FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

Reading in a second language: An eyetracking study

Nicolas Dirix

Supervisor: Prof. Dr. Wouter Duyck
Co-supervisor: Prof. Dr. Marc Brysbaert

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Psychology

Academic year 2017–2018

GHENT
UNIVERSITY

CONTENTS

ACKNOWLEDGEMENTS

De eerste zin schrijven vond ik altijd het moeilijkste. Soms had ik al een heel manuscript klaar, maar stonden er helemaal bovenaan enkel wat kernwoorden over hoe ik mijn verhaal wilde beginnen vertellen. Bij dit onderdeel van mijn proefschrift was dit niet anders, en waren de *keywords* “veel geleerd”, “veel gedaan”, “veel bedanken”. En dat is eigenlijk ook wel de kern van wat mijn doctoraat heeft ingehouden: het is uitdagend en hard werken geweest, ik heb mezelf kunnen specialiseren in een vakgebied én ik heb er ook een steentje aan kunnen bijdragen, maar zonder de hulp en steun van collega’s, vrienden en familie had het een stuk moeilijker geweest dit tot een goed einde te brengen.

Als eerste wil ik graag mijn promotor Wouter Duyck bedanken voor zijn bereidheid dit project met mij aan te gaan en voor zijn onvoorwaardelijke steun aan mijn onderzoek. Ik kreeg steeds veel vrijheid in mijn planning, geoorloofde vertragingen door grote datasets en trage computers werden zonder problemen aanvaard, en ik kan oprecht stellen dat ik elk onderzoeksidee dat ik had tijdens de afgelopen jaren zonder schroom kon voorleggen. Een ander sterk punt waar ik veel aan heb gehad tijdens mijn doctoraat zijn jouw probleemoplossende capaciteiten: voor elk ietwat vreemd resultaat bleek tóch een logische uitleg te zijn, voor elk (te) ambitieus idee een plan van aanpak, en als er toch eens iets was waar je geen pasklaar antwoord op had, wist je aan te wijzen wie dit wel zou hebben. Het was ook fijn een promotor te hebben met een gedeelde voorliefde voor de koers en de Rode Duivels (ik weet niet zeker of ik Nainggolan al heb vergeven voor zijn slechte terugspeelbal op Courtois die we in Barcelona mochten aanschouwen in de Belchica). Mijn vaardigheden als onderzoeker zijn ongetwijfeld ook verbeterd door mijn co-promotor Marc Brysbaert. Marc, het was voor mij een eer met zo’n grote naam uit de psycholinguïstiek te mogen samenwerken. Een halfuurtje in jouw kantoor leerde me ook vaak meer bij dan dagen ploeteren in de literatuur. Ik ben nog steeds vol bewondering hoe je bij het lezen van papers of beluisteren van presentaties er steeds in slaagt om kleine, maar o zo belangrijke details voor de kwaliteit van het onderzoek weet bloot te leggen. Van onze samenwerking onthoud ik dat het belangrijk is aandachtig te blijven voor alle aspecten van het onderzoek en mee te zijn met de hedendaagse ontwikkelingen in het vakgebied.

Vanaf mijn 2^de^ jaar maakte Evy Woumans deel uit van mijn begeleidingscommissie. Beste Evy, ik kon niet alleen profiteren van jouw ervaring en goede raad voor mijn onderzoek, maar ik kon ook steeds bij jou terecht voor een werkgerelateerd of persoonlijk dipje. Dankzij jou heb ik ook van een schitterende conferentie in eigen stad mogen genieten, en je gezelschap op de trip naar Southampton maakte de reis en het verblijf heel wat aangenamer. Ik wil ook graag de overige

leden van mijn begeleidingscommissie, Evelyne Lagrou, Eva Van Assche en Johan Van Braak, bedanken voor de samenwerking en hun bijdrage aan mijn onderzoek.

Natuurlijk wil ik ook even stilstaan bij de collega’s van het Lemma project, met wie ik het nauwst heb samengewerkt tijdens mijn doctoraat:

Heleen, je hebt een ongeëvenaard aanstekelijk enthousiasme. Ik heb veel met jou kunnen delen, ook omdat ik wel wat van mezelf herken in jouw: erg gepassioneerd voor de job, maar soms ook erg gefrustreerd; de nodige dosis zelfkritiek, maar ook de onzekerheid die daar soms mee gepaard gaat. Het was fijn om samen met jou de PhD-weg te mogen ontdekken en bewandelen, maar ook om te leren het werk soms wat meer te relativeren. Ik kon steeds bij jou terecht om een idee af te checken of gewoon een babbeltje te slaan, ik hoop dat dit na afloop van ons project nog steeds het geval kan zijn.

Wouter, ik heb veel bewondering voor je werkethos en onaflatende inzet. Je bent een getalenteerde onderzoeker en de academische wereld kan er alleen maar baat bij hebben jou aan boord te houden. Ik apprecieerde altijd je korte bezoekjes als je even ging bijtanken in het koffielokaal. Op de meer informele momenten wist je me ook steeds te verrassen met je veelzijdigheid, en wie weet komt er ooit nog een moment dat we ’s avonds kunnen genieten van een tas thee (kop of mok voor de Nederlanders) en wat zoetigheid.

Aster, het was leuk om je bureaugenootje te zijn, wat kletsen tijdens rustigere momenten en samen met jou te mogen afzien in de ijskast/sauna/zonnebank die ons kantoor soms kon zijn. De afgelopen twee jaar was je er wat minder, maar dat heeft uiteraard te maken met je fijne gezinnetje in Nederland, waar ik je ook het allerbeste mee wil wensen.

Ellen, Toru en Marcelo, jullie wil ik graag bedanken voor de vele onderhoudende lemma-meetings, het uitwisselen, bespreken of uitwerken van onderzoeksideeën en natuurlijk de occasionele drinks in de Walrus of Spinnekop. Ook wil ik graag de andere promotoren van het project, Rob Hartsuiker en Martin Valcke, bedanken voor hun feedback op mijn experimenten.

Verder wil ik graag nog enkele co-auteurs en collega’s bedanken voor hun hulp de afgelopen jaren. Uschi, zonder jou had ik waarschijnlijk nooit zo veel kunnen bereiken in mijn vierjarig doctoraat. Je hebt me niet alleen begeleid in mijn eerste stappen als onderzoeker aan onze faculteit, maar je hebt ook al zowat het hele voorbereidende werk voor mijn project verricht voordat er nog maar sprake van was. Je was de perfecte mentor om samen aan mijn eerste studie te werken, maar even goed was je te vinden voor een we-sluiten-alle-bars-in-Valencia-

verjaardagsfeestje. Denis, jou wil ik ook graag bedanken voor de bijdrage aan de neighborhood studie, de uitmuntende linear mixed models cursus en de hilarische restaurantbezoekjes met Wouter en Evy in Southampton. Verder wil ik graag nog Emmanuel en Michael bedanken voor hulp bij analyses, Lies en Christophe voor de administratieve en technische zaken, en Eowyn en Elisabeth voor de gezellige babbels.

Ten slotte wil ik graag mijn familie en vrienden bedanken voor hun oprechte interesse in mijn onderzoek. Ik maak hier graag nog plaats voor enkel speciale vermeldingen:

Aan een doctoraat kunnen beginnen en het succesvol afronden heeft ook veel te maken met kansen krijgen en die kansen ten volle benutten. Er zijn geen twee mensen die ik meer kan bedanken om al die kansen mogelijk te maken dan mijn ouders. Mama en papa, zonder jullie steun doorheen mijn schoolse en academische opleiding was er van een doctoraat zelfs geen sprake geweest.

Broertjes Mathias en Andreas, die carrière maken in het échte leven, wil ik graag bedanken voor hun interesse in mijn werk, maar ook om mij de met momenten nodige ontspanning of afleiding te bezorgen en me nog wat bij te leren over zaken die niets met psycholinguïstiek te maken hebben (zoals zolders afbreken en een goede koersfietshouding aannemen).

Uiteraard is er helemaal onderaan plaats voor een speciale eervolle vermelding voor Caroline. Lieveke, ik maak me nu al vrolijk met het idee jouw bedenkelijke blik te zien wanneer je dit leest. Hoe dan ook kan ik er natuurlijk niet omheen dat jij de enige bent die alles heeft meegemaakt: de successen in het afronden van analyses of papers, maar ook de lange dagen en korte nachten die daarvoor soms nodig waren; het enthousiasme over een nieuwe, mooie, uitmuntende grafiek en de ergernis over crashende computers; de spanning om voor het eerst mijn eigen experiment voor te stellen voor een groot publiek, en het liefjesgemis in een drukke periode met veel congressen; samen de publicaties vieren, maar me ook ruimte geven na een moeilijke dag op het werk. Op één of andere manier vond je met een aan het onwaarschijnlijke grenzende vanzelfsprekendheid altijd wel de tijd en energie om te luisteren als ik iets te vertellen had over mijn onderzoek. Bedankt om er voor me te zijn, Lieveke.

Nicolas

Mei 2018

CHAPTER 1

INTRODUCTION[1]

I would like to ask the reader to imagine the last day when he or she has not read a single word. I can imagine this is a hard, if not impossible task. From the moment we learn to recognize written words, probably not a day goes by without reading, be it a newspaper article, subtitles on the television screen, a road sign, text messages, a book chapter, etc. This seemingly easy and straightforward task however entails a series of complex underlying processes (e.g., recognizing the letters, assembling them into a word while keeping in mind the order of the letters and boundaries of the word, activating and selecting the right word in the mental lexicon, combining and integrating multiple words into an understandable sentence, implementing sentences in the context of a narrative, etc.). These processes are studied in the field of psycholinguistics, and some relate directly to one of the main topics of this dissertation, namely visual word recognition.

READING RESEARCH IN PSYCHOLINGUISTICS

Multiple paradigms have been applied in studies on word recognition, one of the most popular being the lexical decision task. In this task, strings of letters are presented on a screen, and participants have to decide whether an existing word or a nonword is presented by pressing buttons. Researchers discovered that the reaction time (RT) to respond to these isolated words is influenced by their characteristics, such as their frequency, length, or age of acquisition (AoA; Butler & Hains, 1979; Hudson & Bergman, 1985; Rubenstein, Garfield, & Millikan, 1970). Although this paradigm has its advantages (e.g., the relatively easy collection and processing of data), it is not always clear which underlying processes are affected by the word characteristics, as a response component is involved and participants could for example develop a response strategy.

An alternative approach to investigate reading processes which has gained in popularity in the last decades is eye-tracking, where the eye movements of participants are recorded while they read (single) sentences or larger chunks of text. This type of reading is also known as “natural reading” in psycholinguistic research, as reading for meaning is central, without specific task demands or instructions other than to read the presented text. An advantage of this paradigm is that multiple measures can be investigated, such as *fixations* (when the eyes stay still on a word), *saccades* (the jumps the eyes make from one word or piece of text to another),

---

[1] Partly based on based on Dirix N., Van de Putte, E., & Duyck, W. (in press). Bilingual Lexical Access. In Heredia, R. R. & Cieślicka, A. B. (Eds). *The Bilingual Brain, Unwrapped.* New York, NY: Psychology press.

skips (when a word is not fixated during the first passage) and *regressions* (a saccade that goes backwards in the text). The different measures are believed to reflect different stages in the reading process, such as *early* word identification stages or *late* comprehension or integration stages (Boston, Hale, Kliegl, Patil, & Vasishth, 2008). The application of this paradigm in reading research also resulted in the development of models of eye movement control, such as the E-Z reader model (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Pollatsek, & Rayner, 2006). This models states that there are two serial stages in the lexical processing of words. The first stage is the *familiarity check*, in which the lexical candidates are activated when a word is fixated. As soon as a word is fixated, the oculo-motor system also initiates the programming of an intra-word saccade. If the familiarity check is completed before the programming of the intra-word saccade, this saccade is cancelled. The second stage is the *verification stage*, in which the lexical identification of the target word is completed. After this stage, attention is shifted towards the next word. As the programming of a saccade takes more time than the attention shift, parafoveal processing of the next word in the sentence is possible. Moreover, if the next word is processed fast enough, it may be skipped entirely. According to this model, the time it takes to complete these stages can be influenced by word characteristics (such as word frequency; e.g., Miellet, Sparrow, & Sereno, 2007; Rayner & Duffy, 1986), but also the predictability of the word as derived from the context (e.g., Rayner, Reichle, Stroud, Williams, & Pollatsek, 2006). This points out a crucial difference between isolated word presentation, as in the lexical decision task, and natural reading studies: the presence and potential influence of surrounding words or a narrative context. An important question is whether results from lexical decision task also generalize to natural reading, where context effects could modulate word characteristic effects.

In light of this issue, Cop, Dirix, Drieghe, and Duyck (2017) set up a unique dataset which contains eye movement data of participants reading an entire novel: The Ghent Eyetracking Corpus (GECO). Part of the participants were English monolinguals; part were Dutch-English bilinguals, who read half of the novel in their first language (L1), the other half in their second language (L2). The GECO data was used for several studies in this dissertation to investigate word characteristic effects in and processes of (bilingual) reading. GECO has also previously been applied in investigations, for example in a global comparison of reading patterns of mono- and bilingual reading (Cop, Drieghe, & Duyck, 2015) and in a study of the word frequency effect (Cop, Keuleers, Drieghe, & Duyck, 2015).

In CHAPTER 2 we used data of the monolingual part of GECO to investigate the AoA effect in natural reading. Previous research has shown that earlier learned words are processed faster than words learned at a later age (e.g., Gerhand & Barry, 1999, for lexical decision;

Juhasz & Rayner, 2006, for single sentence reading). In our study, we investigated whether this particular word characteristic would also influence the eye movement pattern of book reading, and if so, whether the effect would persist throughout all the word recognition stages. The application with this big dataset also entails the benefit of having a large amount of stimuli and range of word characteristics (which could have advantages over small-scaled, factorial experiments, see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004).

BILINGUAL READING

As mentioned before, part of the GECO data (Cop, Dirix, Drieghe, et al., 2017) was gathered from bilingual participants. The study of L2 reading has become an essential sub-domain of psycholinguistic research, which is justified, as for example the European Union estimated that about 54% of the people in the EU are bi- or multilingual (European Union & European Commission for Education and Culture, 2012). This percentage is not expected to drop, as two of the main determinants are migration and the increase in foreign language in education. It thus seems important to investigate whether reading processes in L2 are comparable to those of L1 reading, as it affects a significant part of the population. Cop, Drieghe, et al. (2015) compared sentence reading patterns of the L1 and L2 part of GECO, and found a somewhat impaired L2 processing: longer fixations, more fixations, less skips and smaller saccades in L2 compared to L1 reading (interestingly, they found that the L1 of the bilinguals did not differ from monolingual reading). Cop, Drieghe, et al. related this L2 disadvantage to the stages of the E-Z reader model (Reichle et al., 2006): they for example suggested that the familiarity check is slower in L2, resulting in longer fixation times, but also more intra-word saccades, which in turn leads to smaller saccades and more fixations.

