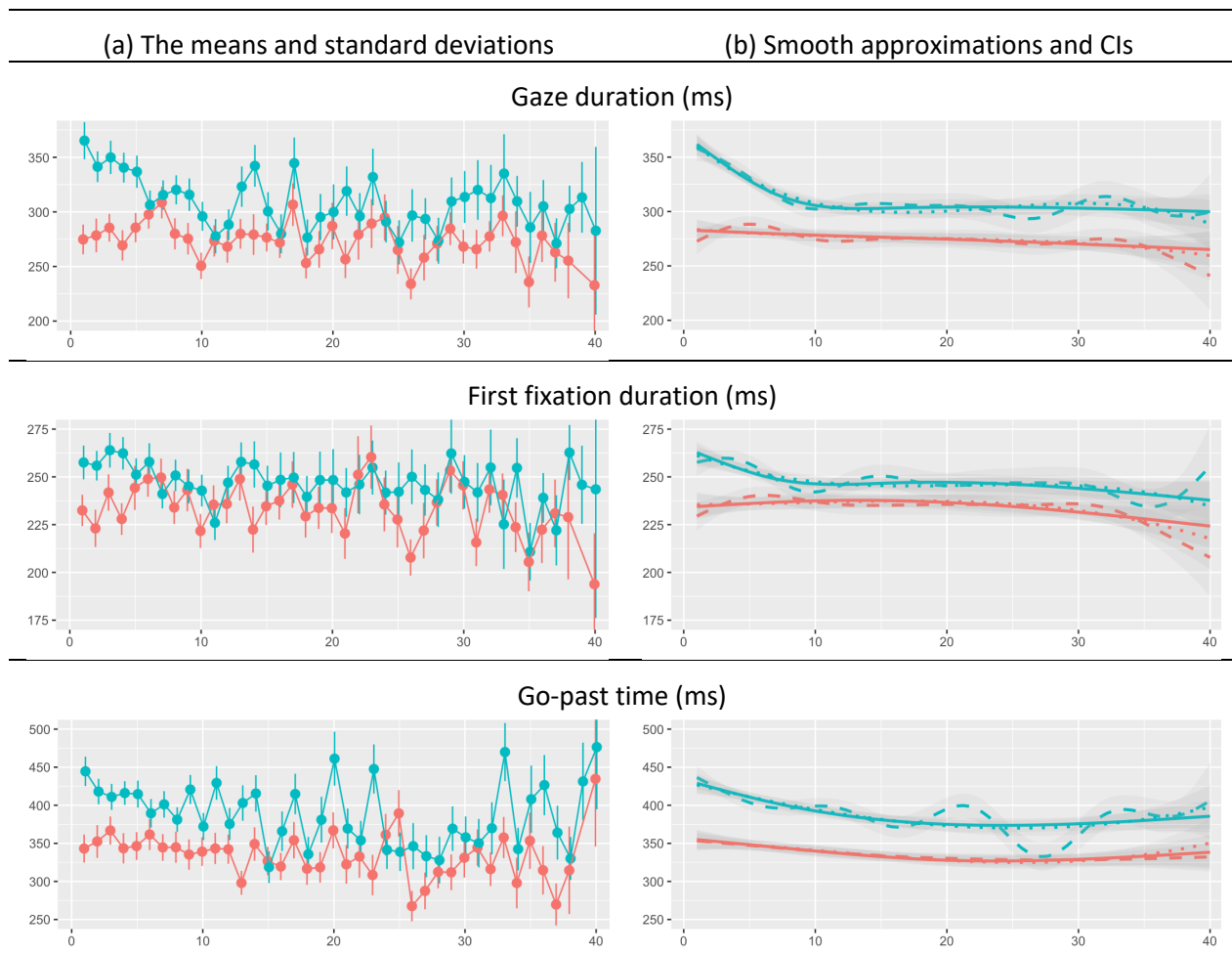
Fixation duration and gaze duration are part of the so-called when-decisions in eye-movement control: how long do the eyes stay on a particular part of the text before being directed elsewhere. Gaze duration is generally considered the most interesting variable for measuring word processing. It indicates how long eyes stay on the word during initial reading (usually on the order of 200-500 ms) and takes into account that words often require more than one fixation to be processed. For example, in Elgort et al.'s (2018) word acquisition study discussed earlier, gaze duration on target words was the primary dependent variable to measure how much attention was paid to the target words over the course of learning (Figure 2). It was a clearly defined, tangible measure to examine the processes involved in anchoring new words in the lexicon and the reduction in recognition effort required as words were learned through repeated exposure.

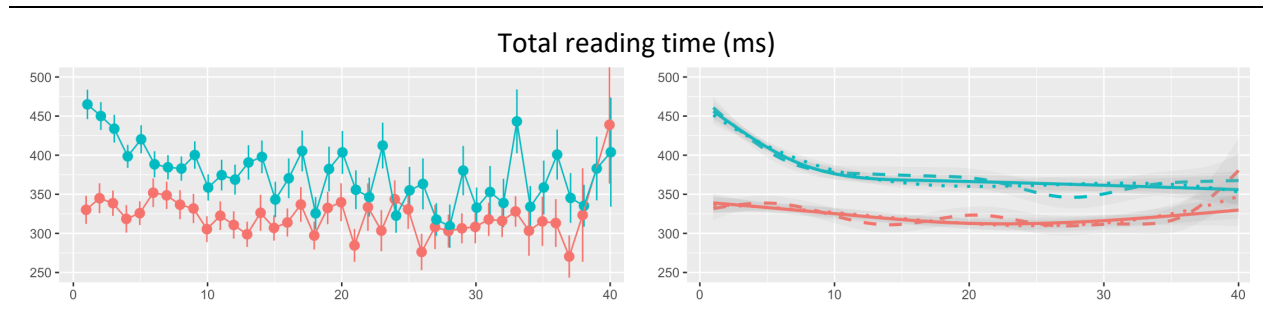
Initial fixation time is also a potentially interesting variable because it can give an idea of how quickly readers notice the difficulty of a word. On first landing, do they already realize that the word is going to

require extra effort? Elgort et al. (2018) observed a small effect on initial fixation duration, but less clearly than on gaze duration, as shown in Figure 2.

