How well do word recognition measures correlate? Effects of language context and repeated presentations.

Nicolas Dirix[1*], Marc Brysbaert[1], and Wouter Duyck[1]

[1]Department of Experimental Psychology, Ghent University, Ghent, Belgium

[*]Corresponding author, Nicolas Dirix, Henri Dunantlaan 2, 9000 Ghent, Belgium,

nicolas.dirix@ugent.be

Abstract

The present study assessed to what extent different word recognition time measures converge, using large databases of lexical decision times and eye tracking measures. We observed a low proportion of shared variance between these measures, which limits the validity of lexical decision times for real-life reading. We further investigated and compared the role of word frequency and length, two important predictors of word processing latencies in these paradigms, and found that these influenced the measures to a different extent.

A second analysis of two different eye tracking corpora compared eye tracking reading times of short paragraphs with reading of an entire book. Our results reveal that correlation between eye tracking reading times of identical words in two different corpora are also low, suggesting that the higher-order language context in which words are presented plays a crucial role. Finally, our findings indicate that lexical decision times better resemble the average processing time of multiple presentations of the same word, across different language contexts.

Keywords: Visual word recognition; eye-tracking; big data studies

.

In the domain of psycholinguistic research, and more specifically in the study of visual word recognition, two of the most applied paradigms are the lexical decision task and eye tracking. In a lexical decision task, participants have to decide whether strings of letters are valid words or not. The time needed to make this decision and produce a yes/no response – the reaction time (RT) – can then be used to investigate influences of word characteristics on these data, such as the frequency or length of the words. The widespread use of the lexical decision task is not surprising, as it is fairly easy to implement, a lot of data can be collected in a relatively short period of time, and analysis and interpretation of the dependent variables (RT and accuracy) are straightforward (Keuleers & Brysbaert, 2011).

In eye tracking research, the eye movements of participants are monitored while they read single sentences or paragraphs. Reading longer passages of text is often referred to as "natural reading" because embedding words in discourse contexts increases correspondence to daily reading situations, where we read for meaning, rather than for lexicality (lexical decision). Eye tracking results in more than one dependent variable. The most commonly investigated measures are first fixation durations (the durations of the first fixation on a word), single fixation durations (the durations of the fixation on a word that is fixated only once), gaze durations (the sum of the durations of fixations on a word before the eyes leave the word) and total reading times (the summed fixation durations of all fixations on a word). These measures are assumed to reflect different stages in the word recognition process (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Rayner, 1998). For example, the first fixation duration is an indicator of the speed of the initial lexical access and word identification. Total reading time reflects higher order processing, such as verification and semantic activation of the word's meaning. As in lexical decision, the influence of word characteristics can be studied by looking for differences between words in all these reading times. As different measures represent different stages in the word recognition process, a detailed pattern of the influence of word characteristics can be revealed. Further advantages are the high spatial and temporal resolution of the equipment (modern eye trackers can record at a sampling rate of up to 2000 Hz with an average accuracy of 0.25 – 0.5 degrees of visual angle) and the ecological validity of the technique, as minimal instructions are required: participants simply have to read the sentences or text presented to them.

Both these tasks have a long history of application in reading research and they were applied to study similar topics in the field. For some of the more well-established effects, similar results were obtained across paradigms: high frequency words are processed faster than low frequency words (e.g., Rubenstein, Garfield, and Millikan (1970) for lexical decision, Rayner and Duffy (1986) for eye tracking), long words take more time to process than short words (e.g., Hudson and Bergman (1985) for lexical decision (but see New, Ferrand, Pallier, & Brysbaert, 2006), Vitu, O'Regan, and Mittau (1990) for eye tracking) and early acquired words are processed faster than late acquired words (e.g., Butler and Hains (1979) for lexical decision, Dirix and Duyck (2017) for eye tracking). Eye movements sometimes provided a more fine-grained pattern of results, where predictors affected measures reflecting initial lexical access stages but not higher order processing or vice versa. In some rare cases opposite results were found with both paradigms. For example, in studies of cross-lingual influences on word recognition, inhibitory effects of first language (L1) cross-lingual neighborhood density were found in a second language (L2) lexical decision task (van Heuven, Dijkstra, & Grainger, 1998), whereas facilitatory effects emerged in eye movements of L2 reading (Dirix, Cop, Drieghe, & Duyck, 2017; Whitford & Titone, 2017). Such differences originate from different task demands, and from the different strategies that yes/no lexical decisions may trigger.

The question to what extent reading times derived from these two paradigms truly converge and represent the same underlying processes, or whether they may be influenced by the same word characteristics differently has been asked before. Schilling, Rayner, and Chumbley (1998) used the same small set of 47 stimuli in a lexical decision, word naming and sentence reading task (eye movements were recorded in the latter), in a factorial design with high and low frequency words. They found moderately high correlations between lexical decision RTs and eye tracking reading times, ranging from .571 to .711. Also, frequency effects correlated between lexical decision RTs and gaze durations (but not with first fixation duration). The authors concluded that similar information on processes of word recognition can be derived from their paradigms (for further assessment of frequency effects across word production and comprehension paradigms, see Gollan et al., 2011).

A high correspondence between lexical decision and eye movement data was also reported by Hoedemaker and Gordon (2014, 2017) in their so-called “ocular lexical decision task”. In this task, participants were presented with sets of three or four letter strings and they had to make a lexical decision by either making a saccade towards the next word when they believed the letter string was a valid word, or pressing a button when they believed it was a non-word. Their eye movements were monitored during this task. Afterwards, the researchers correlated the gaze durations of this task with lexical decision RTs of the English Lexicon Project (ELP), a lexical decision database in which RTs and accuracy scores for more than 40 000 words were collected (Balota et al., 2007). In various versions of their experiment, Hoedemaker and Gordon found correlations between readings times that ranged from .36 to .59, which again suggest there seems to be some degree of overlap in the underlying processes of eye movements and lexical decision. Although the ocular lexical decision task offers a good attempt to reconcile the best of both paradigms, it is important to note that the task still contains the decision component that is typical for lexical decision. This could be a factor contributing to the correlations with the ELP data.

Kuperman, Drieghe, Keuleers, and Brysbaert (2013) built further upon Schilling et al.'s investigation by reanalyzing the latter's dataset (with up-to-date word characteristics) and analyzing three more datasets. One of the purposes of Kuperman et al.'s study was to gain insight in the validity of lexical decision RTs and eye tracking reading times, as neither of the paradigms are without controversy. Lexical decision RTs are not only influenced by the time it takes to recognize the word, but also by a decision-making component, the motor processes required to deliver the manual response and possibly response strategies that may for instance emphasize accuracy or speed. Furthermore, the non-word stimuli can heavily influence the RTs of the target stimuli: effects of word characteristics are downsized if the non-words are less word-like, so that decisions may be based on more low-level factors (Keuleers & Brysbaert, 2010, 2011). For eye tracking reading times, there is a discussion whether the duration of a fixation on a word is only influenced by the currently fixated word, or also by the preceding and the upcoming words; e.g., Engbert, Nuthmann, Richter, & Kliegl, 2005; Kliegl, Nuthmann, & Engbert, 2006). Furthermore, not only the surrounding words, but also the syntactic complexity of the sentence and the predictability of the words derived from the context could have an impact on the eye tracking reading times. Kuperman et al. argued that high correlations between lexical decision RTs and eye tracking reading times would (a) indicate that the same underlying constructs are at play, with minimal influences of specific task requirements, and (b) this would support serial processing accounts of words in text reading, without much influence of the surrounding words.

In their reanalysis of Schilling et al.'s data and an additional small dataset of 80 stimuli (without an orthogonal word frequency manipulation), Kuperman et al. found a very moderate proportion of variance shared between lexical decision RTs and eye tracking reading times, ranging from 21% (additional dataset) to 45% (Schilling et al.'s data) for first fixation durations, and from 19% (additional dataset) to 52% (Schilling et al.'s data) for gaze durations. Interestingly, they also calculated the correlations when the effects of word frequency and length were partialled out. This lowered the degree of shared variance between lexical decision times and eye movement data to 1-15% for first fixations and 5-17% for gaze durations, indicating that word frequency and word length are the dominant factors in the correlations, but also that possibly very little common variance remains between lexical decision times and eye tracking data once these two strong determinants are partialled out.

Besides possible differences between lexical decision and eye tracking, reading studies also differ in their scale, which affects the experimental design. For example, in small-scale psycholinguistic experiments, target variables are often manipulated orthogonally in a factorial design (e.g., high or low frequency crossed with early or late acquired), while other variables are controlled (e.g., word length: only words of 6 letters). In contrast, in megastudies with hundreds or thousands of target words, variables can be investigated continuously as they naturally occur in language. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) argued in favor of the latter approach when studying lexical processing. They advise researchers to be careful when categorizing continuous variables, as this can decrease statistical power or reliability, introduce potential confounds that contaminate the target factors or lead to implicit biases in experimenters and participants. Furthermore, in megastudies chances are lower to come across range restriction issues or side-effects of arbitrary “low” and “high” cut-off values. With respect to eye tracking, there is also the issue of the language context in which target words were presented (i.e., single sentences or longer passages of text that occur in a story or book), as this may affect the eye tracking reading times and the influence of word characteristics (e.g., Radach, Huestegge, & Reilly, 2008; Wochna & Juhasz, 2013; see Kliegl et al., 2006 and Rayner, Pollatsek, Drieghe, Slattery, & Reichle, 2007, for a discussion of this topic).