We revisited the AoA effect in L2 reading in CHAPTER 3. By learning L2 words, an interesting situation occurs: a bilingual learns a new word form for a concept that he or she already learned in L1, with the result that each L2 word has an L2 AoA for its word form, but also an L1 AoA for the underlying concept. By comparing AoA effects in L1 and L2 word processing, it should be possible to assess whether it is the order of learning the word forms (cf. the *mapping* hypothesis, Ellis & Lambon Ralph, 2000), or the organization of the semantic network (cf. the *semantic hypothesis*, Brysbaert, Van Wijnendaele, & De Deyne, 2000) that drives the AoA effect. The mapping hypothesis would predict within-language AoA effects, as the overall order of learning is most important, whereas the semantic hypothesis would predict an effect of L1 AoA on L2 reading, as the order of learning the semantic concepts in L1 is crucial. This was addressed by Izura and Ellis (2002) in a series of lexical decision tasks, who found results in favor of the mapping hypothesis. We decided to study this effect by using GECO data as the first big data study of the L2 AoA effect, but also to have the opportunity to study the influence of L1 and L2 AoA in a more detailed manner by investigating multiple eye-tracking measures. This study also contains an L2 AoA rating experiment, as we needed to gather these ratings ourselves.

Another intriguing topic in bilingual reading is lexical access. Two main hypotheses have been proposed in regard to this issue. The first one assumes that lexical representations of words are stored separate for different languages, meaning that only lexical candidates in the relevant language would be activated. This hypothesis is known as *language selective lexical access*. In contrast, the *language non-selective lexical access* hypothesis proposes that lexical representations in both languages are initially co-activated, also leaving open the possibility of cross-lingual influences. A lot of evidence for language non-selective access comes from studies with cognates. These are words that have the same meaning across languages; they can either be identical (e.g., *ring*) or non-identical (e.g., *boat* – *boot* in English and Dutch) in orthography and/or phonology. Bultena, Dijkstra, and van Hell (2013) found that participants responded faster to cognates than control words in an L2 lexical decision task, showing that the currently irrelevant L1 is co-activated while reading these cognates. In a study with Dutch – English – German trilinguals, the cognate effect was even shown to accumulate with the knowledge of an additional language (Lemhöfer, Dijkstra, & Michel, 2004): responses were fastest to cognates that co-existed in all three languages (e.g., *wind*). Likewise, Van Hell and Dijkstra (2002) tested Dutch – English – French trilinguals on an L1 lexical decision tasks. In both their experiments, they found that Dutch – English cognates were responded to faster than control words, providing evidence that the (non-target) L2 can even be co-activated while performing a task in L1, whereas usually influences of the non-target language are mostly (or only) found in L2. This because L1 is usually the dominant language and is therefore less susceptible to influences of the inferior L2 (Dijkstra & van Heuven, 2002).

A second research line that provides evidence for language non-selective access involves orthographic neighbors. Neighbors are words that resemble each other closely, differing only in one letter. Words can have both intra-lingual (e.g., *purse* and *nurse*) and cross-lingual neighbors (e.g., *purse* and *puree*, mash in Dutch). In a generalized lexical decision (with both L1 and L2 stimuli in the same experiment), van Heuven, Dijkstra and Grainger (1998) found that Dutch – English bilinguals responded slower to English target words with many Dutch neighbors, compared to a small amount of cross-lingual neighbors. Again, it seems that lexical representations of the non-target language are co-activated when words are recognized.

Next to studies with isolated words, there is also evidence for non-selective lexical access in eye-tracking experiments with single sentences or a narrative. This is important, as a sentence is usually presented unilingually, providing a strong cue to which language the words in the sentence belong. A word that is language ambiguous (e.g., a cognate or an inter-lingual homograph) could be discarded of cross-lingual influences when the context is unilingual. Duyck, Van Assche, Drieghe, and Hartsuiker (2007) conducted an eye-tracking experiment with L2 stimuli (identical cognates, non-identical cognates and control words) embedded in sentences. They found a cognate facilitation effect in early and late eye movement measures for the identical cognates, showing that even in the eye movement patterns of reading L2 sentences, and advantage for cognates is present, further supporting language non-selective access. Cognate effects were also studies in the bilingual GECO data by Cop, Dirix, Van Assche, Drieghe, and Duyck (2017). In L2 reading, cognate facilitation was found only on late measures; in L1 reading identical cognate facilitation was present to a smaller extent: only high frequent cognates showed facilitation, again in late measures, showing that even when bilinguals are reading a novel in a strict unilingual context, without other instructions than just reading, the non-target language is co-activated.

In CHAPTER 4, we further investigated cross-lingual neighborhood effects. As the status of cognates is still debated (i.e., do cognates have a representation for each language, or a single representation that is shared among languages?), a cross-lingual neighborhood influence would be more conservative evidence of language non-selective lexical access. In Experiment 1, we attempted to replicate the generalized lexical decision task of van Heuven et al. (1998), which is one of the only studies reporting such an effect. Furthermore, we investigated the effects of cross-lingual neighbors in the bilingual part of GECO to see whether a potential effect of the non-target language could survive the strong unilingual context.

STUDYING IN L2

Studies of bilingual reading found that the L2 reading process is somewhat impaired in comparison to L1 reading (e.g., Cop, Drieghe, et al., 2015). There also seems to be an increase in the popularity of education in L2 (European Union & European Commission for Education and Culture, 2012), for example English as a Medium of Instruction (EMI) in higher education. This poses the question what the impact of education in L2 has on the study process, as for examples a lot of handbooks in higher education are international, English editions.

In CHAPTER 5, we therefore investigated the eye movement patterns of participants reading for different goals (informative reading and studying for a test) in L1 and L2. There were only a few eye-tracking studies that compared the eye movement patterns of different reading goals (e.g., Yeari, van den Broek, & Oudega, 2015), but this was the first study to investigate this in an L2 context. Furthermore, we also wanted to assess the impact of studying in L2 on academic achievement. In related research, Vander Beken and Brysbaert (2017) investigated memory for texts in L1 and L2. They found a recall cost when participants had to study and recall the content of the texts in L2 in comparison to L1, but identical scores on L1 and L2 recognition tests with true/false statements, which suggest that a superficial encoding or access to the memory trace seems to be equal in L1 and L2. We included recognition tests for the content of the texts to see whether we could replicate the findings of Vander Beken and Brysbaert. Additionally, we were interested whether we could predict test scores by reading times, or in other words, whether longer fixation times on a piece of text results in higher scores for questions related to that piece of text.

CORRESPONDENCE BETWEEN READING MEASURES

As mentioned at the beginning of the introduction, two of the most popular paradigms in visual word recognition are lexical decision and eye tracking. Although the results of studies applying these paradigms often point toward similar effects of word characteristics (e.g., facilitating AoA effects in lexical decision, Gerhand & Barry, 1999 and eye tracking, Juhasz & Rayner, 2006), this is not always the case (to anticipate, in CHAPTER 4 we report quite different orthographic neighborhood effects in lexical decision and eye movement data). Next, there is also the difference in the nature of task (for example, the direct measure of fixation times on words vs the RTs measured by the speed of the button press; the presence of surrounding words and language context in sentence or paragraph reading; etc.), which poses the question how well these measures correspond. Kuperman, Drieghe, Keuleers, and Brysbaert (2013) addressed this issue by correlating lexical decision RTs and timed eye movement measures. They found a surprisingly low shared amount of variance between these measures, suggesting that context indeed is of critical importance, and it might even be questioned whether these two paradigms truly tap into the same underlying processes. In CHAPTER 6, we extended the findings of Kuperman et al. by further assessing the convergence between lexical decision and eye-tracking paradigms in different languages with data from recent lexical decision mega studies and eye-tracking corpora. We also investigated the impact of context effects on eye-tracking measures by correlating reading times for identical words in different corpora, and by comparing the fixation times of a single presentation vs. repeated presentations. Finally, we calculated the reliabilities for the lexical decision RTs and various timed eye-tracking measures, and compared the impact of two important predictors in these datasets (word frequency and length).

As L2 reading seems to differ from L1 reading, it did not seem implausible that context effects would have a different influence the eye movement pattern (see for example Gollan et al., 2011). In CHAPTER 6 we therefore also calculated the correlations between L2 lexical decision RTs and L2 eye-tracking reading times and compared them to the L1 results. If correlations for the L2 data would be higher than those of the L1 data, this would suggest that

L2 reading is more affected by individual word characteristics and less by the context than L1 reading; and vice versa if lower correlations for L2 would turn up.

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

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., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*, 215–226.

Bultena, S., Dijkstra, T., & van Hell, J. G. (2013). Cognate and word class ambiguity effects in noun and verb processing. *Language and Cognitive Processes*, *28*(9), 1350–1377. https://doi.org/10.1080/01690965.2012.718353

Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. *Memory & Cognition*, *7*(2), 68–76. https://doi.org/10.3758/BF03197587

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., Dirix, N., Van Assche, E., Drieghe, D., & Duyck, W. (2017). Reading a book in one or two languages? An eye movement study of cognate facilitation in L1 and L2 reading. *Bilingualism: Language and Cognition*, *20*(4), 747–769. https://doi.org/10.1017/S1366728916000213

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

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. https://doi.org/10.1017/S1366728902273010

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

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103–1123. https://doi.org/10.1037/0278-7393.26.5.I103

European Union & European Commission for Education and Culture. (2012). Europeans and their languages.Retrieved from http://ec. europa.eu/public_opinion/archives

Gerhand, S., & Barry, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, *27*(4), 592–602. https://doi.org/10.3758/BF03211553

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

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

Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica*, *23*, 245–281.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7–8), 846–863. https://doi.org/10.1080/13506280544000075

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

Lemhöfer, K., Dijkstra, T., & Michel, M. (2004). Three languages, one ECHO:Cognate effects in trilingual word recognition. *Language and Cognitive Processes*, *19*(5), 585–611. https://doi.org/10.1080/01690960444000007

Miellet, S., Sparrow, L., & Sereno, S. C. (2007). Word frequency and predictability effects in reading French: An evaluation of the E-Z Reader model. *Psychonomic Bulletin and Review*, *14*(4), 762–769. https://doi.org/10.3758/BF03196834

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., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465. https://doi.org/10.1037/0882-7974.21.3.448

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157. https://doi.org/10.1037/0033-295X.105.1.125

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4–22. https://doi.org/10.1016/j.cogsys.2005.07.002

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

Van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review*, *9*(4), 780–789. https://doi.org/10.3758/BF03196335

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

Vander Beken, H., & Brysbaert, M. (2017). Studying texts in a second language: the importance of test type. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728917000189

Yeari, M., van den Broek, P., & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: evidence from eye-movements. *Reading and Writing*, *28*(8), 1071–1097. https://doi.org/10.1007/s11145-015-9561-4

CHAPTER 2

AN EYE MOVEMENT CORPUS STUDY OF THE AGE OF ACQUISITION EFFECT.[1]

The current study investigated the effects of word-level age of acquisition on natural reading. Previous studies, using multiple language modalities, found that earlier learned words are recognized, read, spoken and responded to faster than words learned later in life. Until now, in visual word recognition, experimental materials were limited to single word or sentence studies. We analyzed data of the Ghent Eye-tracking COrpus (GECO; Cop, Dirix, Drieghe, & Duyck, in press), an eye-tracking corpus of participants reading an entire novel, resulting in the first eye movement megastudy of AoA effects in natural reading. We found that the age at which specific words are learned indeed influences reading times, above other important (correlated) lexical variables such as word frequency and length. Shorter fixations for earlier learned words were consistently found throughout the reading process in early (single fixation durations, first fixation durations, gaze durations) and late measures (total reading times). Implications for theoretical accounts of AoA effects and eye movements are discussed.

[1] Dirix, N., & Duyck, W. (2017). An eye movement corpus study of the age of acquisition effect. *Psychonomic Bulletin & Review, 24*(6), 1915-1921. doi: 10.3758/s13423-017-1233-8.

INTRODUCTION

Carroll and White (1973) first discovered that the age at which we learn words influences their processing speed, independent from other language processing determinants. They found shorter latencies for picture naming when words had an earlier age of acquisition (AoA). Since then, AoA effects have been reported in various tasks and language modalities: picture naming (e.g., Belke, Brysbaert, Meyer, & Ghyselinck, 2005), word naming (e.g., Gerhand & Barry, 1999b), masked priming (e.g., Brysbaert, Lange, & Van Wijnendaele, 2000), semantic categorization (e.g., Brysbaert, Van Wijnendaele, & De Deyne, 2000) and lexical decision (e.g., Gerhand & Barry, 1999a). For reviews, see Johnston & Barry (2006) or Juhasz (2005).

Age of Acquisition Hypotheses

Two hypotheses try to explain the mechanism behind the AoA effect. The *semantic hypothesis* claims that AoA effects do not primarily originate from learning lexical word forms, but from their semantic representations. AoA effects then reflect the speed by which these are accessed, as a function of the organization of the representational network (Brysbaert, Van Wijnendaele, et al., 2000; Steyvers & Tenenbaum, 2005). When new concepts are learned, they are linked to the ones already in the network. Early learned words will be more central and better connected in the network, making them more easily accessible. Evidence for this hypothesis comes from the observation that AoA effects become larger when semantic activation of stimuli is necessary; i.e. they are larger in object naming tasks than in lexical decision (Barry, Johnston, & Wood, 2006), and larger in lexical decision than in word naming (Cortese & Khanna, 2007). More direct evidence comes from semantic categorization tasks where AoA effects were found (Brysbaert, Van Wijnendaele, et al., 2000), and from a semantic Simon task (Ghyselinck, Custers, & Brysbaert, 2004). In this last paradigm, participants judged whether words were presented in upper- or lowercase by responding verbally with labels that could be semantically congruent or incongruent with the (irrelevant) meaning of the target ("living" and "nonliving"). The semantic congruency effect was stronger for early acquired words, showing that the meaning of the early learned words was activated faster. The authors conclude that semantics play an important role in the AoA effect.

The second hypothesis is the *mapping* or *connectionist* hypothesis. It originates from simulations with connectionist networks (Ellis & Lambon Ralph, 2000; Monaghan & Ellis, 2010): items that were trained first always had an advantage over later trained items because the early items are learned better. The researchers argue that information which enters a network first, benefits more from the plasticity of the network and alters its connections, or weights, to a stronger extent. As new information keeps on entering the network, the network loses plasticity, making weight changes smaller. Early items thus have a larger impact on the networks final structure. In contrast to the semantic hypothesis, the mapping hypothesis does not situate AoA effects on a single processing level. It could play at the lexical, semantic and/or phonological level. Evidence for this hypothesis comes from tasks where it is shown that learning completely new information (e.g., nonwords, complex patterns, etc.) in several stages results in an order of acquisition (OoA) effect, analogous to the AoA effect (Joseph, Wonnacott, Forbes, & Nation, 2014; Stewart & Ellis, 2008).

Age of Acquisition in Eye-tracking

AoA effects emerge across language modalities, and are therefore also of interest for visual word recognition research, which often use lexical decision tasks (Brysbaert, Lange, et al., 2000). Next to these studies with single word presentations, also eye-tracking has been used to investigate AoA effects in a few rare sentence reading studies (Joseph et al., 2014; Juhasz & Rayner, 2003, 2006). This is highly relevant, given that most words are encountered in a sentence context. It is therefore important to generalize findings from experimental, isolated word recognition, to natural language processing.