Figure 2: Eye movement findings from Elgort et al. (2018). Each row shows a different when-variable for up to 40 repetitions of target words (gaze time, first fixation time, go-past time, total reading time). The left panel shows the observed data; the right panel shows a smoothed approximation illustrating the main differences. In each figure, the upper (green) curve shows the data for the new target words and the lower (red) curve shows the data for a set of high-frequency control words. The data show the longer reading times when participants encounter an unfamiliar target word. The longer reading times were particularly pronounced in the first 10-12 readings of the word and then levelled off to a plateau above that observed for high-frequency words (but in line with that observed for other low-frequency content words in the text). The pattern was clearest for gaze duration and total reading time.

Source: Adapted from Elgort et al. (2018)





The reason why initial fixation duration is often less informative for word processing is that a significant percentage of initial fixations are mislandings due to an error in the execution of the eye movement. Such mislocated fixations are followed by a rapid corrective saccade (more details are given below) and do not say much about language processing. To avoid cases of non-language-related first fixation durations, authors sometimes limit their analysis to the single fixation duration. This is the duration of the first fixation given that the word is fixed only once in the first passage. The idea is that single fixations to a lesser extent include fixations that are not related to language processing.

Other when-variables are go-past time and total reading time. These variables include regressions to previous words (go-past time) or from further words back to the target word (total reading time). Regressions indicate integration problems for words within the sentence and the broader text. In their study, Elgort et al. (2018) found that new words had a stronger effect on total reading time than on go-past time (Figure 2), pointing to integration costs of the new words.

In addition to the when-variables, eye tracking also provides a number of where-variables. These indicate where the eyes land in a text and then go to. The where-variables include: Refixation probability, regression probability, saccade length, skipping rate, and landing position. As we will see below, they are often less informative for text processing than the when-variables. In the study of Elgort et al. (2018), no additional information was obtained from the where-variables.

The multitude of variables reflects the richness of eye-tracking data. At the same time, it increases the risk of researchers performing analyses until a statistically significant effect is found, which is then presented as an important new finding but which has little chance of being replicated in a new study. Therefore, it is good practice to look for convergent information across variables and be mindful of making multiple comparisons when reporting and evaluating findings (von der Malsburg, & Angele, 2017).

As a rule of thumb, you can expect that as reading becomes more difficult (hard text, less proficient reader, demanding comprehension questions), fixation duration and gaze duration increase, forward saccades become shorter, regression probability and time increase, and fewer words are skipped (see Rayner, 2009). No new information is obtained if a study is designed to examine only these effects, because they are well documented and accepted.

The eye-mind hypothesis in reading and its limitations

When researchers study eye movements to examine the cognitive processes underlying language comprehension, they rely on the eye-mind hypothesis (Just & Carpenter, 1980). According to this hypothesis, the eyes are a faithful window to the mind: We assume that word processing begins from the moment the eyes land on the word and continues until the eyes leave the word. In this view, gaze duration is the ecologically most valid measure of word processing (Just & Carpenter, 1980).

As a general principle, the eye-mind hypothesis can be expected to apply. However, it is embedded in a number of other processes and in noise, which means that individual gaze duration is only to some extent informative for ongoing text processing. As we saw above, there is a fairly low correlation between gaze duration and lexical decision time (Dirix et al., 2019; Kuperman et al., 2013). In addition, Dirix et al. (2019) found that the correlation between gaze duration calculated from two different eye movement corpora is also quite low. They reported a correlation of only $r = .18$ between gaze durations calculated from the Dundee corpus and gaze durations calculated from the GECO corpus (based on 1954 shared words).

One reason for the low correlation between gaze duration in the Dundee corpus and the GECO corpus is likely that participants in the Dundee study read several newspaper articles, while participants in the Geco study read an Agatha Christie detective novel, which contained many repeated words and multi-word expressions (Snell & Theeuwes, 2020). Indeed, the average reading rate in the GECO study (360 words per minute in L1) is significantly higher than the reading rate observed in most other studies (260 words per minute for fiction; Brysbaert, 2019).

Another reason why gaze durations have low correlations between studies is that gaze duration is a variable with noise in it. Evidence for this was reported by Dirix et al. (2019), who noted that the correlation between gaze durations in the Dundee and GECO corpus increased when gaze durations were based on multiple occurrences of the words in each corpus. That is, to get a good estimate of gaze duration for a given word, it is necessary to have several different sentences with this word in the corpus. This filters out sentence-specific effects and also reduces the measurement error in individual gaze durations.

In the following sections, we will provide more information about why gaze durations differ between studies and as a function of the text in which a word is presented. It is important to understand these processes so that appropriate conclusions are drawn. To anticipate our findings, it is not good practice to draw conclusions about the absence of correlations when there is little signal in the data (as measured by the reliability of the variables that are correlated). Similarly, it is dangerous to draw strong conclusions from the observation of a significant correlation based on a few cases within a corpus because the risk of uncontrolled variables is then very high. We hope these reports will not discourage researchers from performing secondary analyses on eye-movement corpora, but will make them more aware of the possibilities and pitfalls.

Parafoveal preview benefit

Let us begin by asking whether word processing begins only when the eyes land on the word or whether the word is already partially processed when the eyes are still on the previous word. Such processing is

called parafoveal processing. In vision, a distinction is made between foveal vision and parafoveal vision. Foveal vision refers to a circle of about 2 degrees (8 letters in normal reading) around our line of sight. In this part we see the most detail because our eyes contain many more receptors there than in other parts (parafovea and periphery). The difference in detail that can be seen is why we focus our eyes on something we want to examine and why eye movements during reading consist of a sequence of fixations that follow the text. Parafoveal vision refers to a ring in the visual field around the foveal part. So the question is whether word processing begins only when the word is in foveal vision (i.e., is being fixated) or can also be partially processed when the eyes are still focused on the previous word.