Because of the above issues, Kuperman et al. (2013) did not only assess the convergence between lexical decision and eye tracking in small-scale data (as just reported), but they also calculated correlations of reading times across lexical decision megastudies and eye tracking corpora. For lexical decision, data was obtained from the ELP (Balota et al., 2007). Eye tracking reading times were provided by the Dundee corpus (Kennedy & Pynte, 2005), an eye movement database of participants reading 20 newspaper articles (56,212 word tokens; 9,776 word types in total). Kuperman et al. found a substantially lower correlation for the 6,817 word types common in these databases, compared to the correlations obtained in the factorial/single sentence experiments. The proportion of shared variance, when including word frequency and length, ranged between a surprisingly low 1.3% (for first fixation duration) and 5.8% (for gaze duration) and dropped to an astounding 0.03 - 0.2% when word frequency and word length were partialled out. Similar results were obtained in an analysis of 545 common words in a smaller-scale Dutch Eye-Movement Online Internet Corpus (DEMONIC; Kuperman, Dambacher, Nuthmann, & Kliegl, 2010) and the Dutch Lexicon Project (DLP; Keuleers, Diependaele, & Brysbaert, 2010). Furthermore, Kuperman et al. plotted the word frequency effects for each of the databases they investigated and discovered two things: (a) the frequency effect seems to be smaller in eye tracking times than in lexical decision RTs and (b) the frequency effect shows a floor effect in RTs, but not eye tracking times, for frequencies around 50 per million and higher. Kuperman et al. interpreted these findings as evidence for parallel processing in reading and concluded that language context is an important determinant of reading. Indeed, the correlations for text passage reading were substantially lower than those for single sentence reading and the word frequency effect was modulated by the task and language context.

Although Kuperman et al.'s study provides interesting insights into the contributions of the lexical decision task and eye tracking to study visual word recognition, they also identified some remaining concerns. For example, the authors commented on "... the scarcity of corpus data about eye movements in reading" (p. 578) and believed that "To improve the quality of the eye movement data, it would be better to make sure that each word appears in a number of sentences presented at different times in the study" (p.578). In the current study, we elaborate on these and other issues, by investigating data of recently collected lexical decision megastudies and eye tracking corpora.

The Present Study.

Using megastudies and corpora, the present study aimed to extend Kuperman et al.'s (2013) findings by a) generalization to other datasets, b) investigating convergence of paradigms in second-language (L2) reading, and c) assessing the effect of the higher-level language context that is implied when reading a narrative/book, which is important given the large effects of language context that Kuperman et al. observed. Also, similar to Kuperman et al., we investigated effects of word length and frequency. Finally, in addition we calculated and compared the reliabilities of eye movement data and lexical decision data.

For the eye movements, data was taken from the Ghent Eyetracking COrpus (GECO; Cop, Dirix, Drieghe, & Duyck, 2017), a collection of eye movement data of English monolinguals and Dutch-English bilinguals reading an entire novel. The lexical decision RTs were provided by the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and the Dutch Lexicon Project Two (DLP2; Brysbaert, Stevens, Mandera, & Keuleers, 2016) for English and Dutch, respectively, so that we could assess task convergence for both English and Dutch. BLP was preferred over ELP because the GECO data had been collected on British participants. In line with the results of Kuperman et al. (2013), we expected low correlations between the lexical decision RTs and eye tracking measures, with an additional drop when word frequency and length effects are partialled out.

Next to replicating Kuperman et al. (2013), we also wanted to correlate the L2 reading data of GECO with a big L2 lexical decision task ran in our lab. In the last two decades, lexical decision and eye tracking paradigms also found their way into research on bilingual word recognition, so that it is very relevant to assess task convergence for L2 reading as well. If similar results are obtained in comparison to those in the L1 datasets, this would point to similar general word recognition processes in L2 as in L1 (although with a general delay, see Cop, Drieghe, & Duyck, 2015). However, Gollan et al. (2011) for example found that language context (i.e., the semantic constraint of a sentence) had a different impact on L1 vs L2 reading times. If we find higher correlations in L2 than in L1, this could indicate that the influence of the target words' characteristics is larger in L2 lexical processing; lower correlations could indicate that top-down processing and language context plays an even more important role in L2.

The third goal of this study was to further examine the role of language context, which seems to be of critical importance in the reading process, as suggested by the low correlations between reading times of words presented in isolation and those appearing in sentences (Kuperman et al., 2013). In addition, we investigated the role of multiple presentations of the target stimuli throughout the texts. More specifically, we assessed the effect of the higher-order narrative context inherent to reading a full novel (instead of separate newspaper articles of a limited length in the Dundee corpus). We correlated the timed measures of two eye tracking corpora: GECO (Cop et al., 2017) and the Dundee corpus (Kennedy & Pynte, 2005). If influences of surrounding words and higher-order language context are an important determinant of eye tracking reading times, we would expect these correlations to be fairly low. Furthermore, GECO is also suited to investigate whether multiple presentations would make a difference in the correlations with RTs. The English version consists of 54,364 words, but only 5,012 word types, implying that many words are repeated throughout the novel. We correlated lexical decision RTs with the average eye tracking reading times of words that appeared more than once, but also with the first occurrence of these words. We can expect that multiple readings of a word across different contexts converge toward lexical decision data, and therefore that repeated occurrence data would yield higher correlations across tasks.

Fourth, we further investigated the influence of word frequency and length on the dependent variables across tasks. These variables are proven to be important predictors in lexical decision (e.g., (Balota et al., 2004; Brysbaert & Cortese, 2011; New et al., 2006) and in eye movement research (e.g., Cop, Keuleers, Drieghe, & Duyck, 2015; Kliegl, Grabner, Rolfs, & Engbert, 2004; Kuperman & Van Dyke, 2011). So, even if eye tracking and lexical decision tap into different process, these important predictors should have similar effects across paradigms if they are truly important determinants of word reading. As Kuperman et al. (2013) and authors of the lexicon projects (e.g., Keuleers, Diependaele, et al., 2010; Keuleers et al., 2012) noted, the frequency effect reaches a floor effect at a frequency of approximately 50 per million. This does not seem to be the case in reading times of eye movement data. Furthermore, the frequency effect seems to be modulated by context, as a larger frequency effect was reported in lexical decision RTs than in eye tracking reading times (Kuperman et al., 2013; Schilling et al., 1998). As Kuperman et al.’s study contains the only formal comparison of the frequency effects in lexical decision and eye tracking corpora, we wanted to see whether we could obtain similar results with GECO and the recent lexicon projects. Additionally, we investigated the effect of word length. For lexical decision RTs, a U-shaped word length effect has been reported (New et al., 2006) and in eye movements the linearity of the effect seems to depend on the specific measure (e.g., Schuster, Hawelka, Hutzler, Kronbichler, & Richlan, 2016). Our approach allows us to directly compare differences (in linearity) between the word processing latencies in the dependent variables.

The final goal of this study was to compare the reliabilities of each of the dependent measures by analyzing their internal consistency. This is the first direct comparison of reliabilities of these two paradigms, which could prove to be important as this could teach us to what extent low correlations are due to little overlap in underlying processes or to the potential low reliability of the measures. We estimated the reliabilities with the Intraclass Correlation Coefficient (ICC; McGraw & Wong, 1996; Shrout & Fleiss, 1979). According to Mcgraw and Wong, the ICC can be best understood as "…a measure of the proportion of a variance … that is attributable to objects of measurement" (p. 30); in this case, the "objects of measurement" are the words read by the participants. More specifically, we calculated the ICC(3,k) measure. This type of ICC is applied for estimating the reliability of average measurements (over participants), where each item is seen by all participants, and in which the correspondence between measurements is determined in terms of consistency (as opposed to absolute agreement; See McGraw & Wong, 1996, for an overview of the various ICC measures; also see Revelle, 2018, for how ICC can be based on mixed-effects models that tolerate missing observations). This seems suitable for the current datasets: we wish to estimate the reliability of reading times that are averaged over participants and of datasets where participants were presented with (almost) every item (see Brysbaert et al., 2016 and Keuleers et al., 2012, for a similar approach in the lexicon projects). Another advantage of this particular coefficient is that it is less sensitive to missing data (Courrieu, Brand-D'abrescia, Peereman, Spieler, & Rey, 2011; Courrieu & Rey, 2011), which is appropriate for both lexical decision data, where we have missing data due to errors, and eye movement data (missing data due to word skipping).

Method

Materials

GECO. GECO is a database of eye movements of participants reading an entire novel: “The mysterious affair at Styles” of Agatha Christie (Dutch title: *“De zaak Styles”*; 1920). A group of 19 Dutch dominant bilinguals (with English as L2) read the book half in their L1 and half in their L2. Additionally, a group of 14 British English monolingual participants completed the novel in their mother tongue. For details on the corpus, the participants and the procedure we refer to Cop et al. (2017) and Cop, Drieghe, et al. (2015)

Dundee. The Dundee corpus consists of eye movement data of 10 English and 10 French participants reading 20 newspaper articles of approximately 2,800 word tokens each (see Kennedy and Pynte, 2005, for further information on the material, participants and procedure)**.** For the current study, only the English data was used.