One of the advantages of investigating eye movements is that they can be monitored with high spatial and temporal resolution. They reveal large amounts of information on underlying word recognition processes (Rayner, 1998, 2009). Also, multiple dependent variables are available in eye-tracking. Single fixations are the durations of the fixation of words that were fixated only once. First fixations are the durations of the first fixation of words, regardless of later refixations. Gaze durations are the sum of all fixation durations during the first passage before the eyes focus of the word. These measures are “early” measures of eye-tracking because they reflect initial stages of word recognition. Finally, total reading times are a “late” measure of eye-tracking, as they constitute the sum of all fixations on the target word including refixations. As participants only have to read the presented text, another advantage of eye-tracking is the minimal amount of interference by task demands, in contrast to for example the lexical decision, which includes a decision component that may introduce strategic biases. Eye-tracking does therefore seem to be a promising technique to investigate AoA effects in visual word recognition.

Juhasz and Rayner (2003, 2006) found that the AoA of target words influenced reading times in eyetracking: earlier AoAs lead to shorter fixations. In the 2003 study, this was found in early measures (single fixation and gaze duration); in the 2006 study also in an additional early (first fixation) and late measure (total reading time). The authors argue that this difference is due to the design of the studies: in the 2006 study an orthogonal design with early and late AoA values was applied; in the 2003 study, AoA was treated as a continuous variable. The effects were more pronounced when only extreme AoA values were presented. As both studies presented the target stimuli in sentences, and because semantic activation (i.e., the meaning) of these words is necessary to understand the sentence, Juhasz and Rayner interpreted their results as evidence for the semantic hypothesis.

These pioneering eye-tracking studies on AoA effects are very informative and now require assessments of their generalizability. First, the total amount of target sentences (and words) tested by Juhasz and Rayner (2003, 2006) was limited to respectively 72 and 108. This is typical for an eye-tracking paradigm, but rather small compared to the megastudy approach that we adopted here. Second, the researchers operationalize “natural reading”, their extension of isolated word recognition, as single sentence reading, whereas in daily life we also tend to read longer chunks of text that make a coherent whole. Finally, although the 2003 study with a continuous AoA yielded significant effects, their most convincing results of AoA effects come from orthogonal designs. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) argued that a factorial approach could entail several flaws, such as implicit biases of experimenters and participants, and a reduction of power and reliability when continuous variables are converted to categorical ones. They propose a megastudy approach as a valuable alternative, with large samples of stimuli varying on a broad range of characteristics. For isolated word recognition, this approach has been successfully applied in two studies (Cortese & Khanna, 2007; Cortese & Schock, 2012) that assessed AoA effects in lexical decision data of the English Lexicon Project (ELP, Balota et al., 2007). Both studies found an AoA effect (faster reaction times for earlier AoA) above and beyond other predictors such as word frequency and length. In compliance with these studies, we assessed AoA effects using megastudy data of natural story reading.

Current Study

We investigated AoA effects in the Ghent Eye-tracking COrpus (GECO; Cop et al., in press). This corpus is an eye-tracking database of participants reading an entire novel. GECO has previously successfully been used to investigate for example effects of word frequency (Cop, Keuleers, Drieghe, & Duyck, 2015) and orthographic neighborhood (Dirix, Cop, Drieghe, & Duyck, in press). Here, we used the corpus to investigate the importance of AoA, in addition to other lexical variables, when participants are reading a large body of text, rather than single words or sentences. The corpus contains a monolingual (English) and a bilingual (Dutch and English) part. For the current study we focused on the monolingual data as we wanted to investigate the AoA effect without potential influences of second language knowledge. The monolingual dataset contains about 760 000 words read in total: 14 participants read 54 364 words (5012 unique), embedded in 5 300 sentences. This dataset provides a large variety in target words and a broad range of word characteristics.

We analyzed both early (single fixation, first fixation and gaze duration) and late (total reading time) measures of eye-tracking. The AoA ratings for our stimuli were taken from the database of Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). Such ratings are commonly used in AoA experiments and score well on validity (Brysbaert, in press). Next to AoA, we included other (sometimes correlated) important word recognition predictors in the analysis: word frequency (SUBTLEX-UK; van Heuven, Mandera, Keuleers, & Brysbaert, 2013), length and neighborhood density (CLEARPOND; Marian, Bartolotti, Chabal, & Shook, 2012). Several target words were presented more than once throughout the novel, so we included the predictor “rank of occurrence” to account for repetition effects.

We expected that reading times, on all measures, would be shorter for earlier learned words, in accordance with Juhasz and Rayner (2003, 2006). We did not apply an orthogonal design, but we included interactions between the predictors in the base models. This allowed the interaction of AoA with word frequency, as in Gerhand and Barry (1999a). They found that the AoA effect was larger for low frequent words.

METHOD

Participants and Materials

The stimuli and data of this study were taken from the monolingual GECO part (Cop et al., in press), in which participants read the entire novel “The mysterious affair at Styles” by Agatha Christie. We included all nouns for which an AoA rating was available in Kuperman et al. (2012), but only if at least 75% of the raters made an AoA estimation (to ensure a reliable AoA rating). 7158 nouns (1487 unique) remained in the final selection (see Table 1).

The monolingual participants were 14 undergraduate students at the university of Southampton (8 females, M_{age} = 21.8, SD_{age} = 5.6). Their language proficiency was tested with the LexTALE (Lemhöfer & Broersma, 2012; M = 91.07, SD = 8.92, range = [71.25 – 100]).

Table 1 *Descriptive Statistics for the nouns of the monolingual part of GECO used in the current study, averaged over stimuli (standard deviations between parentheses).*

Word Frequency[a]	Word Length	AoA[b]	Neighborhood Density[c]	Rank of Occurrence
3.99 (0.90)	5.85 (2.23)	6.42 (2.47)	4.75 (5.68)	13.40 (19.71)

[a]Log10 Subtlex frequencies from SUBTLEX-UK (van Heuven et al., 2013); [b]Age of Acquisition of the English words (Kuperman et al., 2012); [c]Total neighborhood densities from CLEARPOND (Marian et al., 2012).

Procedure

Eye movements of the participants were monitored while they read the novel in four separate sessions. The number of chapters was fixed for each session, but the reading tempo within the sessions was self-paced. To ensure that participants were reading for comprehension, multiple choice questions were presented after each chapter. For a detailed overview of the procedure, see Cop et al. (in press).

Eye movement analysis

Each dependent variable was fitted in a linear mixed model using the lme4 package (version 1.1-10) in R (version 3.1.1; R Core Team, 2014). P-values were calculated with lmerTest (2.0-30). Initial models included fixed factors AoA, Word Frequency, Word Length, Neighborhood Density, Language Proficiency and Rank of Occurrence (all continuous), and random intercepts for subjects and words. The random intercepts for subjects were included to ensure that individual differences in genetic, developmental or social factors between subjects were modeled (Baayen, Davidson, & Bates, 2008). The random intercept for words were included to be able to generalize to other nouns, as the current stimuli set is not an exhaustive list of all English nouns. Word Frequency was log transformed with base 10 to normalize its distribution. All continuous variables were centered.

Each dependent variable was also log transformed with base 10. The following procedure was applied to discover the optimal model (Barr, Levy, Scheepers, & Tily, 2013): first a full model including all interactions between the fixed effects (up to three-way) was fitted. Then, the model was backward fitted by excluding the interaction with the smallest t-value. An interaction term was excluded if a model comparison Chi-square test turned out to be not significant, meaning that it did not contribute to the fit. Next, the random effects were forward fitted. They were kept in the model if they contributed to the fit. Finally, the fixed effects were again backward fitted.

RESULTS

The average fixation times are presented in Table 2. We median split the data by AoA and word frequency, just to give an indication of the effect sizes of these crucial predictors. The descriptive statistics indicate that their independent effects are comparable in size.

Outliers were determined as fixation times more than 2.5SD away from the subject means and were removed from the dataset (2.16% for single fixation, 2.37% for gaze duration, 2.80% for total reading time). All final models are presented in Table 3. See Appendix 2A for the first fixation analysis.

Table 2. *Average single fixation duration, first fixation duration, gaze duration and total reading time for early [2.4-7.8] and late [7.9-19] AoA and low [0.01-3.44] and high [3.45-5.85] word frequency, in ms.*

	Age of Acquisition			Word Frequency		
	Early	Late	Effect	Low	High	Effect
Single fixation duration	216	226	10	226	217	9
First fixation duration	218	232	14	232	219	13
Gaze Duration	234	255	21	256	234	22
Total Reading Time	265	301	36	303	266	37

Single Fixation Duration

Only nouns that received a single fixation were selected for this analysis (56.35%). There was a main effect of AoA: single fixations were shorter for words with an earlier AoA. The main effects of word frequency and word length were significant, as was their interaction. Single fixations were shorter for more frequent words, but only for nouns of 4 or more letters ($\chi = 6.17$, df = 1 $p < .05$). The interaction between word length and language proficiency was also significant. Fixations became longer with increasing word length, but this effect diminished for participants who scored 92.65 or higher on the LexTALE ($\chi = 3.84$, df = 1 $p < .05$).

Gaze Duration

The main effect of AoA was significant: gaze durations were shorter for earlier learned words. The main effects of word frequency and word length were significant, as was their interaction. Gaze durations were shorter for higher frequent nouns; post hoc contrasts showed that the effect was significant for even the shortest words (3 letters, $\chi = 5.27$, df = 1 p < .05) but it became larger as word length increased.

Table 3. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final linear mixed effect model for the dependent measures.*

	Single Fixation Duration					Gaze Duration					Total Reading Time				
	β	SE	*t*	*p*		β	SE	*t*	*p*		β	SE	*t*	*p*	
Fixed Effects															
Intercept	2.320	0.014	160.806	<.001	***	2.343	0.017	141.872	<.001	***	2.388	0.019	124.878	<.001	***
Age of Acquisition	0.002	0.001	4.272	<.001	***	0.002	0.001	3.632	<.001	***	0.003	0.001	3.178	.002	**
Word Frequency	-0.008	0.002	-4.305	<.001	***	-0.011	0.002	-4.821	<.001	***	-0.013	0.003	-4.478	<.001	***
Word Length	0.002	0.001	2.320	.025	*	0.006	0.001	4.502	<.001	***	0.009	0.001	7.597	<.001	***
Neighborhood Density	<-0.001	<0.001	-0.693	.489		<-0.001	<0.001	-0.048	.962		<0.001	<0.001	0.578	.563	
Language Proficiency	-0.001	0.002	-0.696	.499		<-0.001	0.002	-0.177	.863		<0.001	0.002	0.221	.829	
Rank of Occurrence	<-0.001	<0.001	-0.810	.418		<-0.001	<0.001	-0.644	.520		<-0.001	<0.001	-3.233	.001	**
AoA * Word Frequency	/	/	/	/		/	/	/	/		-0.002	0.001	-3.944	<.001	***
Word Frequency * Word Length	-0.001	<0.001	-2.957	.003	**	-0.001	<0.001	-2.640	.008	**	/	/	/	/	
Word Length * Language Proficiency	<-0.001	<0.001	-2.651	.018	*	/	/	/	/		/	/	/	/	
	Variance	SD				Variance	SD				Variance	SD			
Random Effects															
Word															
(Intercept)	<0.001	0.018				0.001	0.023				0.001	0.033			
Subject															
(Intercept)	0.003	0.054				0.004	0.062				0.005	0.071			
Age of Acquisition	<0.001	0.001				<0.001	0.001				<0.001	0.001			
Word Frequency	<0.001	0.005				<0.001	0.006				<0.001	0.008			
Word Length	<0.001	0.002				<0.001	0.004				<0.001	0.004			

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Total Reading Time

The main effects of AoA and word frequency were significant, as was their interaction (see Figure 1): Total reading times were faster for an earlier AoA reading times, but only for words with a word frequency up to 4.290 ($\chi = 3.86$, df = 1 $p < .05$). The main effects of word length and rank of occurrence were significant. Reading times were slower with increasing word length, but faster for repeated presentations of a noun.

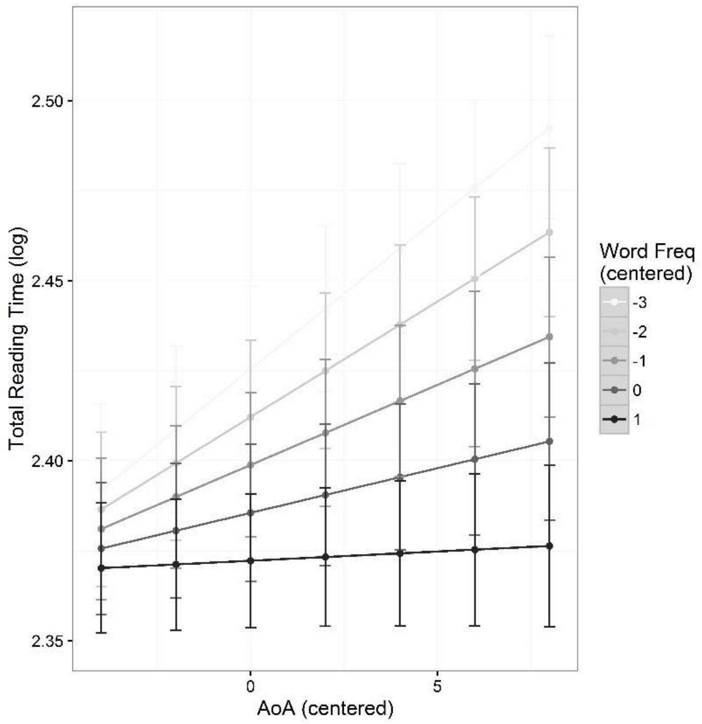

Figure 1. The interaction between AoA (x-axis) and Word Frequency (lines) in Total Reading Times (Y-axis)

DISCUSSION

We investigated AoA effects in the monolingual data of an eye-tracking corpus (GECO; Cop et al., in press). In accordance with a few rare earlier eye-tracking investigations (Juhasz & Rayner, 2003, 2006), we expected faster reading times for earlier learned words. And indeed, we found that AoA had the expected effect on reading times for all four dependent eyetracking measures: earlier learned words were read faster, independent of other lexical variables. Furthermore, we hypothesized that word frequency and AoA could interact. For total reading times, this interaction was indeed significant and in line with previous results (Gerhand and Barry 1999a): the AoA effect was larger for low frequent words.

This study was the first to investigate AoA effects in natural reading. Our results show that the age at which we learn words does not only influence the reading process when encountering single words (e.g., Brysbaert, Lange, et al., 2000) or sentences (Juhasz & Rayner, 2003, 2006), but even when while reading longer pieces of coherent text. The results are also in line with other megastudy investigations of AoA effects on isolated word recognition (e.g., Cortese & Khanna, 2007).