Parafoveal processing can be examined with the boundary change paradigm (Rayner, 1975). In this paradigm, there is an invisible boundary in a sentence before a critical target word. The sentence is initially presented with the target word masked or replaced by another stimulus (e.g. a row of XXXXXX). As long as the eyes do not land on the critical word, the target remains unavailable. Once the eyes cross the invisible boundary before the word, the parafoveal preview is replaced by the target word so that participants see the intended word in foveal vision when their eyes land on it. The contribution of parafoveal information can be measured by comparing this condition with a condition in which the target word was visible all the time.

A robust finding in the eye-movement literature is that fixation times on target words are shorter when the target word was visible in parafoveal vision as opposed to seeing a preview that was partially or completely different (Rayner, 1975). This finding is called the *parafoveal preview benefit*. It shows that word processing already begins before the eyes land on the word.

Explaining the benefit of parafoveal preview has led to what is perhaps the most controversial issue in research on eye movements during reading: Whether lexical processing in reading is serial or whether multiple words are processed in parallel (Starr & Rayner, 2001; for a review, see Rayner, 2009). In the highly influential E-Z Reader model (Reichle et al., 1998), only one word is lexically processed at a time. In this serial model, the completion of lexical identification of the foveal word causes the attention beam to shift to the next word in parafoveal vision. The arrival of the attentional beam initiates lexical processing of the parafoveal word. The arrival of the beam usually occurs before the eyes land on the word, and in this way the model explains the advantage of the parafoveal preview. The preview benefit is due to the processing time available between when the attention beam lands on the foveal word and when the eyes land on the word.

Other models such as the SWIFT model (Engbert et al., 2005) or the OB1 model (Snell et al., 2018) assume that several words are processed in parallel because attention is distributed across multiple words as a gradient. According to these parallel models, fixation times on a word reflect the processing of preceding and upcoming words in addition to the processing of the currently fixated word. Such models predict not only a parafoveal preview benefit, but also an influence of the upcoming word on the fixation times of the currently fixated word. That is, gaze duration on a word is influenced not only by how much it was processed in parafoveal vision during the previous fixation, but also by which word is coming next. Whether such so-called parafoveal-on-foveal effects exist is controversial, however (for a review see Drieghe, 2011; Jensen et al., 2021).

Still other models assume serial processing of lexical units, but argue that the units may consist of more than one word, for example, idioms such as “a storm in a teacup” or spaced compound words such as “teddy bear” (Zang, 2019; see also Cutter et al., 2014; Kliegl, 2007; Yang et al., 2022). In such models, the

multiword units are processed sequentially but the words within a multiword expression are processed in parallel, even when they are separated by spaces.

Independent of the interpretation of the parafoveal preview benefit effect, the existence of such an effect shows that gaze duration on a word is influenced not only by the time it takes to process that word in foveal vision, but also by how much of the word was processed in the previous fixation(s) and - possibly - by which word comes next. This is one of the reasons why gaze durations for words in sentences correlate less with the processing time of isolated words, and why gaze durations on words may differ as a function of the sentence in which the words occur.

Spill-over

Another finding in eye-movement data is that the gaze duration on a word is longer when the word is preceded by a low-frequency word than by a high-frequency word (e.g. Findelsberger et al., 2019; Kennedy et al., 2013; Kennison & Clifton, 1995; Rayner & Duffy, 1986).

An influence of the processing of the preceding word naturally arises from the architecture of a parallel model of eye-movement control. For example, the SWIFT model assumes that saccades are autonomously generated during reading and often occur before the fixed word has been fully processed. A mechanism exists to inhibit saccades when a reader experiences difficulties during word recognition (e.g. when encountering a low-frequency word), but this mechanism does not always prevent premature eye movements. Serial models also postulated mechanisms to account for spill-over effects. In EZ-Reader, for example, a low-frequency word causes the attention beam to shift to the next word later than for a high-frequency word, thus reducing the time window in which parafoveal processing can take place.

Again, important for the present discussion is that even assuming a strictly serial model of lexical processing, it is widely accepted that the duration of gaze on foveal words is influenced by features of the previously seen word. This problem can be circumvented in experimental research by ensuring that the prior word is equivalent between conditions. The best way to account for this factor in secondary analysis is to use a corpus rich enough so that target words are seen in different contexts that are similar on average (which can be controlled by variables related to the prior word).

Word predictability

A third factor influencing the gaze duration of a word, besides the processing difficulty of the word itself, is the predictability of the word within the sentence or text. In a sentence like “The baker rushed the wedding ____ to the reception”, the word “cake” is strongly expected from the preceding context. Predictability has traditionally been assessed via a cloze task in which participants not participating in the eye movement experiments are presented with the sentence up to the critical word and asked to complete the sentence with the first word that comes to mind (Block, & Baldwin, 2010; Federmeier et al., 2007; Taylor, 1953). A word produced frequently is considered predictable. A highly predictable word such as “cake” will be viewed for less time and skipped more often than the word “pies”, even though

the latter word is not considered semantically anomalous (e.g. Balota et al., 1985; Dimigen et al., 2011; Luke & Christianson, 2016).

MacDonald and Shillcock (2003) argued that word predictability need not be based on sentence- or discourse-based expectations (as suggested by the cloze task). The probability of the transition from one word to another already influences the ease with which a word is read. In English texts, the noun "confusion" is more likely to follow the verb "avoid" than the noun "discovery," and readers use such statistical patterns to process words with higher transition probabilities more quickly.