The lexicon projects. The lexicon projects are large-scale lexical decision tasks with tens of thousands of stimuli. There are versions available in multiple languages. For the current study data was taken from the BLP (Keuleers et al., 2012) and the DLP2 (Brysbaert et al., 2016). Each involved some 40 participants per word. See the referenced publications for information on the material, procedure and participants of the lexicon projects.

L2 Lexical Decision Task. In a study of the word-level age-of-acquisition effect in L1 and L2, (Dirix & Duyck, 2017) conducted an L2 lexical decision task including 800 English words of GECO (20 Dutch-English bilingual participants per word). For further information on the stimuli, procedure and participants, see the supplementary materials of Dirix and Duyck (2017).

Results

All analyses were performed in R (version 3.4.1; R Core Team, 2017). Correlations and p-values were calculated with the stats (3.4.1) and Hmisc (4.0-3) packages. A Bonferroni correction for multiple comparisons (for the number of correlations) was applied to all reported *p*-values. Only content words were included in the stimuli selection. Function words could bias the results, as these are mostly on the high-end of the frequency scale and receive in general slower lexical decision responses than other word classes (see Brysbaert et al., 2016). The dependent variables were RTs for lexical decision (LDT) and single fixation durations (SFD), first fixation durations (FFD), gaze durations (GD) and total reading times (TRT) for eye movement measures. Zipf frequencies were taken from the SUBTLEX-UK (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) and SUBTLEX-NL (Keuleers, Brysbaert, & New, 2010) databases for English and Dutch, respectively. For the word frequency and length effects, next to the raw data we also plotted the *z*-transformed values to eliminate the scale differences between the dependent variables (cf. Kuperman et al., 2013).

L1 eye tracking and lexical decision

Monolingual English reading. There were 2,982 word types in common in the English monolingual part of GECO and the BLP (see Table 1). The lowest correlation was between LDT and FFD (r = .166, p <.001), the highest between LDT and TRT (r = .347, p <.001). For the correlations of residualized values with word frequency and length effects partialled out, the pattern was similar, although the correlations with LDT were much lower and even non-significant for SFD and FFD.

Table 1. *Correlations between English GECO reading times and British Lexicon Project reaction times (N = 2,982)*

	LDT	SFD	FFD	GD	TRT	rLDT	rSFD	rFFD	rGD	rTRT
LDT	—	.208	.166	.294	.347	.734	.038	.030	.062	.096
SFD	<.001	—	.819	.708	.574	.049	.964	.782	.661	.512
FFD	<.001	<.001	—	.742	.542	.041	.795	.979	.733	.512
GD	<.001	<.001	<.001	—	.754	.077	.623	.680	.909	.636
TRT	<.001	<.001	<.001	<.001	—	.118	.477	.470	.630	.900
rLDT	<.001	.238	.999	.001	<.001	—	.051	.041	.084	.131
rSFD	1.000	<.001	<.001	<.001	<.001	.172	—	.811	.685	.531
rFFD	1.000	<.001	<.001	<.001	<.001	.880	<.001	—	.748	.523
rGD	.022	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.700
rTRT	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

Pearson correlations above the diagonal, *p* – values (Bonferroni-adjusted for 45 comparisons) for the correlations below the diagonal. LDT = lexical decision time, SFD = single fixation duration, FFD = first fixation duration, GD = gaze duration, TRT = total reading time. The prefix "r" indicates residualized values (with effects of word frequency and word length partialled out).

The effect of word frequency for the raw and *z*-transformed data of the dependent variables is plotted in Figure 1. The effect is larger for lexical decision than for the eye tracking measures, and larger for the TRT and GD than for SFD and FFD. Furthermore, the effect in LDT seems to level off in the region around a Zipf word frequency of 4.5 (which corresponds to 50 per million raw frequency), but it stays linear for the eye tracking measures. These effects persist in the *z*-transformed dataset.

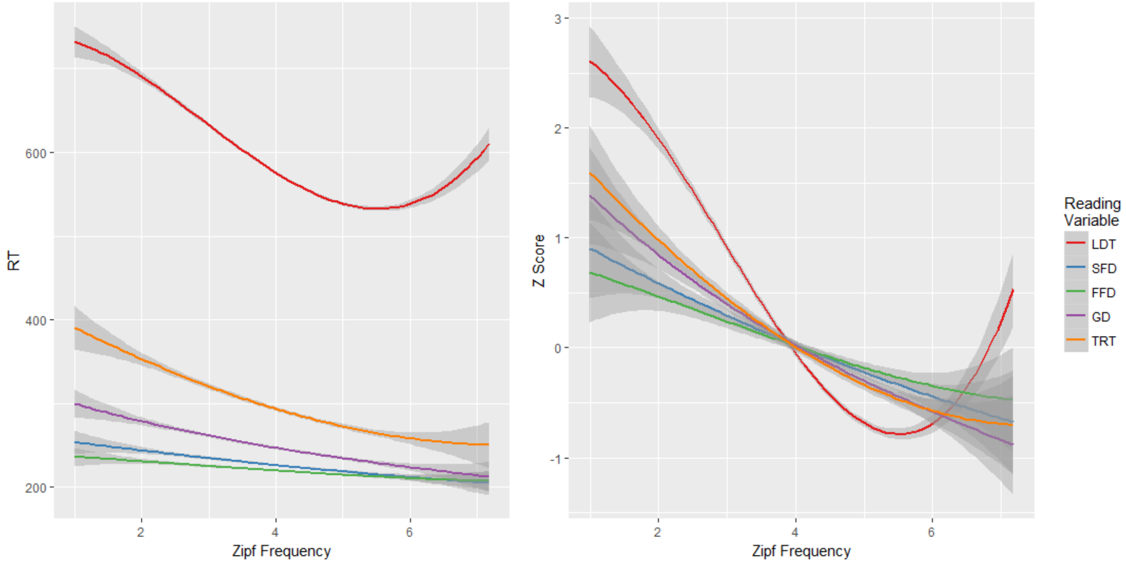

Figure 1. The effect of Zipf word frequency (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of BLP (LDT) and English monolingual GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

The effect of word length is plotted in Figure 2. Word length seems to have the largest impact on TRT and GD, followed by LDT and the smallest effect is found in the SFD and FFD. In terms of linearity, a floor effect for words up to 4-5 letters is present in LDT and both LDT and SFD seem to level off for words of 10 letters and more.

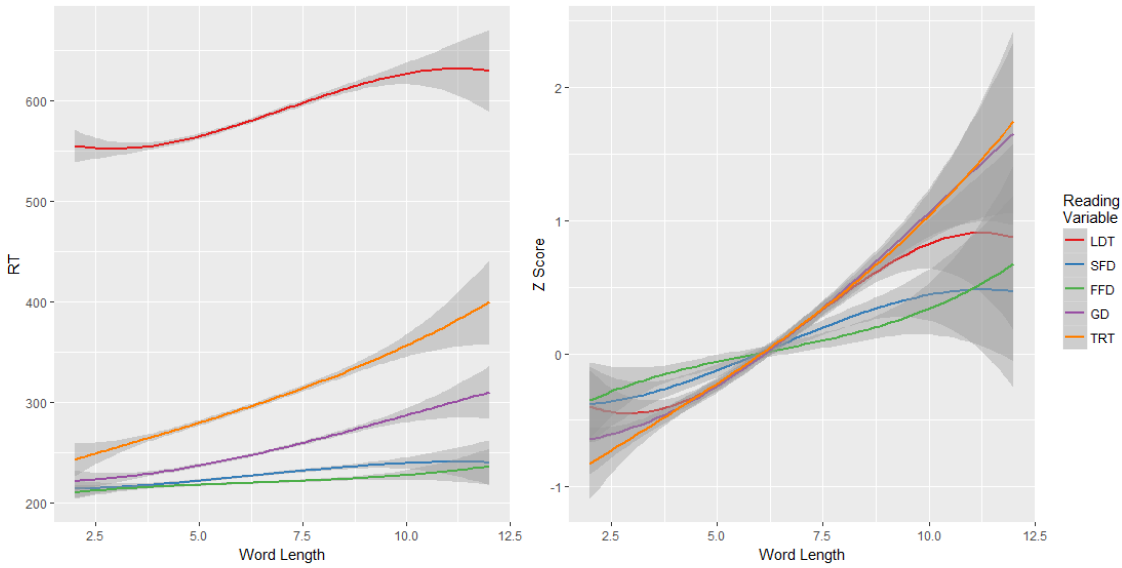

Figure 2. The effect of word length (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of BLP (LDT) and English monolingual GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

L1 Dutch reading. There were 3,188 word types in common among the Dutch L1 part of GECO and the DLP2 (see Table 2). The lowest correlation was between LDT and SFD (r = .140, p <.001), the highest again between LDT and TRT (r = .340, p <.001), which is very similar to monolingual English reading. For the correlations of residualized values with word frequency and length effects partialled out, the pattern was also similar (also to the English monolingual data): much lower correlations of LDT with the eye tracking measures and non-significant ones for SFD and FFD.