Following the reasoning of Juhasz & Rayner (2003, 2006), semantic activation is needed to understand words embedded in sentences, and AoA effects emerged during such reading. AoA effects were found in measures such as single fixations (where the word is read and recognized on a single fixation) and total reading times, for which we assume that semantic activation of the word is then completed. Indeed, the current results could be considered evidence for the semantic hypothesis (Brysbaert, Van Wijnendaele, et al., 2000), where the semantic network organization plays a central role in AoA effects. However, the current results could also be framed in the mapping hypothesis (Ellis & Lambon Ralph, 2000). This hypothesis does not specify which processing level AoA influences, but handles a "first-come, first-served" principle: network weights are altered in favor of items that entered the network earlier. We also observed AoA effects on measures where semantic access of words is not yet assumed to be complete (i.e., first fixation and gaze duration).

Furthermore, this hypothesis predicts that AoA effects should be the strongest in tasks where input-output mappings are arbitrary, such as in picture naming, where there is no systematic mapping between the meaning of the picture and the phonology of the word it represents. On the other hand, AoA effects should be smaller in tasks where input-output mappings are consistent, like in word naming tasks that usually have a reasonably consistent relationship between the orthography and phonology of a word. Evidence for this prediction was provided both in a computational and an experimental study by Lambon Ralph and Ehsan (2006), where the AoA effect was indeed larger for arbitrary mappings than for systematic mappings. In the current study, we found a significant AoA effect in all timed measures of reading, but the averages in Table 3 indicate that the effect is smaller in early measures (which are supposed to reflect early word recognition) than in late measures (which involve semantic processing of the words and thus rely on the arbitrary orthography – semantic mappings). In addition, the mapping hypothesis predicts AoA effects to be present in opaque languages (with arbitrary orthography to phonology mappings). As English is considered an opaque language, our current results are also in line with this prediction.

A third option is that the AoA effect originates from systems that occur in both the semantic and mapping hypotheses, as they are not mutually exclusive. Indeed, whereas the mapping hypothesis describes a functional mechanism, the semantic hypothesis provides a structural explanation. In the data, early learned words have an overall advantage over later learned words, even in early word recognition stages. This can be explained by the mapping hypothesis. However, the meaning of early learned words is also activated faster, possibly because they have a more central place in the lexicon. As our data points towards evidence for both hypotheses, it is likely that they both have a share in the etiology of the AoA effect.

Next to theoretical accounts of the AoA effect, these results are also of importance to eye movement models. An example is the E-Z reader model (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Pollatsek, & Rayner, 2006). According to this model, lexical processing of words occurs in two serial stages. In the *familiarity check*, lexical candidates become active. After completion of this stage, the oculo-motor system starts programming a saccade towards the next word. In the *verification stage,* full lexical identification of the target word is accomplished. After the completion of this stage, attention is shifted towards the next word. This model thus decouples saccade programming from the attention shift. The determining factors for the duration of the two stages are assumed to be word frequency and predictability of the target. However, the current results suggest that also AoA determines the duration of fixations. For example, the familiarity check might be faster for words that are more easily accessible because they have a more central place in the network (semantic hypothesis) or because the network weights are shifted in their advantage (mapping hypothesis), leading to shorter fixations. Future versions of E-Z reader could introduce AoA as a determining factor for fixation times, hereby possibly increasing the explained variance in observed reading times.

In conclusion, we found clear AoA effects in the eye-tracking patterns of monolinguals reading an entire novel, independent and above the influence of other lexical variables. These results generalize the large body of evidence that finds that earlier learned words are processed faster, to natural reading of running text.

REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005

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. doi:10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project.

Behavior Research Methods, *39*(3), 445–459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Barry, C., Johnston, R. A., & Wood, R. F. (2006). Effects of age of acquisition, age, and repetition priming on object naming. *Visual Cognition*, *13*(7-8), 911–927. doi:10.1080/13506280544000101

Belke, E., Brysbaert, M., Meyer, A. S., & Ghyselinck, M. (2005). Age of acquisition effects in picture naming: evidence for a lexical-semantic competition hypothesis. *Cognition*, *96*, B45–54. doi:10.1016/j.cognition.2004.11.006

Brysbaert, M. (in press). Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings “corrected” for frequency. *Quarterly Journal of Experimental Psychology*.

Brysbaert, M., Lange, M., & Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*(1), 65–86.

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*, 215–226.

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, *25*(1), 85–95. doi:10.1080/14640747308400325

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

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (in press). Presenting GECO: An Eye-tracking Corpus of Monolingual and Bilingual Sentence Reading. *Behavioral Research Methods*

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *The Quarterly Journal of Experimental Psychology*, *60*(8), 1072–1082. doi:10.1080/17470210701315467

Cortese, M. J., & Schock, J. (2012). Imageability and age of acquisition effects in disyllabic word recognition. *The Quarterly Journal of Experimental Psychology*, *66*(5), 946–972. doi:10.1080/17470218.2012.722660

Dirix, N., Cop, U., Drieghe, D., & Duyck, W. (in press). Cross-lingual Neighborhood Effects in Generalized Lexical Decision and Natural Reading. *Journal of Experimental Psycholoy:*

Learning, Memory and Cognition.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103–1123. doi:10.1037/0278-7393.26.5.I103

Gerhand, S., & Barry, C. (1999a). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, *27*(4), 592–602. doi:10.3758/BF03211553

Gerhand, S., & Barry, C. (1999b). Age-of-acquisition and frequency effects in speeded word naming. *Cognition*, *73*(2), B27 – B36. doi:10.1016/S0010-0277(99)00052-9

Ghyselinck, M., Custers, R., & Brysbaert, M. (2004). The effect of age of acquisition in visual word processing: Further evidence for the semantic hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 550-554. doi:10.1037/0278-7393.30.2.550

Johnston, R. a., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*, 789–845. doi:10.1080/13506280544000066

Joseph, H. S. S. L., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition, 133*(1), 238–48. doi:10.1016/j.cognition.2014.06.015

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*(5), 684–712. doi:10.1037/0033-2909.131.5.684

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1312–1318. doi:10.1037/0278-7393.29.6.1312

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846–863. doi:10.1080/13506280544000075

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990. doi:10.3758/s13428-012-0210-4

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–43. doi:10.3758/s13428-011-0146-0

Lambon Ralph, M. A., & Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. *Visual Cognition, 13*, 928–948. doi:10.1080/13506280544000110

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS One*, *7*(8), e43230. doi:10.1371/journal.pone.0043230

Monaghan, P., & Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language*, *63*(4), 506–525. doi:10.1016/j.jml.2010.08.003

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. doi:10.1080/17470210902816461

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157. doi:10.1037/0033-295X.105.1.125

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4–22. doi:10.1016/j.cogsys.2005.07.002

Stewart, N., & Ellis, A. W. (2008). Order of acquisition in learning perceptual categories: a laboratory analogue of the age-of-acquisition effect? *Psychonomic Bulletin & Review*, *15*(1), 70–4. doi:10.3758/PBR.15.1.70

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, *29*(1), 41–78. doi:10.1207/s15516709cog2901_3

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2013). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. doi:10.1080/17470218.2013.850521

CHAPTER 3

THE FIRST- AND SECOND-LANGUAGE AGE OF ACQUISITION EFFECT IN FIRST- AND SECOND-LANGUAGE BOOK READING.[1]

The age of acquisition (AoA) effect in first/monolingual language processing has received much attention in psycholinguistic research. However, AoA effects in second language processing were only investigated rarely. In the current study, we investigated first (L1) and second language (L2) AoA effects in a combined eye tracking and mega study approach. We analyzed data of a corpus of eye movements to assess the time course of AoA effects on bilingual reading. We found an effect of L2 AoA in both early and late measures of L2 reading: fixation times were faster for words that were learned earlier in L2. This suggests that the L2 AoA effect has an influence throughout the entire L2 reading process, analogous to the L1 AoA effect. However, we are also the first to find an early effect of L1 AoA on L2 processing: if the L1 translation of the L2 word was learned earlier, the L2 word was also read faster. We discuss the implications of these findings for two important hypotheses that offer an explanation for the AoA effect: the mapping and semantic hypothesis. We propose that the current results suggest an integration between these accounts.

[1] 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. doi: 10.1016/j.jml.2017.07.012.

INTRODUCTION

Through our lifetime, we continuously encounter and learn new words. The age of acquisition (AoA) of words has been identified as an important factor in language processing. A well-established finding, at least in first-language (L1) processing, is that words with an earlier AoA are processed faster than words with a late AoA. This effect has a long history of replications in a multitude of experiments, including different paradigms and techniques.

L1 age of acquisition

In the very first study that revealed an influence of word-level AoA, Carroll and White (1973) found that pictures were named faster when their name was learned at an earlier age. This AoA effect in picture naming has been replicated with different sets of stimuli and in different languages (Belke, Brysbaert, Meyer, & Ghyselinck, 2005; Morrison, Ellis, & Quinlan, 1992; Pérez, 2007) and was also found in word naming studies (Brysbaert, Lange, & Van Wijnendaele, 2000; Gerhand & Barry, 1999b; Morrison & Ellis, 1995).

AoA also influences word recognition: in lexical decision, reaction times (RTs) are faster for earlier acquired words (e.g., Brysbaert, Lange, et al., 2000; Bonin, Chalard, Méot, & Fayol, 2001; Butler & Hains, 1979; Gerhand & Barry, 1999a; Wilson, Cuetos, Davies, & Burani, 2013). Interestingly, in several of these studies (Bonin et al., 2001; Gerhand & Barry, 1999a; Wilson et al., 2013) an interaction was found between AoA and word frequency, with larger the AoA effects for low frequency words.

In two investigations of the English Lexicon Project (ELP; Balota et al., 2007), which consists of lexical decision data for 40 481 English words, the role of word-level AoA was investigated in combination with a large set of other linguistic variables (for example word frequency, length, …; Cortese & Khanna, 2007; Cortese & Schock, 2012). Both studies found an AoA effect, with shorter RTs for earlier learned words. The above interaction between word frequency and AoA also showed up in Cortese and Schock (2012).

Finally, a few studies investigated the AoA effect by means of eye tracking. In this paradigm, the eye movements of participants are recorded while they read pieces of natural text or sentences, without performing an artificial task like lexical decision. In two eye tracking studies, Juhasz and Rayner (2003, 2006) investigated AoA effects in sentence reading. In the 2003 study, AoA and other predictors were included as continuous variables, whereas in the 2006 study an orthogonal design was applied (early vs late AoA). In both studies, early and late timed measures were analyzed, and both yielded significant AoA effects (i.e., shorter fixations for early AoA words). In the 2006 study, an AoA effect was found for all eye tracking measures, whereas the 2003 study only found the AoA effect in early measures (single fixation duration and gaze duration). Juhasz and Rayner argue that the orthogonal design with extreme

AoA values was more sensitive to detect AoA effects in late word processing. These L1 AoA effects in eye tracking were recently replicated in a corpus study by Dirix and Duyck (in press), in which eye movement data of monolinguals reading an entire novel was investigated. L1 AoA effects on 7158 nouns were found in all timed measures (single, first fixation and gaze duration and total reading time), as well as an interaction between AoA and word frequency in total reading times (cf. the lexical decision studies discussed above). Finally, Juhasz, Gullick, and Shesler (2011) investigated the AoA effect with ambiguous words that had an early and late learned meaning (e.g., straw, volume). The sentence context disambiguated the meaning of the target word, and target words received shorter fixations (both in early and late measures) when the early learned meaning of the ambiguous word was relevant.

In sum, the AoA effect seems to be quite robust in the literature on monolingual/L1 language processing. Faster processing of earlier learned words has been found in a large variety of paradigms and in different modalities (see Johnston and Barry (2006) or Juhasz (2005) for reviews). Recent monolingual/L1 mega studies of lexical decision (e.g., Cortese & Schock, 2012) and eye movements (Dirix & Duyck, in press) validated the pioneer findings of smaller scale experiments.

Second language age of acquisition

Although the monolingual/L1 domain now approaches 45 years of AoA research, it has only been 15 years since word-level AoA has been investigated in the field of bilingualism, and studies are very rare. This is remarkable, because there is much more interindividual variability in the age at which words are learned for a second language (L2), so that the variable is possibly of greater relevance than for L1 processing. The majority of the words that we learn in L2 will also be known already in our L1, which creates an interesting situation: L2 words have an L2 AoA (the age at which the word was learned in L2), but also an L1 AoA (the age at which the L1 translation of the L2 word was learned). These L1 and L2 AoAs do not necessarily correspond: words that were learned early in L1 can be learned late in L2 and vice versa. Two main questions were addressed in the few L2 AoA studies that have been carried out. First, researchers investigated whether a word-level AoA effect may indeed be found in L2 processing. Second, it was investigated what mainly drives this AoA effect: the order at which the words were learned in the L1 or L2?

Izura and Ellis (2002) first addressed these questions. In their Experiment 1 (picture naming) and 2 (lexical decision), they found shorter RTs for earlier acquired words in L1 and L2, thus confirming the existence of a L2 AoA effect. To further assess whether it was the L1 or L2 AoA of the words that caused the AoA effect in L2, Izura and Ellis orthogonally manipulated the L1 and L2 AoA of their stimuli in Experiment 4 (lexical decision). Results showed only within-language AoA effects: in L1, RTs were faster for words learned early in L1, irrespective of when the words were learned in L2. Similarly, L2 reading was only influenced by order of acquisition in L2, not L1. The AoA seems to only have an impact within each language. Izura and Ellis (2004) later replicated these findings in both translation judgments and lexical decision. To date, these are the only two visual word recognition studies that investigated both the roles of L1 and L2 AoA in a full orthogonal design. For production, similar within-language AoA effects were also obtained in a bilingual picture naming task (Hirsh, Morrison, Gaset, & Carnicer ,2003).

In a spin-off of AoA research, the order of acquisition (OoA) effect of newly acquired stimuli is investigated. These “laboratory studies of AoA” allow researchers to study the impact of learning new stimuli at different points in time, while characteristics such as frequency can be controlled. Typically, a part of the stimuli set is introduced at the beginning of the study phase (“early acquired”); another part is presented at a later time (“late acquired”). This generally results in processing advantages for earlier learned items. For example, participants were faster to categorize “early” learned abstract checkerboard stimuli than a “later” learned set (Stewart & Ellis, 2008). In studies that involved linguistic material, similar results were obtained. Izura et al. (2011) found that early learned novel words for existing objects were processed faster in a series of behavioral tasks up to 35 days after the learning phase. Joseph, Wonnacott, Forbes, and Nation, (2014) found OoA effects on eye movements: total reading times decreased for novel words between the training and testing phase both for early and late learned items, but this effect was significantly larger for the early trained set.

These OoA studies support the robustness of acquisition effects, as OoA effects emerge even with a minimal delay between the presentation of the early and late stimuli set. Second, Izura et al. (2011) claim that these effects mirror real-life AoA effects, as the advantage for the early learned set can persist for weeks after training. Finally, studies involving linguistic materials could be interpreted as learning vocabulary of a novel language, mapping new lexical forms onto existing semantics, analogous to real life L2 learning.

To summarize, in the previous parts we have shown that L1 AoA is a well-established effect in psycholinguistic research. For L2 processing, some rare studies have confirmed L2 AoA effects, independent of L1 AoA, but the number of studies and stimuli is limited. Also, only isolated L2 word reading was investigated, and AoA eye tracking research for L2 sentence reading is completely lacking, until the present study. Our study will shed light on the specific time-course of AoA effects. Further, we will also argue that this approach may clarify the etiology of the (L1) AoA effect, about which two hypotheses exist.