Frisson et al. (2005) challenged MacDonald and Shillcock's (2003) interpretation, but recent work with large language models has confirmed the importance of word-based statistics (Cevoli et al., 2022; Chandra et al., 2023; Heilbron et al., 2023). Large language models are connectionist networks trained on huge databases that pick up statistical regularities in the word sequences of the language. Commonly used models are GPT and BERT. These models allow the researcher to distinguish between two measures: Uncertainty before a target is encountered and surprise when a target is encountered. The first measure indicates the extent to which the target word is expected based on prior context. The second measure indicates how difficult it is to integrate the target word into the ongoing discourse representation. An unpredictable word is not only unexpected but also more difficult to integrate into the ongoing understanding of the text. Research suggests that both factors influence gaze duration on words (Cevoli et al., 2022; but also see Frisson et al., 2017).

Syntactic complexity

Word predictability refers mainly to the meaning of words and their context. However, sentences also have syntactic structure. Gaze duration on words can thus be affected by the syntactic difficulty of the sentence read. This is especially true for sentences that are temporally ambiguous and constructed in a different way than the reader expects. The latter type of sentences is called garden path sentences. When readers process a sentence like "Since Jay always jogs a mile seems a short distance", fixation times are longer due to processing difficulties unrelated to lexical processing (Frazier & Rayner, 1982; see also Figure 1). Although it is beyond the scope of this article to provide a comprehensive overview of such influences, overviews are available summarising how syntactic, pragmatic and world knowledge factors influence eye movements during reading (e.g. Clifton, Staub & Rayner, 2007).

Word length

Finally, it is important to keep in mind that word length is much more important in eye tracking studies than in any other paradigm. During reading, gaze duration and probability of word skipping are strongly influenced by word length (Brysbaert et al., 2005; Dirix et al., 2019; Siegelman et al., 2022). In isolated word recognition tasks such as lexical decision or naming, word length is also important, but not to the same extent as in eye movement studies. New et al. (2006) investigated the word length effect using the lexical decision times from Balota et al. (2007). They found no word length effect for words of 5-8 letters, facilitative effects for words of 3-5 letters and inhibitory effects for words of 8-13 letters, resulting in a U-shaped curve of the influence of word length on lexical decision times. In contrast, during reading, word

length effects are quite robust, with fixation times getting longer with each letter increase from at least 5 letters (Kliegl et al. 2004; Rayner et al, 1996).

The effects of word length and word frequency are often difficult to disentangle because the two variables correlate with each other. For example, the correlation between word length and word frequency in lexical decision time for 62 thousand English words studied by Brysbaert et al. (2019) was $r = -.44$.

Word length is important not only for target words, but also for the words preceding a target word. Differences in the length of preceding words can lead to different gaze durations on the target word (think of the spill-over effect) and can result in different rates of skipping target words. Therefore, it is good to always consider the length of at least the preceding word.

Where decisions versus when decisions

In the previous sections we discussed why Just and Carpenter's (1980) eye-mind hypothesis is only a rough approximation of how eye movements can be used to estimate processing difficulty, especially at the level of individual words. We have argued that parafoveal preview, spill-over, predictability, syntactic processing, and word length affect the gaze duration on a given word, in addition to the processing effort for the word itself. We have discussed these effects so that researchers can take them into account when considering whether a particular question can be answered by secondary analysis of an eye-movement corpus. As a rule of thumb, it is dangerous to investigate a phenomenon for which there are few stimuli, because these effects can easily be stimulus-specific (this applies to both observational and experimental research). Just as it is good to have many participants, it is good to have a wide variety of stimuli because this averages out the effects of confounding variables. Brysbaert and Stevens (2018) recommend aiming for at least 40 participants and 40 different stimuli per condition.

In addition to the when-variables, eye tracking provides a number of where-variables. These indicate where in the text the eyes land. As we saw, the where-variables include the probability of refixation, the probability of regression, the length of the saccade, the rate of skipping, and the landing position. Usually, less information about word processing is obtained from the where-variables than from the when-variables. This is because the where-variables are strongly influenced by word length and the fact that eyes do not always land where they were intended. Thus, skipping rate and landing position depend strongly on word length and distance of the word from the current fixation location. Word difficulty and predictability have only modest effects (Brysbaert et al., 2005; Heilbron et al., 2023; Rayner et al., 2011). Readers of languages with short words skip words more often than readers of languages with long words (Siegelman et al., 2022).

Where-variables are more informative when words have spatially distributed information, such as long compound words (afterimage, headache, sandstone, washing machine) or formulaic language (best of both worlds, thank you, give it up, you know what I mean). There is good evidence that familiar multi-word units are processed differently than other phrases (Carrol & Conklin, 2014; Siyanova-Chanturia, & Pellicer-Sanchez, 2019; Schmidtke et al., 2018). Because these expressions are read in a sequence of fixations, a combination of when and where measures is likely to be most informative here.

Where-variables are also more informative for short words of 5 letters or less, because these words are skipped very often. As a result, the numbers of fixations on these words tend to be low. In addition, fixations on short words have a high chance of being unintended fixations. When landing positions are plotted as a function of word length, they invariably show a Gaussian distribution (McConkie et al., 1988), indicating that the eyes do not always land on the intended location, but are subject to execution error. Nuthmann and colleagues (2005) made elegant estimates of how often short words are mistakenly skipped or fixated based on analyses of landing place distributions. Their analyses suggest that a significant proportion of fixations on very short words are actually the result of failed skips and are followed by a quick forward saccade. Therefore, the duration of these fixations says little about the language processing going on. A further finding with short words is that function words are skipped more often than content words of the same length (Chamberland et al., 2013; Drieghe et al., 2008), suggesting that they may be a different type of words (Greenberg et al., 2004) or that it is difficult to fully match function and content words on all possible confounding variables (Schmauder et al., 2000).