Table 2. *Correlations between Dutch GECO reading times and Dutch Lexicon Project 2 reaction times (N = 3,188)*

	LDT	**SFD**	**FFD**	**GD**	**TRT**	**rLDT**	**rSFD**	**rFFD**	**rGD**	**rTRT**
LDT	—	.140	.164	.315	.340	.830	.021	.047	.115	.142
SFD	<.001	—	.768	.619	.469	.024	.974	.738	.589	.417
FFD	<.001	<.001	—	.653	.476	.056	.741	.977	.654	.451
GD	<.001	<.001	<.001	—	.779	.121	.527	.583	.871	.616
TRT	<.001	<.001	<.001	<.001	—	.148	.372	.400	.614	.868
rLDT	<.001	1.000	.103	<.001	<.001	—	.025	.057	.138	.171
rSFD	1.000	<.001	<.001	<.001	<.001	1.000	—	.758	.604	.428
rFFD	.434	<.001	<.001	<.001	<.001	.082	<.001	—	.669	.461
rGD	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.707

rTRT	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

Pearson correlations above the diagonal, *p* – values (Bonferroni-adjusted for 45 comparisons) for the correlations below the diagonal. LDT = lexical decision time, SFD = single fixation duration, FFD = first fixation duration, GD = gaze duration, TRT = total reading time. The prefix "r" indicates residualized values (with effects of word frequency and word length partialled out).

The effect of word frequency for raw and *z*-transformed data of Dutch lexical decision and reading is plotted in Figure 3. The effect again is larger for lexical decision than for the eye tracking measures, and larger for TRT and GD than for SFD and FFD. Furthermore, the effect in LDT again seems to level off in the region around 4.5 Zipf frequency, but it remains more linear for the eye tracking measures. These effects persist in the *z*-transformed dataset.

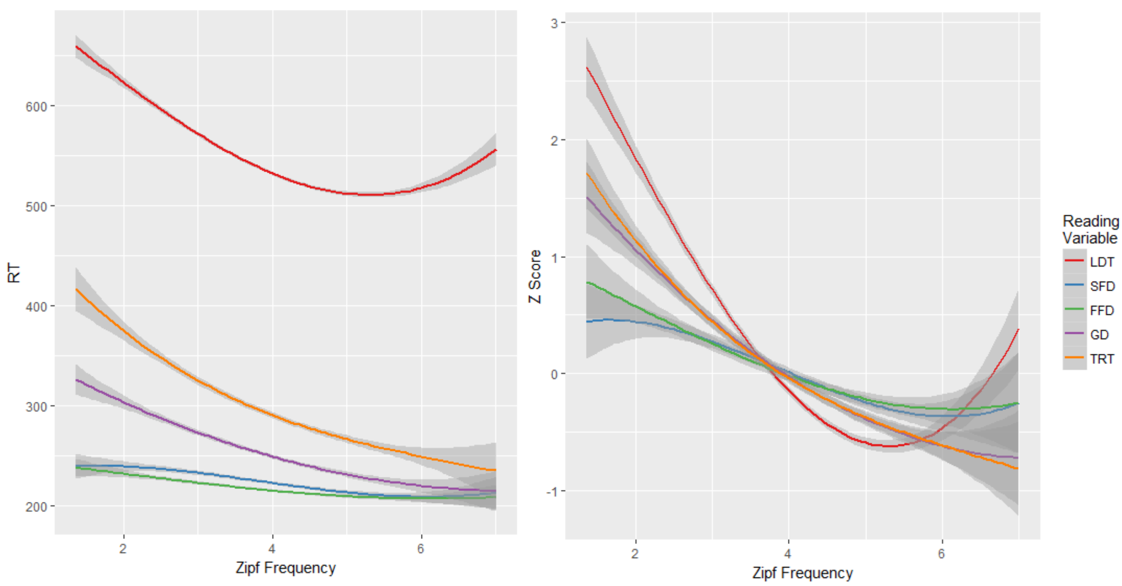

Figure 3. The effect of Zipf word frequency (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of DLP2 (LDT) and Dutch L1 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

The word length effect for the Dutch dataset is plotted in Figure 4. Word length again seems to have the largest impact on TRT and GD, followed by LDT and the smallest effect is found on SFD and FFD. In terms of linearity, a floor effect for words up to 6-7 letters is present in LDT and a ceiling effect can be observed in SFD and FFD for words of 10 letters and more.

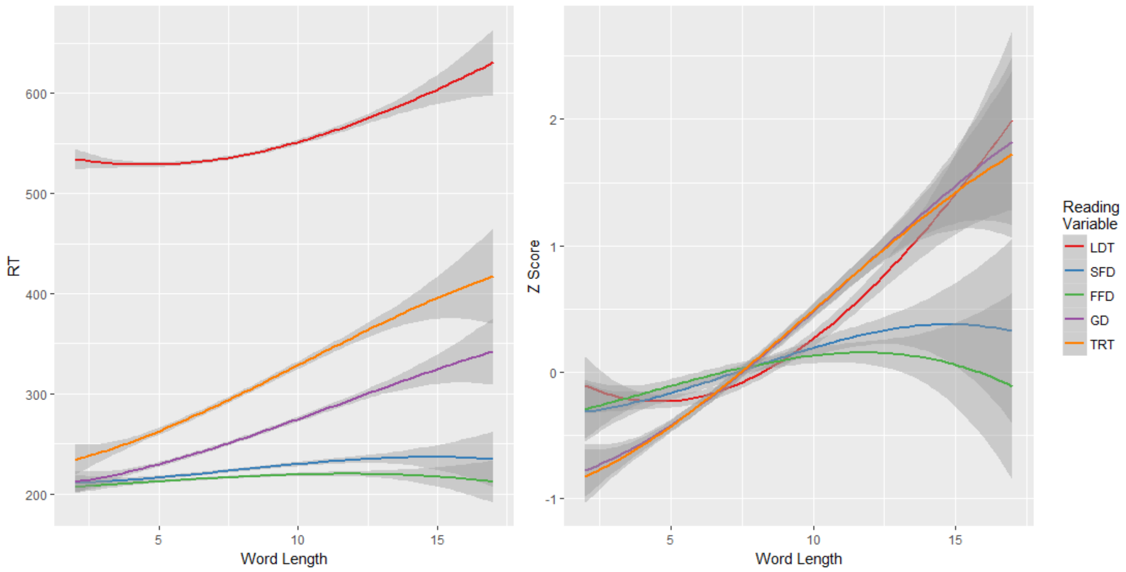

Figure 4. The effect of word length (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of DLP2 (LDT) and Dutch L1 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

L2 reading and lexical decision

There were 791 word types in common in the English L2 reading part of GECO and the L2 lexical decision task (see Table 3). The pattern, but also the magnitude of the correlations was strikingly similar to those of English and Dutch L1 reading data. The lowest correlation was between LDT and SFD (r = .181, p <.001), the highest between LDT and TRT (r = .329, p <.001). For the correlations of residualized values, the correlations were again much lower compared to those of the raw data, those of LDT with SFD and FFD were not significant.

Table 3. *Correlations between L2 English GECO reading times and L2 lexical decision reaction times (N = 791)*

	LDT	SFD	FFD	GD	TRT	rLDT	rSFD	rFFD	rGD	rTRT
LDT	—	.181	.189	.271	.329	.810	.071	.074	.110	.149
SFD	<.001	—	.771	.628	.504	.086	.978	.746	.599	.461
FFD	<.001	<.001	—	.674	.441	.089	.747	.979	.664	.405
GD	<.001	<.001	<.001	—	.743	.125	.561	.621	.915	.633
TRT	<.001	<.001	<.001	<.001	—	.167	.427	.374	.626	.906

rLDT	<.001	.726	.546	.020	<.001	—	.087	.091	.136	.184
rSFD	1.000	<.001	<.001	<.001	<.001	.623	—	.763	.613	.472
rFFD	1.000	<.001	<.001	<.001	<.001	.470	<.001	—	.678	.413
rGD	.084	<.001	<.001	<.001	<.001	.005	<.001	<.001	—	.691
rTRT	.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

Pearson correlations above the diagonal, *p* – values (Bonferroni-adjusted for 45 comparisons) for the correlations below the diagonal. LDT = lexical decision time, SFD = single fixation duration, FFD = first fixation duration, GD = gaze duration, TRT = total reading time. The prefix "r" indicates residualized values (with effects of word frequency and word length partialled out).

The effect of word frequency for raw and *z*-transformed L2 data is presented in Figure 5. The general pattern reoccurs in the L2 data: the effect again is larger for LDT than for the eye tracking measures, and larger for TRT and GD than for SFD and FFD. Furthermore, the effect in LDT again seems to level off, now in the region around 5 Zipf frequency, but it stays more linear for the eye tracking measures. These effects also persist in the *z*-transformed dataset.