The origin of the age of acquisition effect

The first hypothesis about the mechanism behind the AoA effect is the semantic hypothesis. According to this hypothesis, AoA effects originate from the organization of the semantic representational network of words (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Steyvers & Tenenbaum, 2005). As we learn new words or concepts, they are linked to semantic representations we already know. Early learned words take up a more central place in the semantic network, so that they are more easily accessible than later learned words. In a study of semantic networks, Steyvers and Tenenbaum (2005) indeed found that most nodes in the network have few connections, but they are joined through a few nodes with many connections, so-called "hubs" (cf. the early learned words).

There are a few sources of empirical evidence for a semantic locus of the AoA effect. First, earlier learned words were categorized faster in semantic categorization tasks (Brysbaert, Van Wijnendaele et al, 2000; Menenti & Burani, 2007), In a more complex design, Ghyselinck, Custers, and Brysbaert (2004) presented names of living and non-living stimuli, which were either printed in upper- or lowercase. Participants were instructed to judge the letter case of targets words by responding verbally, using the labels "living" and "non-living", so that responses were either congruent or incongruent with the semantic category of the words. Ghyselinck et al. found a larger congruency effect for early than late AoA words. The authors concluded that the meaning of early AoA words is activated faster than that of late AoA words, and proposed that semantics indeed have an important role in the AoA effect. Second, the magnitude of the AoA effect seems to increase with a higher need of semantic activation: it is smallest in word naming tasks, larger in lexical decision tasks and largest in object naming (Barry, Johnston, & Wood, 2006).

The second hypothesis explaining AoA effects is the mapping or connectionist hypothesis (Ellis & Lambon Ralph, 2000; Lambon Ralph & Ehsan, 2006; Monaghan & Ellis, 2010). In a connectionist modeling study, Ellis and Lambon Ralph (2000) conducted a series of simulations involving OoA. They trained networks on learning patterns; one set of patterns was introduced immediately ("early"), one set after a number of training cycles ("late"). Importantly, in analogy with human language learning, the training on the "early" set was continued together with learning the "late" set. They found that in a variety of circumstances the early set was always learned better than the late set. Ellis and Lambon Ralph relate this to AoA effects by the principle of mapping of word forms on meaning representations. Information that enters the network first has the advantage of network plasticity: it can have a large influence on the connections between input (word form) and output (meaning) representations, making them more easily accessible. With new information entering the network, it becomes more settled or entrenched. This allows for a progressive smaller influence on the connection weights and thus a disadvantage for later learned items. The mapping hypothesis does not specify a particular linguistic level at which these effects take place.

A first line of evidence for the mapping hypothesis comes from the OoA investigations. In particular, the study of Stewart and Ellis (2008) shows that even when learning random patterns, without semantics, an OoA effect emerges. Additional evidence comes from the L2 AoA literature. In their 2002 study, Izura and Ellis applied the following reasoning: if words in two languages share semantic representations (e.g., Kroll & Stewart, 1994; Van Hell & De Groot, 1998b), AoA effects of the (shared) semantic representations should transfer from L1 to L2. So, the semantic hypothesis predicts that the AoA effect in L2 should correspond to the age at which the L1 translations of the words were learned. However, the evidence on L2 AoA effects only shows within-language AoA effects, without an influence of L1 AoA on L2 processing (e.g., Hirsh et al., 2003; Izura & Ellis, 2002, 2004). Hence, Izura and Ellis situate the etiology of the AoA effect at the lexical level: when a second language is learned, new mappings (and connections) between in- and output have to be formed. These connections will be subject to the same mechanisms that apply to an L1 AoA effect: an advantage for the early items (because they profit from the network plasticity).

To summarize, whereas one etiological hypothesis about the AoA effect refers to the organization of the semantic network, the other hypothesis assumes a representational plasticity principle. At this point, evidence for both of these hypotheses creates a lack of consensus, although recent studies mostly support the latter. The approach to involve L2 AoA to determine the mechanism behind AoA effect therefore seems very interesting.

The present study

We investigated this matter further by conducting the first bilingual eye tracking sentence reading study of the (L2) AoA effect. The goal of the current study was twofold. First, we wanted to extend the L2 isolated word reading studies (e.g. Izura and Ellis, 2002) by applying eye tracking during natural reading, providing a better insight in the time course of the AoA effect in L2 reading. We were also interested to see whether L2 reading is indeed only influenced by L2 AoA, and not by L1 AoA, which is very informative for the etiology of the AoA effect.

We investigated L1 and L2 AoA effects in fixation time data of the Ghent Eye tracking COrpus (GECO; Cop, Dirix, Drieghe, & Duyck, in press), which contains eye movement data of bilinguals reading an entire novel in L1 and L2. Before analysis, lacking L2 AoA ratings for our stimuli needed to be collected. For L1, vast databases with AoA ratings are freely available (e.g., for Dutch, Brysbaert; Stevens, De Deyne, Voorspoels, & Storms, 2014; for English, Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). AoA ratings are usually collected by asking participants at which age they think they learned words. Although this might seem to be a rather poor indication of AoA, such ratings score well on validity (Brysbaert, in press; De Moor, Ghyselinck, & Brysbaert, 2000). In L2 AoA studies, ratings are typically only gathered for the (few tens of) stimuli presented in the experiment(s).

EXPERIMENT 1

In this experiment, we present the L2 AoA ratings for nearly 5 000 words, including the target stimuli from Experiment 2, using the method of Brysbaert et al. (2014). Participants had to indicate at which age they thought they learned a list of words. These ratings are freely available in the Supplementary Material and may be used in future L2 AoA studies that use similar late Dutch-English bilinguals.

Method

Participants and Materials. 126 undergraduates of Ghent University took part in this experiment (100 female, M_{age} = 18.94 [2.60]). They received course credit for their participation. All participants were unbalanced Dutch – English bilinguals, who received formal English education from age 13 on.

One part of the stimuli consisted of the 1 742 English nouns of the Ghent Eyetracking COrpus (GECO; Cop, Dirix, Drieghe, et al., in press), which was analyzed in Experiment 2. In addition, we selected 3 158 words from the English vocabulary test “wordORnot” of Ghent University (CRR, 2014).[2] To obtain a diverse sample of words, the 61 850 words from this test were divided into ten bins according to word frequency, and then subdivided again in ten bins according to word length. From each of the 100 resulting bins, 31 to 32 words were randomly selected. Combined with the GECO nouns, this resulted in 4 900 words. These were divided into six lists which were all matched on average word frequency and word length: two consisting of the GECO nouns and four of the remaining words.

Procedure. Each participant rated one of the word lists in an excel sheet. They were asked to indicate for each word at which age they believed they learned it, in analogy with L1 AoA rating studies (e.g., Brysbaert et al., 2014; Kuperman et al., 2012). The specific instructions can be found in Appendix 3A. Participants were also encouraged to complete the list in good conscience and not to fill in random numbers, as their data could not be used if it correlated poorly with the average ratings. All participants needed maximum one hour to complete their list.

[2] The authors would like to thank Emmanuel Keuleers for providing the stimuli list.

Results and Discussion

The data of four participants correlated less than .60 with the average AoA ratings. It can be expected that AoA ratings for L2 words show a smaller inter-individual consistency than for L1 words, as people may start to learn the language at a different age, but with such low correlations we cannot exclude the possibility of random responses. Therefore, we excluded these ratings from the dataset (cf. Ghyselinck, De Moor, & Brysbaert, 2000). The final dataset thus included 20 – 21 ratings for each word list.

The average correlation between the ratings was .76 (sd = .05), which is indeed somewhat lower than reported in L1 AoA studies (around .90). However, as mentioned before, this does make sense as the L2 learning onset differs more between participants that for L1 learning. Furthermore, as formal English education only starts at age 13 in Flanders, vocabulary acquisition before that age depends largely on which words participants encounter in their daily life.

In Figure 1, the distribution of the L2 AoA ratings is presented. This resembled a normal distribution, as was the case in the large scale L1 AoA ratings (e.g. Brysbaert et al., 2014)

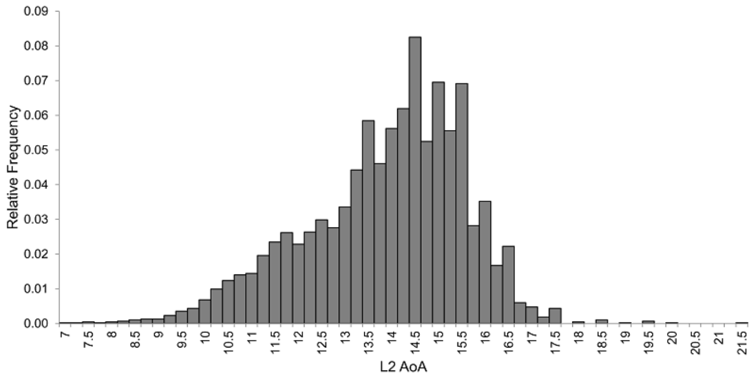

Figure 1. Frequency distribution of the L2 AoA ratings of Experiment 1.

We also visualized the relation between our L2 ratings and (a) their Dutch L1 translation AoA ratings (Figure 2; L1 AoA ratings from Brysbaert et al., 2014) and (b) their word frequency (Figure 3; SUBTLEX-UK frequencies from van Heuven, Mandera, Keuleers, & Brysbaert, 2013), as these were often reported as (highly) correlated in previous research.

L2 AoA was moderately correlated with L1 AoA (r = .52), which shows that word learning order roughly corresponds across languages, although some later learned L1 words may be earlier learned L2 words, and vice versa.

The correlation between L2 AoA and word frequency was somewhat higher (r = -.66). This further confirms the established relationship between these two lexical variables, and shows that also in L2 learning, high frequency words are learned earlier.

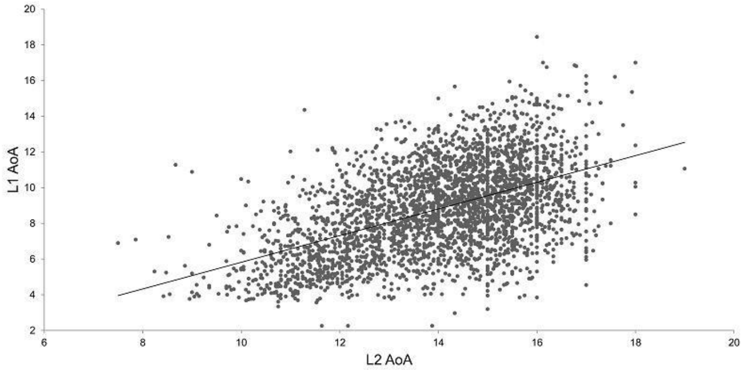

Figure 2. Scatter plot of the L2 AoA ratings of the words and the L1 AoA of their translation.

In conclusion, L2 ratings seem to show more inter-individual variability than L1 AoA ratings, but they show a lot of resemblance in terms of their characteristics: their distribution is similar and their relation with other lexical variables is in line with what could be expected. As such, they may be considered valid measures of the age at which our participants learned the L2 words for our L2 eye tracking analyses.

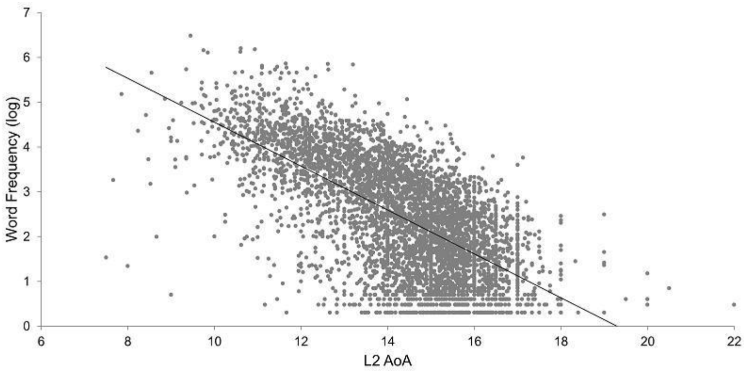

Figure 3. Scatter plot of the L2 AoA ratings of the words and their log word frequency.

EXPERIMENT 2

We investigated the L1 and L2 AoA effect in various reading measures of GECO (Cop, Dirix, Drieghe, et al., in press). This corpus consists of eye movement data of participants reading an entire book in their L1 (Dutch) and L2 (English). It has previously been successfully applied in investigations on sentence-level bilingual reading (Cop, Drieghe, & Duyck, 2015), the word frequency effect (Cop, Keuleers, Drieghe, & Duyck, 2015), the cognate effect (Cop, Dirix, Van Assche, Drieghe, & Duyck, in press) and cross-lingual orthographic neighborhood effects (Dirix, Cop, Drieghe, & Duyck, in press).

This study is the first investigation of the L2 AoA effect during reading of meaningful, longer passages of natural text, using eye tracking. While the studies of for example Izura and Ellis (2002, 2004) were of critical importance to the field and introduced L2 AoA as a concept, this eye tracking study offers additional value and insight to AoA research. First, eye tracking allows “natural reading”, as the participants only have the instruction to read the presented text, without further influence of task demands, or decision or response processes. Second, a detailed analyses of the time course of (AoA) effects can be made, as different timed measures represent different underlying processes (Rayner, 1998, 2009). So-called early measures (i.e., single fixation duration, first fixation duration and gaze duration) reflect lexical access, or selecting the correct representation of a word in the memory and accessing it. Late measures, such as total reading time, reflect higher order processes such as the semantic activation of the word, word verification and sentence comprehension. In analogy with the monolingual AoA studies by Juhasz and Rayner (2003, 2006) and Dirix and Duyck (in press), we investigated first fixation durations (the duration of the first fixation on a word), single fixation durations (the duration of the first fixation on words that were only fixated once), gaze durations (the duration of all fixations on a word before the eyes move to the right of the word) and total reading times (the summed duration of all fixations on a word, including refixations after regressions). By analyzing these measures, we can investigate whether AoA effects in L2 reading on both early and late processes of word recognition are similar to those in L1 reading, or whether the time course of L1 and L2 AoA differs across languages.

Furthermore, this is the first mega study of L2 AoA. By including a large amount of stimuli and data points, the variability in included word characteristics allows to assess their independent and simultaneous effects as continuous predictors in the analyses. As most previous studies orthogonally manipulated AoA, this is also beneficial to our insights in the AoA effect and the reliability of earlier results. Indeed, Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) argue that there are several disadvantages of applying a factorial design to study lexical processing. For example, there can be a decrease in statistical power and reliability when categorizing a continuous variable, or there are potentially contaminating factors that confound the factors of the design, which makes it difficult to estimate the influence of a categorical variable in a small set of stimuli.