The error in the execution of eye movements is also why landing positions are difficult to interpret unless you can make a distribution over many observations. When many landing positions are analysed, they invariably form a normal distribution. The longer the saccade, the larger the standard deviation of landing positions (McConkie et al., 1988, 1994). The middle of a word or somewhat before the centre is the ideal landing position. If the eyes land far away from it, they are likely to make a rapid corrective saccade (Vitu et al., 2001). As we saw above, this is why initial fixation duration usually correlates less with language processing effort than gaze duration or single fixation duration.

Finally, where-decisions have a strong impact at the end and beginning of a text line, as there is a large return sweep to be made from one line to the next. These are likely to affect the gaze duration of the first and last word of a text line and therefore it is common to remove these words from eye movement analyses unless one is interested in the return sweep itself (Parker et al., 2023; Slattery & Vasilev, 2019).

Considerations for analysing eye movement corpora

One way to summarise what we have covered so far is to think about what factors to take into account when considering a secondary analysis of an eye movement corpus. What type of analysis is likely to be worthwhile?

- Eye movement data are noisy and therefore many observations are needed to obtain stable estimates of word processing times. Averaging reading times across multiple presentations increases correlations (Dirix et al., 2019) and therefore it is recommended to ensure that target words are available in multiple contexts. Variation in context gives more control over the impact of the previous word (spill-over, regression) and the next word (parafoveal preview).
- Researchers increasingly use large language models to estimate the predictability of words in text (Cevoli et al., 2022; Chandra et al., 2023; Heilbron et al., 2023). This seems to work rather well although further investigations involving direct comparisons with other measures of predictability are welcome.
- Because decisions where to move the eyes depend on the length of words in a sentence, it is important to make sure that you can control the length of target words and take into account the length of at least one word before the target word.

- Sometimes a word is skipped incorrectly (either due to a planning or execution error). This is evidenced by the fact that the skipped saccade is followed by a quick regression to the skipped word. These (rather common) cases create difficult choices for gaze duration calculation.
- Because return sweeps are long eye movements, critical words are ideally not the first or last words of lines of text.
- Unless researchers are interested in function words, it is better not to include these words in the analysis, as their large number threatens to obscure patterns present only for content words. For a stop list of function words, see, for example, Nothman et al. (2018).
- Eye movement variables are not interchangeable. Indicate which variable is the most interesting for your research question and make detailed, testable predictions for this particular variable. Adjust the significance level for subsequent exploratory analyses for the fact that you are making multiple comparisons (Von der Malsburg & Angele, 2017).

A call to build an expanding eye movement database by providing open access to materials and data

For the reasons discussed in the previous sections, the most interesting eye movement data for secondary analysis are large corpora, in which participants read thousands of words (see Table 1 for some of these studies). While such large studies are much needed and researchers are encouraged to collect more of them, there is an additional, probably more feasible way to assemble a database of eye movement data for secondary analysis. Several eye movement studies are published every week in which participants read sentences or paragraphs, with data for a few hundred words. These can be combined to gradually build up a large database of diverse material. Given the ease with which datasets can be stored and put online these days (e.g., through the open science framework, github, figshare, a university repository, or a research council repository), such an undertaking is quite possible. A collection of small data sets is interesting for secondary analysis because it includes more variation than is typically possible in a single study (Strand & Brown, 2023). With a little discipline, we can build a nice database in a few years, with additional data being added every few days (see Aydoğın et al., 2023, for an example of such an endeavour in another field). Ideally, the field agrees on a uniform format so that new information can be easily integrated and analysed. An important consideration in this regard is to ensure that you can provide access to the text you used. Sometimes authors discover at the very end of a study that they may violate copyrights if the text material is published as part of a corpus (e.g., because a published book was used).

Another way to increase the database is to also add data from self-paced reading studies. Eye movements provide the highest degree of detail, but a lot of (converging) information can be gathered from reading times of words, phrases, sentences, and paragraphs. These times can be obtained with self-paced reading. In this technique, participants have to press a key to see the information piece by piece, and the time between key presses is measured (Marsden et al., 2018; Mitchell, 1984; Patterson & Nicklin, 2023). The reading times can then be compared to eye movement data. Frank et al. (2013; see also Amenta et al., 2023) did so for a study they conducted and reported a correlation of $\rho = .21$ between reading times in word-by-word self-paced reading study and gaze durations for the same words in an eye tracking study (see below for the interpretation of this correlation). Vasilev et al. (2022) reported a reading rate of 165 words per minute ($SD = 50$) in participants reading short non-fiction paragraphs

word-by-word.¹ This is slower than the typical reading rate for non-fiction texts (240 words per minute; Brysbaert, 2019), but still as fast as listening to speech (Brysbaert & Vantieghem, 2023).

Corpora of self-paced reading times have been published by Frank et al. (2013), Smith and Levy (2013), Futrell et al. (2017), and Huang et al. (2023). A further interesting dataset may be Boyce and Levy (2023), who proposed the maze task (in which participants repeatedly need to choose between two options on how to continue a sentence) as an alternative to self-paced reading.

The need for reliability information

Because secondary analysis of existing materials is typically a correlational analysis, it is important to have information about the reliability of the variables. Above we saw that Frank et al. (2013) found a correlation of $\rho = .21$ between self-paced word reading times and gaze durations for the same words.

To properly interpret this correlation, it is important to know how reliable each variable was. If each variable was measured with a reliability of .9, then the observed correlation of .21 seems quite low and may indicate that different processes are involved in reading plain text and word-for-word reading. On the other hand, if each variable was only measured with a reliability of .3 (because of the large variability in both variables), then a correlation of .21 looks pretty good. You can get an estimate of how good by correcting the observed correlation via the equation: corrected correlation = observed correlation divided by the square root of the test reliabilities. Thus, if test reliability is .3, the expected correlation for perfectly reliable tests is $.21 / \sqrt{.3 \times .3} = .70$, suggesting that both variables measure largely the same processes. In contrast, if the reliability of the Frank et al. variables was .9, then the maximum expected correlation would be merely $.21 / \sqrt{.9 \times .9} = .23$, implying that the two variables have little in common. Unfortunately, Frank et al. (2013) did not give the reliability of their variables.