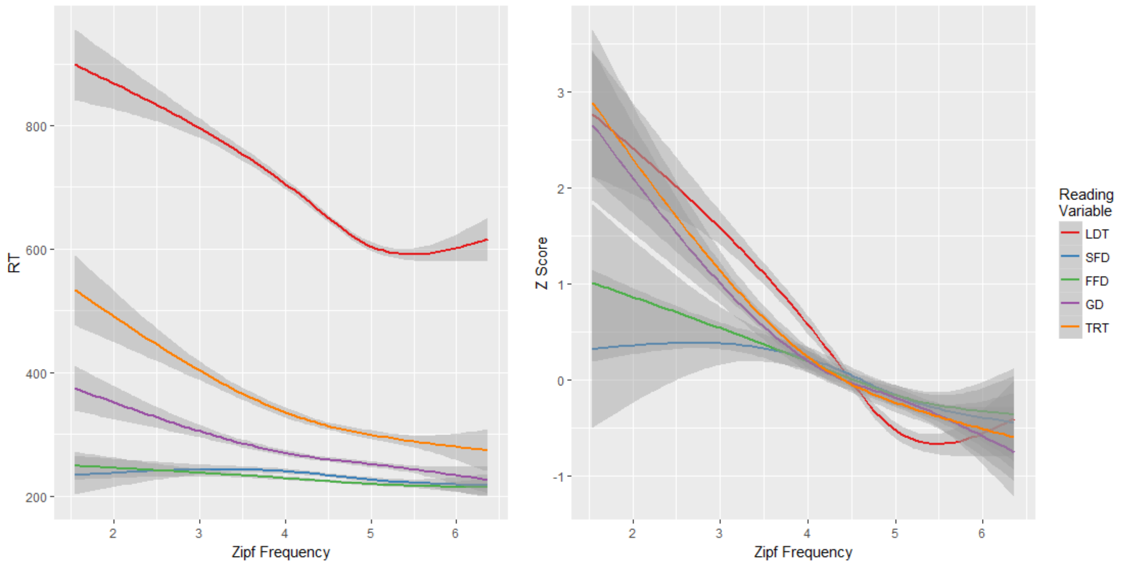

Figure 5. The effect of Zipf word frequency (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of the L2 lexical decision task (LDT) and English L2 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

The word length effect for the L2 dataset is plotted in Figure 6. Similar to L1 data, word length has the largest impact on TRT and GD, followed by LDT, and least on SFD and FFD. The floor effect in LDT again appears for words up to 5 letters, but now there seems to be a similar floor effect in SFD and FFD. The ceiling effect in SFD and FFD also seems to emerge somewhat earlier, for words of length 8 and more.

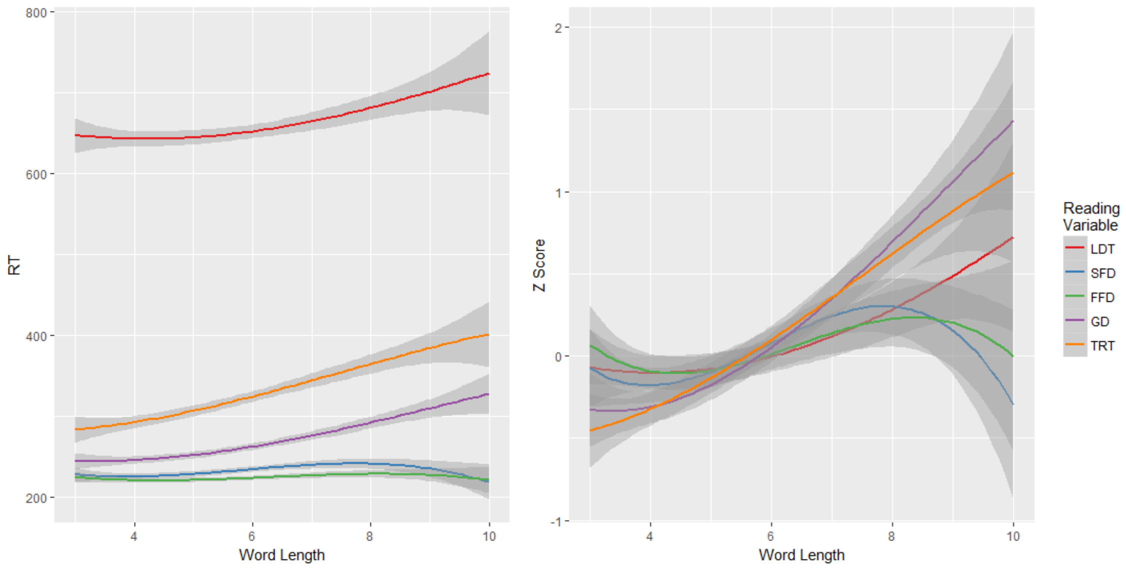

Figure 6. The effect of word length (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of the L2 lexical decision task (LDT) and English L2 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

The influence of language context and repeated presentations

GECO – Dundee correlations. The correlations and *p*-values between the eye tracking measures of GECO and the Dundee corpus are presented in table 4. There were 1,954 word types in common in these corpora. The correlations between the raw eye tracking reading times were very low (even lower than the correlations of eye tracking reading times and LDT), ranging from .048 for SFD to .187 for TRT, even though both are from eye tracking corpora. Only the correlations for GD and TRT reached significance. The proportion of shared variance ranges from 0.01 to 0.16% when word frequency and length effects are partialled out, but none of the correlations between the residualized values were significant.

Table 4. *Correlations between monolingual English GECO reading times and Dundee corpus reading times (N = 1,954)*

	SFD	**FFD**	**GD**	**TRT**	**SFDd**	**FFDd**	**GDd**	**TRTd**	**rSFD**	**rFFD**	**rGD**	**rTRT**	**rSFDd**	**rFFDd**	**rGDd**	**rTRTd**
SFD	—	.831	.711	.577	**.081**	.053	.097	.112	.973	.803	.679	.533	.021	.007	.003	.025
FFD	<.001	—	.742	.566	.079	**.048**	.072	.085	.813	.984	.742	.550	.033	.012	.007	.025
GD	<.001	<.001	—	.751	.111	.070	**.180**	.178	.644	.695	.922	.653	.022	.004	.018	.026
TRT	<.001	<.001	<.001	—	.110	.072	.197	**.187**	.506	.516	.653	.923	.021	.006	.037	.038
SFDd	**.060**	.081	<.001	<.001	—	.856	.677	.588	.021	.033	.023	.022	.964	.82	.636	.537

Pearson correlations above the diagonal, *p* – values (Bonferroni-adjusted for 120 comparisons) for the correlations below the diagonal. LDT = lexical decision time, SFD = single fixation duration, FFD = first fixation duration, GD = gaze duration, TRT = total reading time. The suffix “d” indicates variables from the Dundee corpus. The prefix “r” indicates residualized values (with effects of word frequency and word length partialled out). Correlations and *p*-values between the same variables of the two corpora are in **bold**.

FFDd	1.000	**1.000**	.306	.240	<.001	—	.712	.580	.007	.012	.004	.006	.832	.979	.704	.554
GDd	.004	.242	**<.001**	<.001	<.001	<.001	—	.839	.003	.007	.017	.037	.598	.652	.906	.731
TRTd	<.001	.030	<.001	**<.001**	<.001	<.001	<.001	—	.023	.023	.026	.037	.511	.520	.741	.918
rSFD	<.001	<.001	<.001	<.001	1.000	1.000	1.000	1.000	—	.826	.698	.548	**.022**	.007	.003	.026
rFFD	<.001	<.001	<.001	<.001	1.000	1.000	1.000	1.000	<.001	—	.754	.559	.034	**.012**	.008	.025
rGD	<.001	<.001	<.001	<.001	1.000	1.000	1.000	1.000	<.001	<.001	—	.708	.024	.004	**.019**	.029
rTRT	<.001	<.001	<.001	<.001	1.000	1.000	1.000	1.000	<.001	<.001	<.001	—	.023	.006	.041	**.041**
rSFDd	1.000	1.000	1.000	1.000	<.001	<.001	<.001	<.001	**1.000**	1.000	1.000	1.000	—	.850	.660	.557
rFFDd	1.000	1.000	1.000	1.000	<.001	<.001	<.001	<.001	1.000	**1.000**	1.000	1.000	<.001	—	.719	.566
rGDd	1.000	1.000	1.000	1.000	<.001	<.001	<.001	<.001	1.000	1.000	**1.000**	1.000	<.001	<.001	—	.807
rTRTd	1.000	1.000	1.000	1.000	<.001	<.001	<.001	<.001	1.000	1.000	1.000	**1.000**	<.001	<.001	<.001	—

First occurrence vs repeated presentations. We found 1,915 word types in common in the BLP and monolingual English GECO that were presented more than once throughout the novel. The correlations between the average eye tracking reading times of all word occurrences, the first occurrence and LDT are presented in Table 5. The general pattern of lower correlation between LDT and timed eye movement measures for SFD/FFD and higher correlations for TRT/GD appears in both the "all occurrences" and "first occurrence" datasets. However, there is an increase of about .10 in the correlations with LDT when all occurrences are taken into account compared to only the first occurrence, which results in an increase in shared variance from 1.4 to 5.3% for FFD and 7.1 to 13.3% for TRT. There is also an increase in the correlations of the residualized values (except for TRT), although they remain very low.

Note that the shared variance of eye tracking reading times between the first occurrence and all occurrences of the same word is about 27 to 39% for raw data, this stays approximately the same for residualized values (26% - 32%). To make sure that the correlations were not limited to the first reading vs. all readings, we decided to run an additional analysis to investigate the correlations between eye movement measures at different occurrences of the same word within the GECO data. For the 1,915 words that were presented at least twice in the corpus, we selected two random occurrences. This random selection was applied at the participant level, so that the selected presentations of words that occurred more than twice were different ones for each participant. The correlations ranged between .116 for SFD and .371 for TRT (see Table 6), which is higher than those between GECO and DUNDEE, but lower than the correlations between the first and all later occurrences.