Our results are also crucial for the discussion on the etiology of the AoA effect, as the semantic and mapping hypotheses make different predictions for the L2 AoA effect (cf. Izura &

Ellis, 2002). The semantic hypothesis predicts that the L1 AoA of the translation of L2 words should influence word processing in L2, given that L2 words are mapped onto existing L1/semantics (Kroll & Stewart, 1994). In this case, L1 AoA should have an effect both on L1 and L2 reading; L2 AoA should have no effect at all. This predicts an effect of L1 AoA on all fixation durations in L1 reading, similar to previous monolingual research; L2 reading should also be influenced by L1 AoA. In contrast, the mapping hypothesis predicts that L2 reading should only be influenced by the age at which the words were learned in L2. In this case, the AoA effects should only operate within languages: L1 AoA should only influence L1 processing; and L2 AoA should influence L2 processing. As the mapping hypothesis does not specify a single level at which AoA effects can occur, we would expect within-language AoA effects on both early and late measures.

We also considered the possibility that L1 or L2 AoA could interact with other word characteristics, such as word frequency (Dirix & Duyck, in press; Gerhand & Barry, 1999a; Wilson et al., 2013). We covered this by including interactions between the predictors in our primary statistical models. Next to word frequency, we also included word length, cross- lingual orthographic overlap and rank of occurrence as word characteristics. Orthographic overlap was included because Van Assche, Drieghe, Duyck, Welvaert and Hartsuiker (2011) showed that reading times are shorter as cross-lingual orthographic overlap between translation equivalents increases. This was operationalized by calculating the Corrected Levenshtein Distance[3] between each noun and its translation (Schepens, Dijkstra, & Grootjen, 2012). Rank of occurrence was included to control for repetition effects as some of the words are repeated multiple times throughout the novel. Finally, L1 and L2 proficiency were included as participant characteristics.

Method

Participants and Materials. The following criteria were used to select which nouns of the bilingual part of GECO (Cop, Dirix, Drieghe, et al., in press) were included in the current study: (a) only nouns for which an AoA rating of the word and its translation were available (in Brysbaert et al. (2014) for Dutch; in the new collected ratings for English, see Supplementary Material); (b) as we wanted reliable AoA estimates, we only selected words that at least 75% of the raters knew (similar to Izura & Ellis, 2002); (c) identical cognates or interlingual homographs were excluded. This resulted in 1069 unique Dutch nouns and 966 unique English nouns. See Table 1 for the characteristics of these nouns. The participants of GECO were 19 unbalanced Dutch-English bilinguals (17 females, Mage = 21.2, SDage = 2.2). They all read the

[3] The formula used for calculating the Corrected Levenshtein Distance:
$Orthographic\ Overlap = 1 - \frac{Distance}{Length}$

entire novel “The mysterious affair at Styles” by Agatha Christie (1920; in Dutch: “De zaak Styles). The L1 (Dutch) and L2 (English) proficiency of the participants was rated with the Dutch and English version of the LexTALE (Lemhöfer & Broersma, 2012). For Dutch, their average LexTALE score was 92.43 (SD = 6.34, range = [73.75 – 100]). For English, their average LexTALE score was 75.63 (SD = 12.87, range = [51.25 – 98.75]).

Table 1. *Descriptive Statistics for the GECO nouns analyzed in the current study, averaged over stimuli per language (standard deviations between parentheses).*

	Word Frequency[a]	Word Length	L1 AoA[b]	L2 AoA[c]	Rank of Occurrence[d]	CLD[e]
Dutch	1.51 (1.01)	6.49 (2.35)	6.83 (1.93)	12.35 (1.46)	18.53 (33.59)	0.33 (0.26)
English	1.63 (0.95)	5.91 (2.12)	7.04 (2.11)	12.33 (1.55)	12.32 (16.88)	0.37 (0.30)

[a]Log10 Subtlex frequencies per million words: SUBTLEX-NL for Dutch words (Keuleers, Brysbaert, & New, 2010), SUBTLEX-UK for English words (van Heuven et al., 2013); [b]For “Dutch”, this means the L1 AoA of the Dutch words, for “English” the L1 AoA of the Dutch translation of the English words (Brysbaert et al., 2014); [c]For “Dutch”, this means the L2 AoA of the English translation of the Dutch words, for English the L2 AoA of the English words (from the Experiment 1 ratings, see Supplementary Material).[d]The average amount of repetitions of each word throughout the novel. [e]The average amount of orthographic overlap, expressed by the Corrected Levenshtein Distance.

Procedure. All participants read the entire novel while their eye movements were recorded, spread over 4 separate sessions. Half of the novel was read in Dutch (L1), the other half in English (L2); the order was counterbalanced. Multiple-choice questions were presented after each chapter to ensure participants were paying adequate attention and reading for comprehension. We refer to Cop, Dirix, Drieghe, et al. (in press) and Cop, Keuleers, et al. (2015) for further details on the procedure.

Analyses of Eye Movements. The timed measures were fitted in a linear mixed model using the lme4 (version 1.1-11) and the lmerTest (version 2.0-30) packages in R (version 3.2.4, R Core Team, 2016). The initial models contained the fixed factors L1 AoA (continuous), L2 AoA (continuous), Language (Dutch or English), Word Frequency (continuous) per million, Word Length (continuous), L1 Proficiency (continuous), L2 Proficiency (continuous), the Rank of Occurrence (continuous) and Orthographic Overlap (continuous). Word frequency was log transformed with base 10 to normalize its distribution. All continuous variables were centered to reduce collinearity between main effects and interactions. Next to the fixed factors, a random intercept per subject and per word was included in all initial models (Baayen, Davidson, &

Bates, 2008). This was done to ensure that differences between subjects concerning genetic, developmental or social factors were modeled on one hand, and because our stimuli set does not contain all possible nouns in a language on the other.

A separate analysis was carried out for each dependent variable (i.e., the timed measures). First, the timed measure was log transformed with base 10 to normalize its distribution (see Lo & Andrews, 2015 for an alternative approach). Second, a full model including the two random clusters and all interactions (up to 3-way) was fitted. By backward fitting of the fixed effects, forward fitting of the random effects and again backward fitting of the fixed effects, the optimal model was discovered (Barr, Levy, Scheepers, & Tily, 2013).

An omnibus analysis was conducted for each dependent variable, so that we did not split up the analysis for each language in the first stage of the analysis. As a consequence, we considered effects that did not interact with the factor language as equally large for L1 and L2 reading. If there however was a significant interaction with language, we also conducted language-separate analyses to interpret this interaction. Finally, the Variance Inflation Factor (VIF) was calculated for each model to estimate the influence of multicollinearity on the regression coefficients. A VIF larger than 5 indicates moderate influence, larger than 10 is considered to be problematic (Fox & Weisberg, 2010).

Results

Single Fixation Duration. Only nouns that received exactly one fixation were selected for this analysis (54.60% of the data). Furthermore, single fixation durations that differed more than 2.5 standard deviations from the subject means were considered as outliers and excluded (2.35%). The outcome of the final model for single fixation durations is presented in Table 2. The maximum VIF for this model was 3.789.

Effects of L1 AoA. Across languages, there was no significant main effect of L1 AoA. The three-way interaction between L1 AoA, language and word length was significant (β = 0.0008, se = 0.0004, t = 2.133, p < .05).

Effects of L2 AoA. The main effect of L2 AoA was not significant. There was however a significant interaction between L2 AoA and Language (β = 0.0043, se = 0.0015, t = 2.863, p < .01; see Figure 4) and between L2 AoA and word frequency (β = -0.0018, se = 0.0007, t = -2.577, p < .05). There was a facilitating effect of L2 AoA on nouns with a log word frequency per million of 1.482 or less (χ^2 = 3.85, df = 1, p < 0.05). Finally, there was a significant interaction between L2 AoA and L1 proficiency (β = -0.0002, se < 0.0001, t = -2.431, p < .01) and a marginally significant one between L2 AoA and orthographic overlap (β = 0.0035, se < 0.0019, t = 1.822, p < .1). Post hoc contrasts showed that the effect of L2 AoA was significant when the L1 proficiency score of the participants was lower than 89.20 (χ^2 = 3.86, df = 1, p <

0.05; see Figure 11). No significant effects were found in the contrasts for the L2 AoA and orthographic overlap interaction.

L1 Reading. *Effects of L1 AoA.* Separate analyses for each language showed that the interaction between L1 AoA and word length was not significant for L1 reading (β = -0.0004, se = 0.0003, t = -1.475, p > .1).

Effects of L2 AoA. There was no main effect of L2 AoA on L1 reading (β = 0.0014, se = 0.0008, t = 1.433, p > .1).

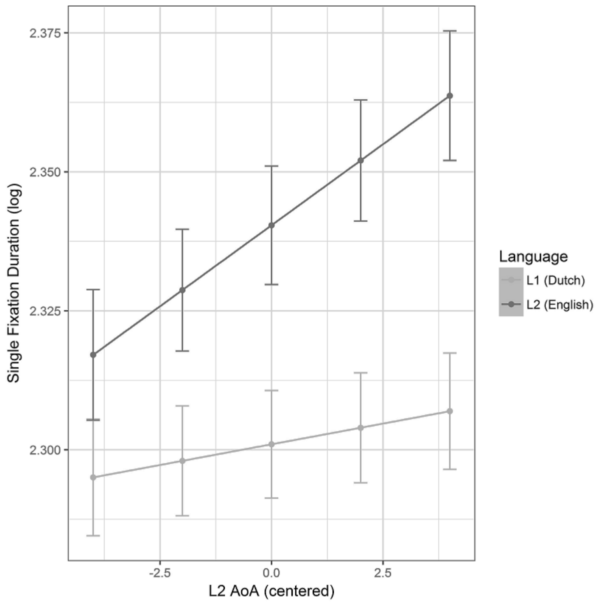

Figure 4. The interaction between Language (lines) and L2 AoA (x-axis) for single fixation durations (y-axis). Error bars represent standard errors.

L2 Reading. *Effects of L2 AoA.* Separate analyses for each language showed that there was facilitation for earlier L2 AoA in L2 reading (β = 0.0057, se = 0.0013, t = 4.454, p < .001; see Figure 4).

Effects of L1 AoA. The L2 analysis showed that the interaction between L1 AoA and word length was significant for L2 reading (β = 0.0006, se = 0.0003, t = 2.028, p < .05). Post-hoc contrasts showed that there was a marginally facilitating effect of an earlier L1 AoA, but only for words that contained 10 or more letters (χ^2 = 2.76, df = 1, p < 0.1).

Table 2. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for single fixation duration for bilingual reading.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.3018	0.0097	237.792	<.001	***
L1 Age of Acquisition	0.0011	0.0008	1.409	.159	
L2 Age of Acquisition	0.0015	0.0010	1.483	.138	
Language	0.0394	0.0041	9.714	<.001	***
Word Frequency	-0.0042	0.0021	-1.998	.050	.
Word Length	0.0027	0.0009	2.959	.005	**
L1 Proficiency	-0.0019	0.0015	-1.229	.235	
L2 Proficiency	0.0005	0.0008	0.642	.529	
Rank of Occurrence	-0.0001	<0.0001	-1.470	.142	
Orthographic Overlap	-0.0021	0.0029	-0.736	.462	
L1 AoA * Language	-0.0013	0.0010	-1.281	.200	
L1 AoA * Word Length	-0.0003	0.0003	-1.083	.279	
L2 AoA * Language	0.0043	0.0015	2.863	.004	**
L2 AoA * Word Frequency	-0.0018	0.0007	-2.577	.010	*
L2 AoA * L1 Proficiency	-0.0002	0.0001	-2.431	.015	*
L2 AoA * Orthographic Overlap	0.0035	0.0019	1.822	.069	.
Language * Word Frequency	-0.0050	0.0026	-1.928	.054	.
Language * Word Length	0.0011	0.0009	1.140	.254	
Word Frequency * Word Length	-0.0020	0.0006	-3.231	.001	**
L1 AoA * Language * Word Length	0.0008	0.0004	2.133	.033	*
Language * Word Frequency * Word Length	0.0023	0.0009	2.580	.010	**

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0003	0.0167
Subject		
(Intercept)	0.0017	0.0417
Language	0.0002	0.0155
Word Frequency	<0.0001	0.0056
Word Length	<0.0001	0.0027
Word Frequency * Word Length	<0.0001	0.0008

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

First Fixation Duration. First fixation durations that differed more than 2.5 standard deviations from the subject means were considered as outliers and excluded from the dataset

(2.29%). In comparison to single fixation duration, this measure included all first fixations on the target nouns, irrespective of later refixations. The maximum VIF was 4.174. The final model is presented in Table 3.

Effects of L1 AoA. There was a general significant main effect of L1 AoA (β = 0.0017, se = 0.0007, t = 2.397, p < .05): first fixation durations were shorter for nouns with an earlier L1 AoA. The two-way interaction between L1 AoA and language was marginally significant (β = -0.0016, se = 0.0009, t = -1.713, p < .1). Furthermore, the three-way interaction between L1 AoA, language and word length was significant (β = 0.0007, se = 0.0003, t = 1.995, p <.05; see Figure 5).

Effects of L2 AoA. The main effect of L2 AoA across languages was not significant, but the interaction with language was marginally significant (β = 0.0025, se = 0.0013, t = 1.889, p < .1; see Figure 6). The interaction between L2 AoA and word frequency was significant (β = -0.0013, se = 0.0006, t = -2.081, p < .05): a facilitatory effect of L2 AoA was present for nouns with a word frequency of 1.280 or less (χ^2=3.86, df=1, p < 0.05). Finally, the interaction between L2 AoA and L1 proficiency was significant (β = -0.0001, se = 0.0001, t = -2.104, p < .05). Post hoc contrasts showed that the effect of L2 AoA was significant when the L1 proficiency score of the participants was lower than 87.25 (χ^2 = 3.85, df = 1, p < .05; see Figure 11).

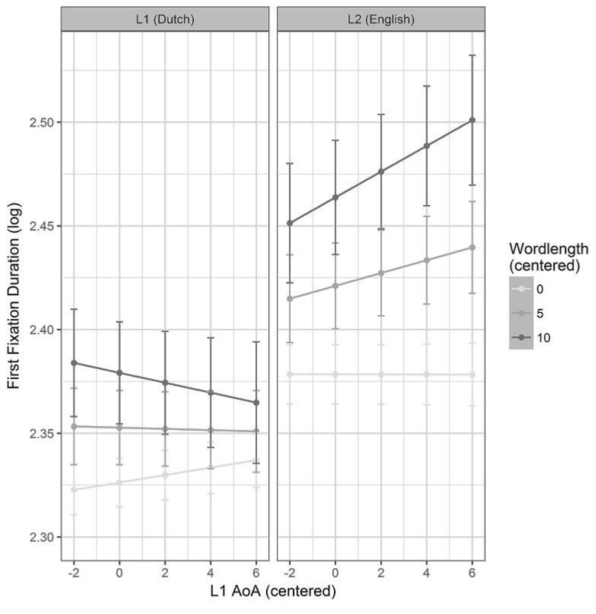

Figure 5. The interaction between L1 AoA (x-axis) and word length (lines) in each language (panels) for first fixation durations (y-axis). Error bars represent the standard error.

L1 Reading. *Effects of L1 AoA.* In the L1 analysis, there was a facilitatory effect of L1 AoA on L1 nouns (β = 0.0016, se = 0.0007, t = 2.489, p < .05). The interaction between L1 AoA and word length was not significant for L1 reading (β = -0.0003, se = 0.0002, t = -1.310, p > .1).