Reliability can be calculated when the raw data are available, but it is good to always provide this information when making a dataset available. There are many ways to estimate the reliability of variables (Flora, 2020; Revelle & Condon, 2019), with slightly different results, but the differences are usually small and the different indices rarely contradict each other. The main reason for low reliability is a limited number of observations (a small sample of participants reading a small number of stimuli). This is another reason why it is interesting to combine data from many studies. Each study will produce an error-prone estimate, but the combination will produce a better (more stable) estimate.

Conclusion

Secondary analysis of eye movement corpora has great potential for vocabulary research because eye movements during reading are the closest thing to visual language processing in natural conditions. To exploit this potential, two conditions must be met. First, the research community must have access to sufficient material so that stable (replicable) patterns can be observed. Second, researchers must know how to analyse these data so that they correctly interpret the presence (and absence!) of patterns in the data. We hope the present article will help achieve these ambitions.

¹ Reading rate was not included in the article, but was kindly provided by the authors upon request.

References

- Ambridge, B., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, 21(2), 174-193.
- Amenta, S., Hasenäcker, J., Crepaldi, D., & Marelli, M. (2023). Prediction at the intersection of sentence context and word form: Evidence from eye-movements and self-paced reading. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-022-02223-9>
- Angele, B., Schotter, E. R., Slattery, T. J., Tenenbaum, T. L., Bicknell, K., & Rayner, K. (2015). Do successor effects in reading reflect lexical parafoveal processing? Evidence from corpus-based and experimental eye movement data. *Journal of Memory and Language*, 79, 76-96.
- Aydoğan, T., Karşilar, H., Duyan, Y. A., Akdoğan, B., Baccarani, A., Brochard, R., ... & Balci, F. (2023). The timing database: An open-access, live repository for interval timing studies. *Behavior Research Methods*. Advanced Online publication <https://doi.org/10.3758/s13428-022-02050-9>.
- Bacon, F. (1620). *Novum Organum. Oxonii: E Typographeo Clarendoniano*.
- Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The Interaction of Contextual Constraints and Parafoveal Visual Information in Reading. *Cognitive Psychology*, 17, 364 – 390.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445-459.
- Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2013). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual Word Recognition Volume 1: Models and methods, orthography and phonology* (pp. 90-115). Psychology Press.
- Berzak, Y., Nakamura, C., Smith, A., Weng, E., Katz, B., Flynn, S., & Levy, R. (2022). CELER: A 365-participant corpus of eye movements in L1 and L2 English reading. *Open Mind*, 6, 41-50.
- Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, 42(3), 665-670.
- Borovsky, A., Kutas, M., & Elman, J. (2010). Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition*, 116(2), 289-296.
- Boyce, V., & Levy, R. (2023). A-maze of natural stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, 2(1): X, pp. 1–34. DOI: <https://doi.org/10.5070/G6011190>
- Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109, 104047.
- Brysbaert, M., Drieghe, D., & Vitu, F. (2005). Word skipping: Implications for theories of eye movement control in reading. In G. Underwood (Ed.), *Cognitive processes in eye guidance* (pp. 53-77). Oxford University Press.
- Brysbaert, M., Keuleers, E., & Mandera, P. (2021). Which words do English non-native speakers know? New supranational levels based on yes/no decision. *Second Language Research*, 37(2), 207-231.

- Brysbaert, M., Mander, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2), 467-479.
- Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, 1: 9, 1–20, DOI: <https://doi.org/10.5334/joc.10>.
- Brysbaert, M., Stevens, M., Mander, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology* 7, 1116. doi: 10.3389/fpsyg.2016.01116
- Brysbaert, M., & Vantighem, A. (2023). No correlation between articulation speed and silent reading rate when adults read short texts. *Psychologica Belgica*, 63, 82–91.
- Carrol, G., & Conklin, K. (2014). Eye-tracking multi-word units: some methodological questions. *Journal of Eye Movement Research*, 7(5), 1-11. DOI 10.16910/jemr.7.5.5
- Cevoli, B., Watkins, C., & Rastle, K. (2022). Prediction as a basis for skilled reading: insights from modern language models. *Royal Society Open Science*, 9(6), 211837.
- Chamberland, C., Saint-Aubin, J., & Légère, M. A. (2013). The impact of text repetition on content and function words during reading: Further evidence from eye movements. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(2), 94-99.
- Chandra, J., Witzig, N., & Laubrock, J. (2023, May). Synthetic predictabilities from large language models explain reading eye movements. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications* (pp. 1-7).
- Clifton, C., Jr., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–371). Elsevier. <https://doi.org/10.1016/B978-008044980-7/50017-3>
- Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2), 602–615.
- Coskun, M., Kuperman, V., & Rueckl, J. (2023). Long-lag repetition priming: No evidence for morphological effects. *The Mental Lexicon*. Advanced online publication at <https://doi.org/10.1075/ml.21014.cos>.
- Cutter, M. G., Drieghe, D., & Liversedge, S. P. (2014). Preview benefit in English spaced compounds. *Journal of experimental psychology. Learning, memory, and cognition*, 40(6), 1778–1786. <https://doi.org/10.1037/xlm0000013>
- 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), 552-572.
- Dirix, N., Brysbaert, M., & Duyck, W. (2019). How well do word recognition measures correlate? Effects of language context and repeated presentations. *Behavior Research Methods*, 51, 2800-2819.
- Drieghe, D. (2011). Parafoveal-on-foveal effects on eye movements during reading. In S.P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford Handbook on Eye Movements* (pp. 839-855). Oxford University Press.