Table 5. *Correlations between monolingual English GECO reading times of words with more than one occurrence, the reading times of their first occurrence and British Lexicon Project reaction times (N = 1,915)*

	LDT	**SFD**	**FFD**	**GD**	**TRT**	**rLDT**	**rSFD**	**rFFD**	**rGD**	**rTRT**	**SFD1**	**FFD1**	**GD1**	**TRT1**	**rSFD1**	**rFFD1**	**rGD1**	**rTRT1**
LDT	—	.230	.215	.344	.364	.797	.059	.076	.105	.124	.117	.115	.225	.268	.041	.048	.077	.122
SFD	<.001	—	.849	.743	.615	.070	.952	.801	.684	.538	.545	.450	.427	.365	.508	.420	.354	.287
FFD	<.001	<.001	—	.760	.579	.092	.818	.972	.754	.549	.445	.516	.414	.321	.418	.492	.368	.271
GD	<.001	<.001	<.001	—	.801	.115	.625	.675	.870	.638	.395	.401	.603	.490	.334	.355	.478	.355
TRT	<.001	<.001	<.001	<.001	—	.133	.486	.486	.631	.859	.319	.293	.483	.623	.255	.246	.347	.488
rLDT	<.001	.444	.012	<.001	<.001	—	.074	.095	.132	.155	.051	.06	.093	.145	.051	.060	.097	.153
rSFD	1.000	<.001	<.001	<.001	<.001	.268	—	.842	.718	.565	.529	.439	.355	.287	.534	.442	.372	.302
rFFD	.189	<.001	<.001	<.001	<.001	.008	<.001	—	.776	.565	.426	.503	.362	.266	.430	.507	.379	.279
rGD	.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.734	.380	.406	.525	.387	.384	.409	.550	.408
rTRT	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.294	.284	.386	.540	.297	.286	.404	.568
SFD1	<.001	<.001	<.001	<.001	<.001	1.000	<.001	<.001	<.001	<.001	—	.827	.683	.503	.991	.818	.675	.488
FFD1	<.001	<.001	<.001	<.001	<.001	1.000	<.001	<.001	<.001	<.001	<.001	—	.741	.507	.82	.994	.745	.503
GD1	<.001	<.001	<.001	<.001	<.001	.012	<.001	<.001	<.001	<.001	<.001	<.001	—	.747	.651	.716	.955	.690
TRT1	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.468	.481	.686	.951
rSFD1	1.000	<.001	<.001	<.001	<.001	1.000	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.826	.681	.492
rFFD1	1.000	<.001	<.001	<.001	<.001	1.000	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.75	.506
rGD1	.151	<.001	<.001	<.001	<.001	.005	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.722
rTRT1	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

Pearson correlations above the diagonal, *p* – values (Bonferroni-adjusted for 153 comparisons) for the correlations below the diagonal. LDT = lexical decision time, SFD = single fixation duration, FFD = first fixation duration, GD = gaze duration, TRT = total reading time. The suffix "1" indicates the reading times of the first occurrence of the words from GECO. The prefix "r" indicates residualized values (with effects of word frequency and word length partialled out).

Table 6. *Correlations between two random occurrences of the same word in the English GECO reading times (N = 1,915)*

	SFD1	FFD1	GD1	TRT1	SFD2	FFD2	GD2	TRT2
SFD1	—	.809	.634	.460	**.116**	.124	.163	.176
FFD1	< .001	—	.746	.517	.135	**.146**	.146	.155
GD1	< .001	< .001	—	.738	.186	.176	**.279**	.294
TRT1	< .001	< .001	< .001	—	.192	.166	.287	**.371**
SFD2	< .001	< .001	< .001	< .001	—	.827	.673	.514
FFD2	< .001	< .001	< .001	< .001	< .001	—	.734	.529
GD2	< .001	< .001	< .001	< .001	< .001	< .001	—	.722
TRT2	< .001	< .001	< .001	< .001	< .001	< .001	< .001	—

Pearson correlations above the diagonal, *p* – values (Bonferroni-adjusted for 28 comparisons) for the correlations below the diagonal. SFD = single fixation duration, FFD = first fixation duration, GD = gaze duration, TRT = total reading time. The suffix "1" indicates the reading times of the first random occurrence of the words; the suffix "2" indicates the reading times of the second random occurrence. Correlations between the same variables of the two occurrences are in **bold**. *n* = 1915

Reliabilities of the datasets

The low correlations between GECO and the lexical decision megastudies raise the question whether these are caused by the fact that they reveal different reading processes, or that (some of) the different measures may not very reliable. The ICC(3, k) values for each of the dependent variables of GECO, BLP, DLP2 and the L2 lexical decision task analyzed in the current study are presented in Table 7. For the GECO data, the average reading times of all presentations per person were included to calculate the reliabilities. A consistent pattern emerges, indicating that the internal consistency of eye movement measures representing one fixation (SFD and FFD) is the lowest, followed by LDT, and the highest reliability values present themselves for TRT and GD. We also applied a correction for attenuation (based on the ICC values) on the correlations between the LDT and eye movement measures[1]. This

[1] The formula for this correction is $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$, where r_{xy} is the correlation between variable *x* and variable *y*, r_{xx} is the reliability of variable *x* and r_{yy} is the reliability of variable *y*. This correction can only be applied to independent variables and, therefore, cannot be calculated for the correlations between the different eye movement measures.

correction suggests that these correlations are probably somewhat underestimated due to the internal consistencies of the datasets, although they do not increase dramatically.

Table 7. *Correlations, reliabilities and correlations corrected for attenuation of the Dutch L1 datasets (GECO and DLP2), English L1 datasets (GECO and BLP), and English L2 datasets (GECO and L2 lexical decision)*

	Dutch (L1)					**English (L1)**					**English (L2)**				
	LDT	**SFD**	**FFD**	**GD**	**TRT**	**LDT**	**SFD**	**FFD**	**GD**	**TRT**	**LDT**	**SFD**	**FFD**	**GD**	**TRT**
LDT	*.782*	.140	.164	.315	.340	*.816*	.208	.166	.294	.347	*.744*	.181	.189	.271	.329
SFD	.220	*.517*	.768	.619	.469	.302	*.579*	.819	.708	.574	.269	*.611*	.771	.628	.504
FFD	.253	—	*.536*	.653	.476	.234	—	*.616*	.742	.542	.282	—	*.605*	.674	.441
GD	.381	—	—	*.875*	.779	.353	—	—	*.853*	.754	.337	—	—	*.874*	.743
TRT	.406	—	—	—	*.894*	.406	—	—	—	*.894*	.401	—	—	—	*.906*

Pearson correlations above the diagonal, ICC(3, k) values on the diagonal, correlations corrected for attenuation below the diagonal. The correction can only be calculated for variables that were independently collected, so not for the various eye movement measures. LDT = lexical decision time, SFD = single fixation duration, FFD = first fixation duration, GD = gaze duration, TRT = total reading time.

Discussion

By analyzing large datasets from recent eye movement and lexical decision corpora, we attempted to accomplish five goals. First, we wanted to generalize Kuperman et al.'s findings (2013) to larger corpora and other datasets, establishing whether indeed the proportion of shared variance between passage eye tracking reading times and lexical decision RTs is low, especially when controlling for the effects of word frequency and length. Second, we investigated L2 eye tracking reading times and LDT RTs, to see whether similar results are found in L2 processing. The third goal was to investigate the influence of language context (narratives) and repeated presentations by comparing two eye movement corpora and the eye tracking reading times of the first vs all occurrences of words presented more than once, respectively. The fourth goal was to compare the roles of two important predictors of word processing latencies in these paradigms: word frequency and word length. Finally, we assessed the internal consistencies of each of the measures investigated in the current study, in order to investigate whether low correlations reveal that different tasks tap into different reading processes, rather than low reliability. We discuss each of these topics below.

Correlations between lexical decision RTs and eye movement measures

In general, the pattern of correlations we observed between BLP/DLP2 RTs and English/Dutch GECO reading times was highly similar to the results reported in Kuperman et al. (2013): a fairly low correlation overall, and an important contribution of word frequency and length effects to these correlations. A minor difference was that we consistently found the highest correlations between LDT and TRT, whereas in Kuperman et al.'s study the highest correspondence was found between LDT and GD. Their reasoning that LDT possibly includes semantic processing, thus corresponding more to late eye movement measures, also applies to our data. Furthermore, we considered the option that the correlations in our study could be even lower as the text material of GECO consists of a novel rather than the newspaper articles in the Dundee corpus, and hence constitute an even more elaborated higher-order language context. In contrast, the correlations in our study turned out to be slightly higher than Kuperman et al.'s (except for Dutch SFD) with differences ranging from .044 to .117. One possible reason might be the slightly better fit between databases because of the geographical correspondence of the participants: British students for BLP and English GECO and Dutch (Flemish) students for DLP2 and Dutch GECO, whereas US students took part in the ELP and British students in the Dundee study. This geographical correspondence has indeed been found earlier, as for example British SUBTLEX-UK (van Heuven et al., 2014) word frequencies accounted for 3% more variance in BLP data (Keuleers et al., 2012) than their US equivalents (SUBTLEX-US; Brysbaert & New, 2009). An alternative explanation could be a difference in the number of word repetitions in the texts of the corpora, which we discuss below.