Effects of L2 AoA. We found no main effect of L2 AoA on L1 nouns (β = 0.0013, se = 0.0009, t = 1.486, p > .1).

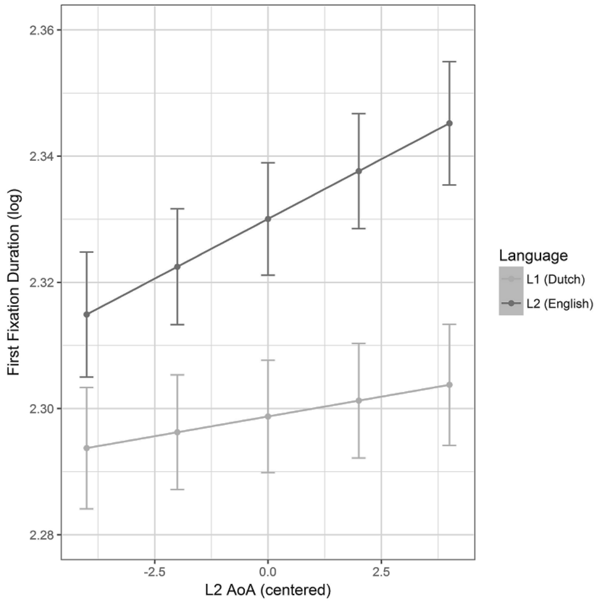

Figure 6. The interaction between Language (lines) and L2 AoA (x-axis) for first fixation durations (y-axis). Error bars represent standard errors.

L2 Reading. *Effects of L2 AoA.* The L2 analysis showed that there was a facilitating effect of L2 AoA for L2 nouns (β = 0.0036, se = 0.0011, t = 3.194, p < .01; see Figure 6).

Effects of L1 AoA. There was no main effect of L1 AoA on L2 nouns (β = <-0.0001, se = 0.0006, t = -0.009, p > .05). However, the interaction between L1 AoA and word length was significant in L2 (β = 0.0005, se = 0.0010, t = 1.964, p < .05; see Figure 5): there was a facilitatory effect for nouns of 12 or more letters when the L1 AoA was earlier (χ^2 = 3.88, df = 1, p < 0.05).

Table 3. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for first fixation duration for bilingual reading.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.2990	0.0089	258.241	<.001	***

L1 Age of Acquisition	0.0017	0.0007	2.397	.017	*
L2 Age of Acquisition	0.0013	0.0009	1.383	.167	
Language	0.0313	0.0017	18.767	<.001	***
Word Frequency	-0.0036	0.0017	-2.149	.034	*
Word Length	0.0014	0.0007	1.954	.055	.
L1 Proficiency	-0.0023	0.0020	-1.196	.249	
L2 Proficiency	0.0009	0.0010	0.926	.368	
Rank of Occurrence	<0.0001	<0.0001	-0.501	.616	
Orthographic Overlap	-0.0032	0.0026	-1.234	.217	
L1 AoA * Language	-0.0016	0.0009	-1.713	.087	.
L1 AoA * Word Length	-0.0002	0.0002	-1.016	.310	
L2 AoA * Language	0.0025	0.0013	1.889	.059	.
L2 AoA * Word Frequency	-0.0013	0.0006	-2.081	.038	*
L2 AoA * L1 Proficiency	-0.0001	0.0001	-2.104	.036	*
Language * Word Frequency	-0.0033	0.0023	-1.466	.143	
Language * Word Length	-0.0001	0.0008	-0.115	.909	
Word Frequency * Word Length	-0.0006	0.0005	-1.202	.229	
L1 AoA * Language * Word Length	0.0007	0.0003	1.995	.046	*
Language * Word Frequency * Word Length	0.0013	0.0007	1.712	.087	.

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0003	0.0166
Subject		
(Intercept)	0.0015	0.0384
Word Frequency	<0.0001	0.0034
Word Length	<0.0001	0.0021

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Gaze Duration. Gaze durations that differed more than 2.5 standard deviations from the subject means were considered as outliers and excluded from the dataset (2.67%). The final model had a maximum VIF of 3.894; it is presented in Table 4.

Effects of L1 AoA. Across languages, there was a marginally significant main effect of L1 AoA ($\beta = 0.0018$, se = 0.0009, $t = 1.939$, $p < .1$): gaze durations were shorter for an earlier L1 AoA. The three-way interaction between L1 AoA, language and word length was again significant ($\beta = 0.0010$, se = 0.0004, $t = 2.452$, $p < .05$; see Figure 7).

Effects of L2 AoA. There was no general significant main effect of L2 AoA. There was a significant interaction between L2 AoA and Language ($\beta = 0.0041$, se = 0.0017, $t = 2.335$, $p < .01$; see Figure 8). L2 AoA also interacted significantly with orthographic overlap ($\beta = 0.0047$, se = 0.0022, $t = 2.135$, $p < .05$). Post-hoc contrasts showed that the L2 AoA effect was larger for words with a CLD of 0.51 or higher ($\chi^2 = 3.87$, df = 1, $p < 0.05$). The interaction between L2 AoA and word frequency was marginally significant (β = -0.0014, se = 0.0008, $t = -1.776$, $p < .1$). Contrast revealed that there was a facilitatory effect of an earlier L2, which was only significant for nouns with a log word frequency up to 1.265 ($\chi^2 = 3.85$, df = 1, $p < 0.05$).

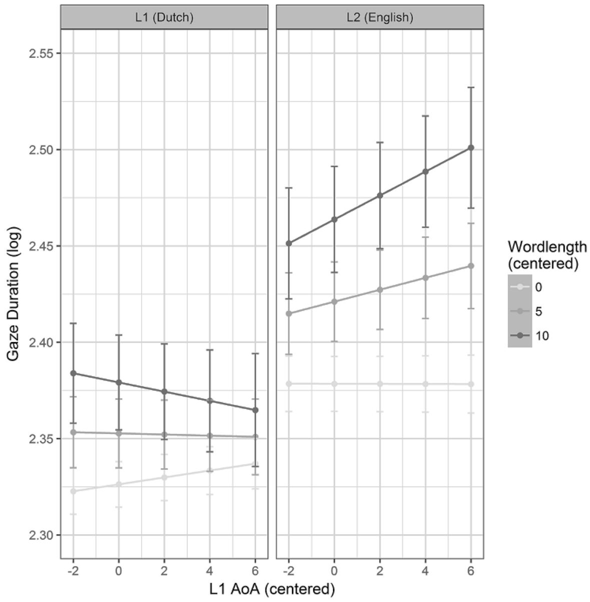

Figure 7. The interaction between L1 AoA (x-axis) and word length (lines) in each language (panels) for gaze durations (y-axis). Error bars represent standard errors.

L1 Reading. *Effects of L1 AoA.* Separate analyses for each language showed that there was a facilitatory effect of L1 AoA on L1 reading (β = 0.0017, se = 0.0008, $t = 2.005$, $p < .05$). The interaction between L1 AoA and word length was marginally significant in L1 (β = -0.0005, se = 0.0003, $t = -1.694$, $p < .1$). Post-hoc contrasts revealed no significant effects.

Effects of L2 AoA. There was no main effect of L2 AoA on L1 reading (β = 0.0017, se = 0.0011, t = 1.502, p > .1).

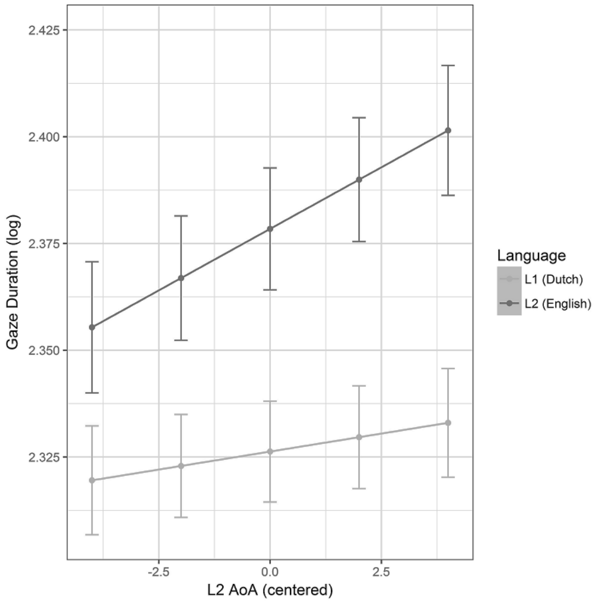

Figure 8. The interaction between Language (lines) and L2 AoA (x-axis) for gaze durations (y-axis). Error bars represent standard errors.

L2 Reading. *Effects of L2 AoA.* The L2 analysis showed that the facilitatory effect of an earlier L2 AoA was significant for L2 reading (β = 0.0058, se = 0.0015, t = 3.781, p < .001; see Figure 8).

Effects of L1 AoA. The interaction between L1 AoA and word length was significant for L2 reading (β = 0.0007, se = 0.0003, t = 2.086, p < .05; see Figure 7). Post-hoc contrasts in L2 revealed that the facilitatory effect of an earlier L1 AoA was only significant for nouns with 14 characters or more (χ^2 = 3.87, df = 1, p < 0.05).

Table 4. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for gaze duration for bilingual reading.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.3269	0.0118	197.070	<.001	***
L1 Age of Acquisition	0.0018	0.0009	1.939	.053	.
L2 Age of Acquisition	0.0017	0.0012	1.416	.157	
Language	0.0522	0.0060	8.716	<.001	***
Word Frequency	-0.0065	0.0024	-2.744	.008	**
Word Length	0.0053	0.0015	3.623	<.001	**
L1 Proficiency	-0.0005	0.0012	-0.442	.663	
L2 Proficiency	-0.0003	0.0006	-0.532	.601	

Rank of Occurrence	0.0000	0.0000	-1.179	.238	
Orthographic Overlap	-0.0033	0.0033	-0.998	.319	
L1 AoA * Language	-0.0018	0.0012	-1.500	.134	
L1 AoA * Word Length	-0.0004	0.0003	-1.338	.181	
L2 AoA * Language	0.0041	0.0017	2.335	.020	*
L2 AoA * Word Frequency	-0.0014	0.0008	-1.776	.076	.
L2 AoA * Orthographic overlap	0.0047	0.0022	2.135	.033	*
Language * Word Frequency	-0.0046	0.0030	-1.544	.123	
Language * Word Length	0.0032	0.0011	3.083	.002	**
Word Frequency * Word Length	-0.0025	0.0007	-3.609	<.001	***
L1 AoA * Language * Word Length	0.0010	0.0004	2.452	.014	*
Language * Word Frequency * Word Length	0.0030	0.0010	3.090	.002	**

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0006	0.0253
Subject		
(Intercept)	0.0026	0.0508
Language	0.0006	0.0240
Word Frequency	<0.0001	0.0062
Word Length	<0.0001	0.0055
Word Frequency * Word Length	<0.0001	0.0010

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Total Reading Times. Total reading times that differed more than 2.5 standard deviations from the subject means were considered as outliers and removed from the dataset (2.89%). The final model for total reading times is presented in Table 5. The maximum VIF for this model was 3.894.

Effects of L1 AoA. The main effect of L1 AoA was significant across languages (β = 0.0029, se = 0.0011, t = 2.608, p < .01). Total reading times were shorter for nouns with an earlier L1 AoA. The interaction between L1 AoA and language was significant (β = -0.0034, se = 0.0015, t = -2.329, p < .05; see Figure 9). The three-way interaction between L1 AoA, language and word length was marginally significant (β = 0.0009, se = 0.0005, t = 1.770, p <.1).

Effects of L2 AoA. The main effect of L2 AoA did not reach significance, but there was a significant interaction between L2 AoA and language (β = 0.0087, se = 0.0021, t = 4.158, p < .001; see Figure 10). The interaction between L2 AoA and L1 proficiency was again significant (β = -0.0003, se = 0.0001, t = -2.874, p < .01). Post hoc contrasts showed that the effect of L2

AoA was significant when the L1 proficiency score of the participants was lower than 85.60 (χ^2 = 3.85, df = 1, $p < 0.05$; see Figure 11).

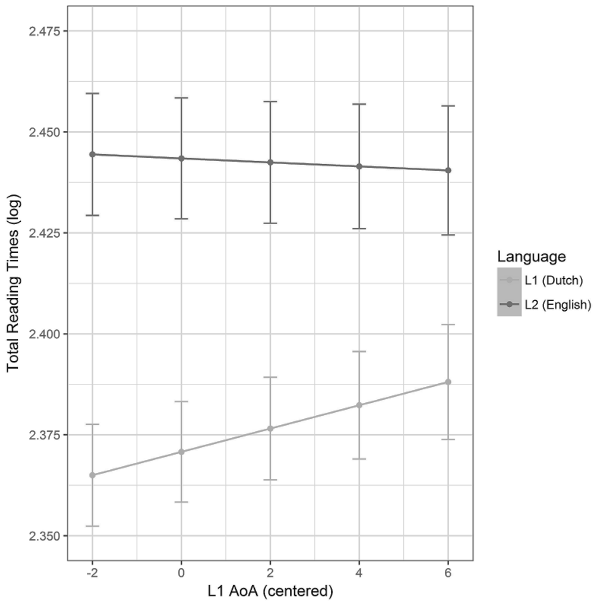

Figure 9. The interaction between Language (lines) and L1 AoA (x-axis) for total reading times (y-axis). Error bars represent standard errors.

L1 Reading. *Effects of L1 AoA.* The separate L1 analysis showed that the facilitatory effect of L1 AoA was significant for L1 reading (β = 0.0027, se = 0.0010, t = 2.608, p < .01; see Figure 9). Furthermore, the interaction between L1 AoA and word length was not significant for L1 nouns (β = -0.0005, se = 0.0004, t = -1.320, p > .1).

Effects of L2 AoA. The effect of L2 AoA was not significant for L1 reading (β = 0.0015, se = 0.0013, t = 1.150, p > .1).

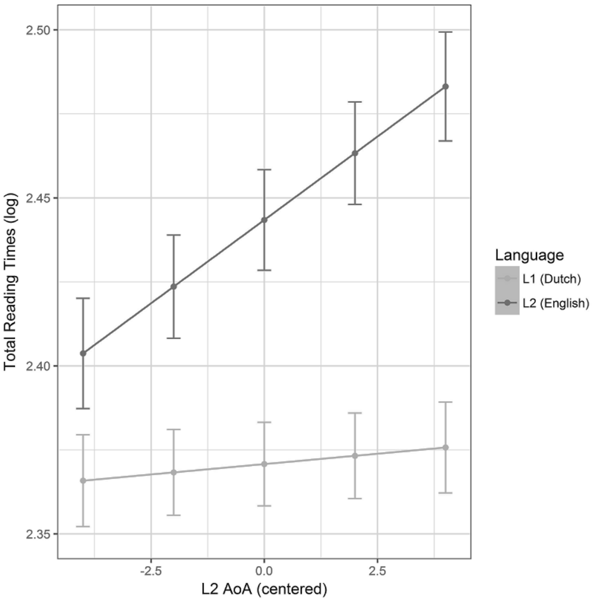

Figure 10. The interaction between Language (lines) and L2 AoA (x-axis) for total reading times (y-axis). Error bars represent standard errors.