- Drieghe, D., & Chan Seem, R. (2022). Parafoveal processing of repeated words during reading. *Psychonomic Bulletin & Review*, 29(4), 1451-1460.
- Drieghe, D., Pollatsek, A., Staub, A., & Rayner, K. (2008). The word grouping hypothesis and eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1552–1560.
- Driver, M. (2022). Emotion-laden texts and words: The influence of emotion on vocabulary learning for heritage and foreign language learners. *Studies in Second Language Acquisition*, 44(4), 1071-1094.
- Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 40(2), 341-366.
- Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, 112(4), 777-813.
- Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, 1146, 75-84.
- Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes*, 56(7), 485-495.
- Findelsberger, E., Hutzler, F., & Hawelka, S. (2019). Spill the load: Mixed evidence for a foveal load effect, reliable evidence for a spillover effect in eye-movement control during reading. *Attention, Perception, & Psychophysics*, 81, 1442-1453.
- Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484-501.
- Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28(7), 1109-1115.
- Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4), 1182-1190.
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2), 178–210. [https://doi.org/10.1016/0010-0285\(82\)90008-1](https://doi.org/10.1016/0010-0285(82)90008-1)
- Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, 95, 200-214.
- Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of Contextual Predictability and Transitional Probability on Eye Movements During Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 862–877.
- Futrell, R., Gibson, E., Tily, H., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2017). The natural stories corpus. arXiv preprint arXiv:1708.05763.

- Gao, C., Shinkareva, S. V., & Desai, R. H. (2023). Scope: The South Carolina psycholinguistic metabase. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-01934-0>
- Greenberg, S. N., Healy, A. F., Koriat, A., & Kreiner, H. (2004). The GO model: A reconsideration of the role of structural units in guiding and organizing text on line. *Psychonomic Bulletin & Review*, 11(3), 428-433.
- Heilbron, M., van Haren, J., Hagoort, P., & de Lange, F. P. (2023). Lexical processing strongly affects reading times but not skipping during natural reading. *Open Mind* 2023, 7, 757–783. doi: https://doi.org/10.1162/opmi_a_00099
- Hollenstein, N., Barrett, M., & Björnsdóttir, M. (2022). The Copenhagen Corpus of eye tracking recordings from natural reading of Danish texts. arXiv preprint arXiv:2204.13311.
- Hollenstein, N., Rotsztein, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data*, 5, 180291.
- Huang, K., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2023, April 21). Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark. <https://doi.org/10.31234/osf.io/z38u6>
- Jensen, O., Pan, Y., Frisson, S., & Wang, L. (2021). An oscillatory pipelining mechanism supporting previewing during visual exploration and reading. *Trends in Cognitive Sciences*, 25(12), 1033-1044.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354.
- Kamienkowski, J. E., Carbajal, M. J., Bianchi, B., Sigman, M., & Shalom, D. E. (2018). Cumulative repetition effects across multiple readings of a word: Evidence from eye movements. *Discourse Processes*, 55(3), 256-271.
- Kennedy, A., Pynte, J., Murray, W. S., & Paul, S. A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, 66(3), 601-618.
- Kennison, S. M., & Clifton, C. (1995). Determinants of parafoveal preview benefit in high and low working memory capacity readers: implications for eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), 68-81.
- Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, and Reichle (2007). *Journal of Experimental Psychology: General*, 136(3), 530.
- 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. <https://doi.org/10.1080/09541440340000213>
- Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *Quarterly Journal of Experimental Psychology*.

- 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, 563-580.
- Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., ... & Usal, K. A. (2022). Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, 1-35. doi:10.1017/S0272263121000954
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language*, 22(1), 15-30.
- Liu, X., Wisniewski, D., Vermeulen, L., Palenciano, A. F., Liu, W., & Brysbaert, M. (2022). The representations of Chinese characters: Evidence from sublexical components. *Journal of Neuroscience*, 42(1), 135-144.
- Liversedge, S., Gilchrist, I., & Everling, S. (Eds.). (2011). *The Oxford handbook of eye movements*. Oxford University Press.
- Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, 88, 22-60.
- Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50, 826-833.
- Mak, M., & Willems, R. M. (2019). Mental simulation during literary reading: Individual differences revealed with eye-tracking. *Language, Cognition and Neuroscience*, 34(4), 511-535.
- Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, 39(5), 861-904.
- McConkie, G. W., Kerr, P. W., & Dyre, B. P. (1994). What are 'normal' eye movements during reading: Toward a mathematical description. In J. Ygge and G. Lennerstrand (Eds.), *Eye movements in reading* (pp. 315 – 327). Oxford, England: Elsevier Science.
- McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations on words. *Vision Research*, 28(10), 1107-1118.
- McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, 43(16), 1735-1751.
- McKinley, J., & Rose, H. (Eds.). (2019). *The Routledge handbook of research methods in applied linguistics*. Routledge.
- Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In D.E. Kieras, & M.A. Just (Eds.), *New Methods in Reading Comprehension Research* (pp. 69–89). Erlbaum.
- New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13(1), 45–52. <https://doi.org/10.3758/BF03193811>