In correspondence with the conclusions of Kuperman et al. (2013), the current results provide further evidence that lexical decision RTs are not a very good predictor for timed eye movement measures. Both paradigms partially tap into different reading processes, and lexical decision may include additional decision-making strategies. Also, the language context that is inherent to eye tracking (which almost always uses sentences instead of isolated words, and in the present study even an entire story) provides top-down influences on reading that minimizes the effects of word characteristics such as word frequency and length: they are of less importance when reading longer passages of texts compared to single sentences (cf. Radach et al., 2008). Recently, a context modulation of word characteristic effects was also found in a lexical decision task when intermixed with a self-paced reading (Teng, Wallot, & Kelty-Stephen, 2016), lending further support for the importance of the context in which target words are presented.

An additional factor that could partially explain the low correlations between LDT and eye movement measures is the differences in sample sizes between the datasets, but also the different participant samples for each of the databases. The number of participants indeed differs between the lexicon projects (some 40 readers per word) and the eye tracking corpora (10 – 20 observations per word token). The smaller number of participants in the eye movement studies was compensated by the fact that many word tokens were seen several times, likely resulting in more stable estimates.

Another issue may be the fact that LDT and eyetracking were done by different participants. As Carter and Luke (2018) noted, "…*who* is reading may be a larger determinant of eye movement behaviors in reading than *what* is being read." (p.487). They showed that there are considerable differences in reading speed between participants, which could influence the correspondence between tasks with different participant samples. It seems plausible that the correlations could be higher with an identical sample of participants performing both tasks: this was the case in the study by Schilling et al. (1998), where the reported correlations were indeed higher than those of the current study. This dataset of Schilling et al. is however limited in the number of stimuli (47 words), its factorial design, and the fact that the target words were presented in isolated, uninformative sentences. Unfortunately, no dataset of participants performing both a large-scale lexical decision and eye tracking task is available (yet) to check to what extent this would indeed increase the correspondence between the measures. Furthermore, taking into account (a) the high internal consistencies of GECO and the lexicon projects, which show that variations in reading times (between different words) are similar across participants, (b) the fairly high correspondence of the BLP and ELP (a correlation of .77 for z-transformed RTs; Keuleers et al., 2012), and (c) the fact that DLP2 and the Dutch part of GECO were collected at the same university using the same participant pool, it seems somewhat unlikely that having the same participant sample complete both tasks would dramatically increase the correlations.

A final (small) factor that may contribute to the limited correlation between LDT and eye movement measures is the fact that the contributions of the two most prominent word characteristics differ between tasks. Word frequency has a stronger effect in LDT, whereas word length is more dominant in eye movement measures.

Convergence across reading tasks for L2 reading

Next to the Dutch and English L1 data, we also analyzed the convergence of English L2 eye tracking times and lexical decision RTs, for the first time. We were interested to see whether the pattern of correlations was similar to that of L1 data, as for example (Gollan et al., 2011) reported different effects of semantic constraint on L1 vs. L2 eye tracking measures, indicating that the language context could be of more importance in L2 reading. The pattern of correlations in L2 was however strikingly similar to that of L1, indicating that the influence of context and word characteristics manifest themselves in similar ways for second-language reading, although L2 processing is usually slower and word-level effects tend to be more pronounced than in L1 (e.g., larger word frequency effects in L2 compared to L1; Brysbaert, Lagrou, & Stevens, 2017; Cop, Keuleers, et al., 2015; Duyck, Vanderelst, Desmet, & Hartsuiker, 2008)

The influence of language context and repeated presentations

We investigated the role of different language contexts by correlating eye movement data from two corpora, contrasting reading of newspaper articles with the semantic context of a full book. The correlations between the Dundee corpus (Kennedy & Pynte, 2005) and GECO (Cop et al., 2017) were surprisingly low, as these are reading times of identical words in a similar paradigm, with shared variances ranging between 0.2% (FFD) and 3.5% (TRT). Furthermore, the shared variance between the first and all occurrences of the same word in GECO (27% for FFD; 39% for TRT) also indicated that even within the same corpus, the local language context is of importance, as the first reading times correlated only moderately with the reading times of later occurrences. In an additional analysis we further investigated this by correlating the reading times of two random occurrences of the same word within GECO. This resulted again in fairly low correlations, with shared variances ranging between 1% (SFD) to 14% (TRT). There were fewer observations per word in this analysis compared to the one including all occurrences, possibly resulting in a less reliable estimate of the reading times, which could partially explain the lower proportion of shared variance. Furthermore, as this analysis concerns reading times of the same participants reading identical words, but embedded in two different sentences, these results also further point towards the crucial role of words surrounding the target words, (such as predictability of the target word or spill-over effects) or the broader top-down language context of the narrative. In terms of eye movement control, these results seem to be in line with models that include some parallel processing (e.g., Engbert et al., 2005; Kliegl et al., 2006).

Next, we found that averaging eye tracking reading times across repeated presentations increased correlations between LDT and GECO measures. So, in future eye movement corpora studies researchers are recommended to make sure that target words are presented several times in different contexts. Eye movement research seems to need multiple presentations of target words in order to approximate effects of word-level variables like they are observed in lexical decision. Two factors are likely to contribute to this. First, averaging reading times across various language contexts arguably yields a more context-free reading estimate. Second, averaging reading times decreases the noise in the variable and leads to a more stable estimate. Alternatively, it could be argued that the influence of word-level variables is overestimated in lexical decision, because words are presented out of context and must be separated from non-existing alternatives. So, the optimal paradigm may depend on the research question. If the goal is to assess the potential of effects of experimental manipulations of word-level variables (like frequency or length), independent from real-life context, orthogonal designs with a lexical decision task is preferable. If, however, the goal is to assess the relevance of certain language variables for natural reading, eyetracking is more suitable.

Note that averaging reading times across multiple contexts and repetitions may also explain why we observed slightly higher correlations between LDT and eye tracking times than Kuperman et al. (2013). Both English eye movement corpora contain approximately 56,000 word tokens, but these correspond to some 10,000 word types in the Dundee corpus (which Kuperman et al. analyzed) versus 5,000 for GECO. Indeed, in the subset of words we analyzed, there were on average more presentations in multiple contexts in GECO (M = 11.76) than in Dundee (M = 8.95; t = 2.497, $p < .05$). Hence, it is plausible that GECO measures approximate LDT better, because averaging across sentence contexts yields an estimate that is relatively more context-independent and because more observations lead to more stable estimates.

Finally, an alternative explanation for the low correlations between the two eye tracking corpora could be that different participants took part in the studies. In a recent study, Carter and Luke (2018) showed that the reading times of participants reading 40 paragraphs in two reading sessions (20 paragraphs per session), separated by a month, were very consistent: they reported correlations of .93 for FFD and .72 for TRT between the two sessions. This suggests that if the same participants were to read the texts of both GECO and Dundee, correlations might be higher (in line with our discussion of the correlations between LDT and eye tracking reading times). Important to note however is that the reading times reported by Carter and Luke are an overall average per participant across words, which yields a general measure of individual reading speed. Here, we looked at the stability of reading times per word. It could be that the global or overall reading speed per person is not very informative for the reading times of specific, individual words. In further support of this claim, we calculated the correlation between the average overall TRT, across words, of the first and last reading session of the monolingual participants of GECO[2]; it was .978 ($N = 14$, $p < .001$). In contrast, the correlations between reading times of random occurrences of two identical words in GECO were quite low, which suggests indeed that the stability of reading rate per person contains little information about the correspondence of individual word reading times in different reading sessions / language contexts.

Word frequency and word length effects

The processing of words embedded in a discourse context is influenced by top-down factors (semantic language context, grammatical restrictions, etc.) that minimize the importance of word-level variables on reading times. Indeed, confirming the results of

[2] Participants were required to read the novel in 4 separate sessions (Cop et al., 2017).

previous studies, the word frequency effect is larger for lexical decision than for timed eye tracking measures (Kuperman et al., 2013), and seems to show a floor effect in lexical decision RTs, starting at a Zipf word frequency of approximately 4.5 (50 per million raw frequency; Keuleers, Diependaele, et al., 2010).

The word length effect also reached a floor effect for lexical decision RTs for the short words (4-7 letters; the onset of the floor effect seemed to be earlier in DLP2 than BLP). We could not replicate the U-shaped curve reported by New et al. (2006) for ELP, but this might be due to the scarcity of short words in our analyzes (the confidence intervals were indeed larger on the short end of the word length scale). Another reason might be that ELP contained more longer words than BLP, so that participants were more surprised when a short word was presented in ELP than in BLP. Relatively speaking, the word length effect was larger in gaze duration and total reading time than in lexical decision. In FFD and SFD the effect seemed to be smaller than in LDT and also showed a ceiling effect, starting at around 9-10 letters, in line with the well-known observation that long words are often fixated more than once.

The effects in the L2 data were very similar to those of L1, and the floor effect of word frequency in LDT was reached roughly in the same region (around 50-100 per million). A floor effect of word length also appeared in SFD and FFD. It is probably the case that the speed limit of visual word processing was reached earlier in L2, as L2 processing seems to be occurring at a slower rate (e.g., Cop, Drieghe, et al., 2015; Duyck, Van Assche, Drieghe, & Hartsuiker, 2007).

All of the above discussed word-level effects and differences between the dependent variables were not due to scale differences, as they persisted in the standardized z-value data.

In general, it is important to note here that these important determinants of reading times still exerted effects across paradigms, notwithstanding the interesting differences discussed above. Even if eyetracking and lexical decision tap into different processes to an important degree, at least this confirms the relevance of these variables when studying reading, across paradigms. Also here, the optimal paradigm to assess such effects depends on the focus of the research question: the context-free, pure effect of frequency manipulations may reveal themselves more clearly in an isolated LDT, whereas the relevance of such manipulations and effects for natural reading may require eyetracking data, narrative materials, and multiple observations.

Reliability of the variables

It is important to know whether low convergence between eye tracking and LDT data results from the fact that both paradigms differentially tap into different reading processes, or whether some of the measures may suffer from low psychometric reliability. A variable cannot correlate more with another variable than with itself. To this end, we assessed and compared the reliability of the datasets analyzed in the current study (this was not done by Kuperman et al., 2013). The ICC values of the subsets of BLP (Keuleers et al., 2012) and DLP2 (Brysbaert et al., 2016) data are comparable to the values of the entire datasets reported in the referenced studies. For the reading times of GECO subset (Cop et al., 2017), the reliabilities of GD (.85 for L1 reading of English) and TRT (.89 for L1 reading of English) were similar to those of the full dataset, and in fact higher than the respective reliabilities for the LDT. The lower reliabilities for LDT can probably be explained by the higher standard deviations in lexical decision times and the fact that each words was only seen once per participant (Brysbaert & Stevens, 2018). The high reliabilities in GD and TRT movement measures indicate that the time needed to fully process a word seems to be highly consistent across participants. Reliabilities were remarkably lower for SFD and FFD, which are eye movement duration measures representing only a single fixation. This could be due to landing errors in first fixations, differences in reading strategies during the first encounter of a word or differences in individual characteristics. Indeed, Kuperman and Van Dyke (2011) found that individual differences accounted for more variability in early word processing stages than word characteristics. The L2 data again showed a similar pattern to that of the L1 datasets. These high within-task reliabilities show that the low correlations observed across tasks are likely to be due to task-specific processing demands and language context influences, and not to suboptimal measurement of the variables (although improvements are always possible).

Conclusion

The present study showed that reading times from different paradigms (LDT vs eye tracking) diverge considerably, across multiple languages and large corpora/databases, and both in L1 and L2 reading. Also across eye tracking corpora, correlations of reading times were low, although within-task reliability was high, illustrating the strong effect of language context. When aggregating eye tracking measures across multiple representations and contexts, convergence with LDT increased. These results indicate that reading research should be aware of the impact of task-specific language context on the manifestation of word-level effects.

Acknowledgements

This research was funded by a concerted research action, Grant No. BOF13/GOA/032, from Ghent University.

The authors would like to thank Dr. Alan Kennedy for sharing the data of the Dundee corpus, and Dr. Michael Stevens for sharing his R code for calculating ICC values.

References

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), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12. https://doi.org/10.16910/jemr.2.1.1

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology*, *64*(3), 545–559. https://doi.org/10.1080/17470218.2010.503374

Brysbaert, M., Lagrou, E., & Stevens, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, *20*(03), 530–548. https://doi.org/10.1017/S1366728916000353

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(9), 1–20. https://doi.org/10.5334/joc.10

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. https://doi.org/10.1037/xhp0000159

Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. *Memory &*

Cognition, *7*(2), 68–76. https://doi.org/10.3758/BF03197587

Carter, B. T., & Luke, S. G. (2018). Individuals' eye movements in reading are highly consistent across time and trial. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(3), 482–492. https://doi.org/10.1037/xhp0000471

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. https://doi.org/10.3758/s13428-016-0734-0

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye Movement Patterns in Natural Reading: A Comparison of Monolingual and Bilingual Reading of a Novel. *PLOS ONE*, *10*(8), e0134008. https://doi.org/10.1371/journal.pone.0134008

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, *22*, 1216–1234. https://doi.org/10.3758/s13423-015-0819-2

Courrieu, P., Brand-D'abrescia, M., Peereman, R., Spieler, D., & Rey, A. (2011). Validated intraclass correlation statistics to test item performance models. *Behavior Research Methods*, *43*(1), 37–55. https://doi.org/10.3758/s13428-010-0020-5

Courrieu, P., & Rey, A. (2011). Missing data imputation and corrected statistics for large-scale behavioral databases. *Behavior Research Methods*, *43*(2), 310–330. https://doi.org/10.3758/s13428-011-0071-2

Dirix, N., Cop, U., Drieghe, D., & Duyck, W. (2017). Cross-Lingual Neighborhood Effects in Generalized Lexical Decision and Natural Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(6), 887–915. https://doi.org/10.1037/xlm0000352

Dirix, N., & Duyck, W. (2017). The first- and second-language age of acquisition effect in first- and second-language book reading. *Journal of Memory and Language*, *97*, 103–

120. https://doi.org/10.1016/j.jml.2017.07.012

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: evidence for nonselective lexical access. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(4), 663–679. https://doi.org/10.1037/0278-7393.33.4.663

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review*, *15*(4), 850–855. https://doi.org/10.3758/PBR.15.4.850

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.

Gollan, T. H., Slattery, T. J., Goldenberg, D., Van Assche, E., Duyck, W., & Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: the frequency-lag hypothesis. *Journal of Experimental Psychology. General*, *140*(2), 186–209. https://doi.org/10.1037/a0022256

Hoedemaker, R. S., & Gordon, P. C. (2014). It takes time to prime: Semantic priming in the ocular lexical decision task. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 2179–2197. https://doi.org/10.1037/a0037677

Hoedemaker, R. S., & Gordon, P. C. (2017). The onset and time course of semantic priming during rapid recognition of visual words. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(5), 881–902. https://doi.org/10.1037/xhp0000377

Hudson, P. T. W., & Bergman, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language*, *24*(1), 46–58. https://doi.org/10.1016/0749-596X(85)90015-4

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153–168. https://doi.org/10.1016/j.visres.2004.07.037

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon*, *6*(1), 34–52. https://doi.org/10.1075/ml.6.1.02keu

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: a new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–50. https://doi.org/10.3758/BRM.42.3.643

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

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

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35. https://doi.org/10.1037/0096-3445.135.1.12

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology*, *63*(9), 1838–1857.

https://doi.org/10.1080/17470211003602412

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. *The Quarterly Journal of Experimental Psychology*, *66*, 563–580. https://doi.org//10.1080/17470218.2012.658820

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42–73. https://doi.org/10.1016/j.jml.2011.03.002.Effects

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

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

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological Research*, *72*(6), 675–688. https://doi.org/10.1007/s00426-008-0173-3

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

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., Pollatsek, A., Drieghe, D., Slattery, T. J., & Reichle, E. D. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert

(2006). *Journal of Experimental Psychology: General*, *136*(3), 520–529. https://doi.org/10.1037/0096-3445.136.3.520

Revelle, W. (2018). *Package 'psych'*. Available at http://personality-project.org/r/psych/psych-manual.pdf.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494. https://doi.org/10.1016/S0022-5371(70)80091-3

Schilling, H. E. H., 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. https://doi.org/10.3758/BF03201199

Schuster, S., Hawelka, S., Hutzler, F., Kronbichler, M., & Richlan, F. (2016). Words in Context: The Effects of Length, Frequency, and Predictability on Brain Responses During Natural Reading. *Cerebral Cortex*, *26*(10), 3889.2-3904. https://doi.org/10.1093/cercor/bhw184

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlation: Uses in Assessing Rater Reliabitlity. *Psychological Bulletin*, *86*(2), 420–428.

Teng, D. W., Wallot, S., & Kelty-Stephen, D. G. (2016). Single-Word Recognition Need Not Depend on Single-Word Features: Narrative Coherence Counteracts Effects of Single-Word Features that Lexical Decision Emphasizes. *Journal of Psycholinguistic Research*, *45*(6), 1451–1472. https://doi.org/10.1007/s10936-016-9416-4

van Heuven, W. J. B., Dijkstra, T., & Grainger, J. (1998). Orthographic Neighborhood Effects in Bilingual Word Recognition. *Journal of Memory and Language*, *39*(3), 458–483. https://doi.org/10.1006/jmla.1998.2584

van Heuven, W. J. B., 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. https://doi.org/10.1080/17470218.2013.850521

Vitu, F., O'Regan, J. K., & Mittau, M. (1990). Optimal landing position in reading isolated words and continuous text. *Perception & Psychophysics*, *47*(6), 583–600. https://doi.org/10.3758/BF03203111

Whitford, V., & Titone, D. (2017). Lexical entrenchment and cross-language activation: Two sides of the same coin for bilingual reading across the adult lifespan. *Bilingualism: Language and Cognition*, in press. https://doi.org/10.1017/S1366728917000554

Wochna, K. L., & Juhasz, B. J. (2013). Context length and reading novel words: An eye-movement investigation. *British Journal of Psychology*, *104*(3), 347–363. https://doi.org/10.1111/j.2044-8295.2012.02127.x