- Nuthmann, A., Engbert, R., & Kliegl, R. (2005). Mislocated fixations during reading and the inverted optimal viewing position effect. *Vision Research*, 45 (17), 2201 – 2217.
- Nothman, J., Qin, H., & Yurchak, R. (2018, July). Stop word lists in free open-source software packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)* (pp. 7-12).
- Pan, J., Yan, M., Richter, E. M., Shu, H., & Kliegl, R. (2022). The Beijing Sentence Corpus: A Chinese sentence corpus with eye movement data and predictability norms. *Behavior Research Methods*, 54, 1989-2000.
- Parker, A. J., Räsänen, M., & Slattery, T. J. (2023). What is the optimal position of low-frequency words across line boundaries? An eye movement investigation. *Applied Cognitive Psychology*. DOI: 10.1002/acp.4036
- Patterson, A. S., & Nicklin, C. (2023). L2 self-paced reading data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, 2(1), 100045.
- Perfetti, C. A., & Bell, L. (1991). Phonemic activation during the first 40 ms of word identification: Evidence from backward masking and priming. *Journal of Memory and language*, 30(4), 473-485.
- Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, 49(2), 407-417.
- Pynte, J., & Kennedy, A. (2006). An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. *Vision Research*, 46(22), 3786-3801.
- Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology*, 7(1), 65-81. [https://doi.org/https://doi.org/10.1016/0010-0285\(75\)90005-5](https://doi.org/https://doi.org/10.1016/0010-0285(75)90005-5)
- Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8), 1457-1506. <https://doi.org/10.1080/17470210902816461>
- Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3), 191-201. <https://doi.org/10.3758/bf03197692>
- 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. <https://doi.org/10.1037/0096-1523.22.5.1188>
- Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2), 514.
- Reichle, E.D., Pollatsek, A., Fisher, D.L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105, 125-157.
- Revelle, W., & Condon, D. M. (2019). Reliability from α to ω : A tutorial. *Psychological Assessment*, 31(12), 1395-1411.

- Schmauder, R, Morris, R. K., & Poynor, D.V. (2000). Lexical processing and text integration of function and content words: Evidence from priming and eye fixations. *Memory & Cognition*, 28 (7), 1098 – 1108.
- 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), 421-439.
- Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H. D., Alexeeva, S., Amenta, S., ... & Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54, 2843-2863.
- Siegelman, N., Elgort, I., Brysbaert, M., Agrawal, N., Amenta, S., Arsenijević Mijalković, J., ... & Kuperman, V. (2023). Rethinking First Language–Second Language Similarities and Differences in English Proficiency: Insights From the ENGLISH Reading Online (ENRO) Project. *Language Learning*.
- Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, 1(7), 261-267.
- Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (Eds.). (2019). *Understanding Formulaic Language: A Second Language Acquisition Perspective*. Routledge.
- Snell, J., & Theeuwes, J. (2020). A story about statistical learning in a story: Regularities impact eye movements during book reading. *Journal of Memory and Language*, 113, 104127.
- Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological review*, 125(6), 969-984.
- Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging*, 15(2), 225–231. <https://doi.org/10.1037/0882-7974.15.2.225>
- Starr, M. S., & Rayner, K. (2001). Eye movements during reading: Some current controversies. *Trends in Cognitive Sciences*, 5(4), 156–163
- Sui, L., Dirix, N., Woumans, E., & Duyck, W. (2023). GECO-CN: Ghent Eye-tracking COrpus of sentence reading for Chinese-English bilinguals. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-01931-3>
- Slattery, T. J., & Vasilev, M. R. (2019). An eye-movement exploration into return-sweep targeting during reading. *Attention, Perception, & Psychophysics*, 81(5), 1197-1203.
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302-319.
- Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging*, 15(2), 225-231.
- Strand, J. F., & Brown, V. A. (2023). Spread the word: enhancing replicability of speech research through stimulus sharing. *Journal of Speech, Language, and Hearing Research*, 66(6), 1967-1976.
- Sui, L., Dirix, N., Woumans, E., & Duyck, W. (2023). GECO-CN: Ghent Eye-tracking COrpus of sentence reading for Chinese-English bilinguals. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-01931-3>

- Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.
- Tse, C. S., Chan, Y. L., Yap, M. J., & Tsang, H. C. (2023). The Chinese Lexicon Project II: A megastudy of speeded naming performance for 25,000+ traditional Chinese two-character words. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-02022-z>
- Vasilev, M. R., Hitching, L., & Tyrrell, S. (2022, August 29). What makes background music distracting? Investigating the role of song lyrics using self-paced reading. <https://doi.org/10.31234/osf.io/nmdt3>
- 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.
- Von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, 94, 119-133.
- Yang, J., Van den Bosch, A., & Frank, S. L. (2022). Unsupervised text segmentation predicts eye fixations during reading. *Frontiers in Artificial Intelligence*, 5, 11.
- Zang C. (2019). New Perspectives on Serialism and Parallelism in Oculomotor Control During Reading: The Multi-Constituent Unit Hypothesis. *Vision*, 3(4), 50
- Zhang, G., Yao, P., Ma, G., Wang, J., Zhou, J., Huang, L., ... & Li, X. (2022). The database of eye-movement measures on words in Chinese reading. *Scientific Data*, 9, 411.

Table 1: Selection of eye movement corpora available for analysis (for a full list and links to the data, see <http://crr.ugent.be/programs-data/megastudy-data-available>).

Study	Stimuli	Word tokens	N participants	Language
Pan et al. (2021)	Sentences	1,685	60	Chinese
Sui et al. (2023)	Book	59,409	30	Chinese
Zhang et al. (2022)	Sentences	8,551	1718	Chinese
Hollenstein et al. (2022)	Speeches	34,987	22	Danish
Cop et al. (2017)	Book	59,716	19	Dutch
Mak & Willems (2019)	Stories	7,790	102	Dutch
Siegelman et al. (2022)	Texts	2,231	45	Dutch
Berzak et al. (2022)	Sentences	320,360	69	English
Berzak et al. (2022)	Sentences	320,360	296	English L2
Cop et al. (2017)	Book	54,364	14	English
Cop et al. (2017)	Book	54,364	19	English L2
Hollenstein et al. (2018)	Sentences	21,629	12	English
Kuperman et al. (2023)	Expository texts	1,653	543	English L2
Luke & Christianson (2018)	Paragraphs	2,689	84	English
Pynte & Kennedy (2006)	Newspaper articles	52,000	10	English
Siegelman et al. (2022)	Expository texts	2,109	46	English
Sui et al. (2023)	Book	59409	30	English L2