L2 Reading. *Effects of L2 AoA.* In the separate L2 analysis it was revealed that the facilitatory effect of an earlier L2 AoA was significant for L2 reading (β = 0.0095, se = 0.0017, t = 5.494, p < .001; see Figure 10).

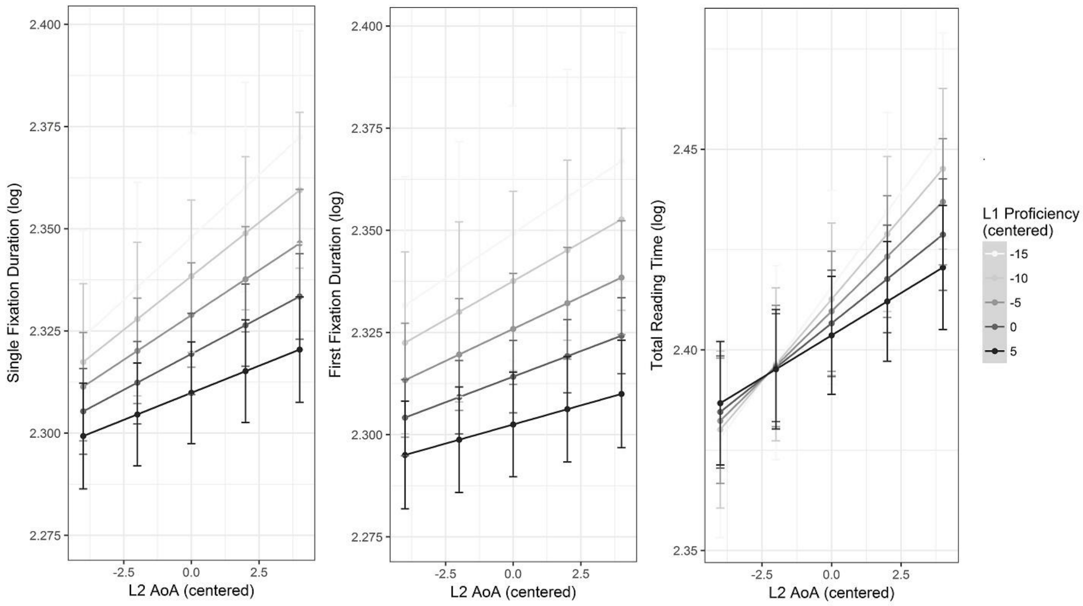

Figure 11. The interaction between L2 AoA (x-axis) and L1 proficiency (lines) for single fixation duration (left panel), first fixation duration (middle panel) and total reading time (right panel). Error bars represent standard errors.

Effects of L1 AoA. The facilitatory effect of L1 AoA was not significant in L2 (β = -0.0006, se = 0.0010, t = -0.587, p > .1). Furthermore, the interaction between L1 AoA and word length was also not significant for L2 reading (β = 0.0005, se = 0.0004, t = 1.470, p >.1).

Table 5. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for total reading time for bilingual reading.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.3726	0.0125	190.535	<.001	***
L1 Age of Acquisition	0.0029	0.0011	2.608	.009	**
L2 Age of Acquisition	0.0012	0.0014	0.905	.366	
Language	0.0727	0.0073	9.959	<.001	***
Word Frequency	-0.0066	0.0031	-2.122	.038	*
Word Length	0.0070	0.0015	4.524	<.001	***
L1 Proficiency	-0.0006	0.0013	-0.451	.658	
L2 Proficiency	-0.0002	0.0006	-0.337	.741	
Rank of Occurrence	-0.0001	0.0000	-2.695	.007	**
Orthographic Overlap	-0.0009	0.0040	-0.226	.821	
L1 AoA * Language	-0.0034	0.0015	-2.329	.020	*
L1 AoA * Word Length	-0.0004	0.0004	-1.045	.296	
L2 AoA * Language	0.0087	0.0021	4.158	<.001	***
L2 AoA * L1 Proficiency	-0.0003	0.0001	-2.874	.004	**
Language * Word Frequency	-0.0083	0.0035	-2.336	.020	*
Language * Word Length	0.0043	0.0013	3.430	.001	***
Word Frequency * Word Length	-0.0050	0.0007	-6.688	<.001	***
L1 AoA * Language * Word Length	0.0009	0.0005	1.770	.077	.
Language * Word Frequency * Word Length	0.0042	0.0011	3.716	<.001	***

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0009	0.0305
Subject		
(Intercept)	0.0029	0.0534
Language	0.0008	0.0292
Word Frequency	0.0001	0.0092
Word Length	<0.0001	0.0055

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Discussion

L1 Reading. In all timed measures except for single fixation durations, there was a significant main effect of L1 AoA on L1 reading: words with an earlier L1 AoA received shorter fixations than words with a later L1 AoA. These effects are largely consistent with previous monolingual AoA research in general (e.g., Brysbaert, Lange, et al., 2000; Carroll & White, 1973; Cortese & Khanna, 2007; Gerhand & Barry, 1999b; Morrison et al., 1992) and the few AoA eye tracking studies in particular (Dirix & Duyck, in press; Juhasz & Rayner, 2003, 2006). Even in the natural reading of long texts, AoA consistently and reliably influences word recognition throughout early and late stages of processing.

Furthermore, across reading languages, we encountered an interaction between L2 AoA and word frequency on two timed measures. In monolingual AoA investigations, an interaction between AoA and word frequency has been reported on a few occasions (e.g., Bonin et al., 2001; Dirix & Duyck, in press; Gerhand & Barry, 1999a; Wilson et al., 2013), with a stronger facilitatory AoA effect for low frequent words. The L2 AoA by word frequency interaction followed the same pattern.

To summarize, the AoA effects in L1 reading are partially consistent with previous research (i.e., the facilitating L1 AoA effect on L1 reading), but we also discovered some minor L2 AoA influences on L1 reading.

L2 Reading. For all of the timed measures that we analyzed, there was a significant interaction between L2 AoA and language: a facilitatory effect of L2 AoA was only present in L2 reading. These results are in line with previous L2 AoA studies (Hirsh et al., 2003; Izura & Ellis, 2002, 2004), which also found a facilitatory L2 AoA effect on L2 isolated word processing.

Furthermore, for single and first fixation duration, the interaction between word frequency and L2 AoA was significant: the AoA effect was again larger for low frequent words. This is consistent with the findings of monolingual studies (Dirix & Duyck, in press; Gerhand & Barry 1999a; Wilson et al. 2013). Wilson et al. argue that this interaction can be explained through the processing speed of high vs low frequency words: orthographic familiarity is higher for high frequency words, so that they are more easily and rapidly accessible. For low frequency words, lower familiarity and processing speed leaves more room for an additional influence of other word characteristics, such as a faster access for early AoA words.

An interaction between L2 AoA and L1 proficiency was present in single fixation duration, first fixation duration and total reading time. The L2 AoA effect was less pronounced when the L1 proficiency of the participants was higher. A similar interaction has been found in the word frequency study by Cop, Keuleers, et al. (2015) using the same database and participant characteristics, between L1 proficiency and L2 word frequency. They argued that the L1 proficiency measure probably entails more than L1 exposure, possibly a general language skill or aptitude. In analogy with their proficiency – word frequency interaction, it is indeed not unreasonable to assume that more language proficient participants not only show reduced frequency effects, but also reduced AoA effects.

There was also a significant interaction between L2 AoA and orthographic overlap in gaze durations: the AoA effect was larger for words with a high amount of orthographic overlap with their translational equivalent. As words with high orthographic overlap are accessed more easily (e.g., Van Assche et al., 2011), it could be that they receive an additional boost when they have already resided for a long time in the representational network. Alternatively, cognate-like words may have larger semantic overlap across languages (Van Hell & De Groot, 1998a), so that they should yield a larger AoA effect if part of that effect originates from semantics.

Surprisingly, whereas L1 AoA did not have an influence on L2 processing in previous research (e.g., Hirsh et al., 2003; Izura & Ellis, 2002), we are the first to find an influence of L1 AoA on L2 reading: for all timed measures except single fixation duration, there was a (marginally) significant main effect of L1 AoA on L2 reading: words with an earlier L1 AoA received shorter fixations than words with a later L1 AoA. After further inspection, it seemed that the L1 AoA effect especially arose for L2 words that take longer to process (i.e. longer words of at least nine to twelve letters): there was a facilitatory effect on single/first fixation and gaze duration when the translation of these words were learned early in L1. This is plausible given that the L1 AoA effect on L2 reading is assumed to originate from shared semantics across languages, which takes time to activate during reading, especially for longer words.

In conclusion, an earlier learning age of L2 words facilitates L2 reading. In addition, L1 AoA also seems to play a role in several measures of L2 natural reading.

GENERAL DISCUSSION

The age at which we learn words influences their processing speed (e.g., Brysbaert & Ghyselinck, 2006; Gerhand & Barry, 1999a; Juhasz & Rayner, 2006; Morrison et al., 1992). This mechanism also applies to L2 (Hirsh et al., 2003; Izura & Ellis, 2002, 2004), although earlier findings are limited to isolated L2 word processing. In the current study, we analyzed eye movement data of a corpus of bilingual natural reading (GECO; Cop, Dirix, Drieghe, et al., in press). Our first goal was to investigate L1 and L2 AoA effects in L1 and L2 reading using eye tracking, in order to provide a detailed analysis of the time course of AoA effects. Of particular interest was the effect of L2 (and potentially L1) AoA on L2 reading. Furthermore, we wanted to test the predictions of the semantic and mapping hypotheses, in order to clarify the origin of the AoA effect.

Consistent with previous monolingual research (Dirix & Duyck, in press; Juhasz & Rayner, 2006), we found an L1 AoA effect for L1 reading on both early (first fixation and gaze duration) and late measures (total reading times). It seems that AoA has an influence throughout the entire reading process, making it easier for earlier learned words to access the representations of words in lexical memory on the one hand, and to activate their meaning and integrate them into sentences on the other hand.

The effects of L2 AoA on L2 processing were consistent with the previous research on isolated word reading: fixation times were shorter for words that were learned earlier in L2. The current study however was the first investigation providing evidence from eye movements, showing that the L2 AoA effect affects the entire time course of L2 word recognition (in analogy with L1 AoA and L1 reading): L2 words that are learned earlier yield benefits for eye tracking measures that reflect initial lexical access, as well as for measures that reflect semantic access and integration. This is consistent with the notion that the origin of the AoA effect may situate itself at different representational levels.

Interestingly, we are also the first to find a cross-lingual AoA influence on L2 reading: in the early reading stages (single/first fixation and gaze duration), longer L2 words were processed faster when their L1 translation was learned early. This is consistent with a semantic etiology of the AOA effect: if one assumes that L2 translational equivalents are mapped onto the existing semantic representations that also serve L1 (Duyck & Brysbaert, 2004; Kroll & Stewart, 1994; Van Hell & De Groot, 1998b), L2 processing should indeed be influenced by L1 AoA, because that measure reflects when the semantic representation that the L2 word is mapped onto, was created (Izura & Ellis, 2002). As noted, this effect interacted with word length: processing is slower for longer words (especially in L2), so it could be that only for these words sufficient time surpasses for this semantic activation to occur. Only then the L1 AoA influence, which originates from the semantic organization of the word network, may influence L2 word recognition.

This cross-lingual AoA effect contrasts with earlier investigations of L2 AoA, who only reported L2 AoA effects on L2 processing (Hirsh et al., 2003; Izura & Ellis, 2002, 2004). There are several reasons to explain this discrepancy. First, there is a potential influence of task characteristics. Whereas participants simply have to read the presented text in natural reading, in other paradigms there can be influences of decision components or answer strategies. Kuperman, Drieghe, Keuleers, and Brysbaert (2013) indeed show that shared variance between lexical decision RTs and eye movement measures may be surprisingly low. This could mean that these two tasks, although they both involve visual word recognition, partially tap into different processes. Second, because we included a large amount of stimuli and the AoA variables as continuous, as opposed to factorial designs, our approach might be more sensitive to discover the subtle effects of L1 AoA on L2 reading. Third, this approach also allowed us to include complex interactions. The small L1 AoA influence on L2 reading was found in the interaction with word length and language. Finally, note that the lexical decision tasks of Izura and Ellis (2002, 2004), who only found within-language AoA effects, likely involve semantics, in order to determine whether the letter string corresponds to an existing meaning. In the present study, the eye tracking measures that reflect later stages of word recognition (e.g. total reading times) also only showed within-language AoA effects, similar to Izura and Ellis, and in contrast with the early eye tracking measures that reflect initial lexical access. In Appendix 3B, we present data from a lexical decision task with the target words of the current study, in which we replicate the null cross-lingual AoA effect of Izura and Ellis.

Finally, we found that L2 AoA has an influence on processing of very low-frequent L1 words. A possible explanation may lie in the higher activation threshold of low-frequency words (McClelland & Rumelhart, 1981). Because we know that lexical access in bilinguals is non-selective (Dijkstra & Van Heuven, 2002; Van Assche, Duyck, Hartsuiker, & Diependaele, 2009), and following the rationale of the semantic hypothesis (i.e., organization of words in a semantic network; (Brysbaert, Van Wijnendaele, et al., 2000; Steyvers & Tenenbaum, 2005), it could be that exposure to L2 words that are learned early affects the position of the underlying semantic representation in the network sufficiently if that representation was very weak to start with (very low-frequent L1 words). Note that this L2 AoA effect was stronger for participants with low L1 proficiency scores. It could be that especially for low-proficient participants, who are supposed to have weaker representations of low frequency words (e.g., Cop, Keuleers, et al., 2015), the L2 AoA order indeed influences these low frequency representations even more strongly, as argued before.

Hypotheses of the AoA effect

We outlined two important hypotheses explaining the AoA effect. The semantic hypothesis situates the origin of AoA effects in the accessibility of semantic representations. It predicts that L2 reading should be affected by L1 AoA, if one assumes that L1 and L2 translational equivalents share semantic representations (Kroll & Stewart, 1994; Van Hell & De Groot, 1998b). In contrast, the mapping hypothesis postulates that it is only the order in which information enters a network determines AoA effects, as there is a plasticity (and processing) advantage for early entered information. As new input-output mappings (corresponding to word form and meaning) have to be installed when learning a new language, within-language AoA effects should emerge: the L2 AoA effect should be in accordance with the age at which the words were learned in L2.

In our study, we found results that support both these hypotheses. First, and most clearly, there is a within-language effect of (L2) AoA on L2 processing: on all timed measures, L2 reading was faster when the words were learned earlier in L2. This finding supports the mapping hypothesis, as the L2 learning order is determining here. The mapping hypothesis does not specify a particular linguistic level at which AoA effects could arise, and we indeed found that L2 AoA influenced both measures that reflect lexical access (single/first fixation and gaze duration) and access to the meaning or verification of the words (total reading time). However, there was also a limited but reliable effect of L1 AoA on L2 reading in our data, which supports the semantic hypothesis. For longer L2 words (only), processing is speeded if their L1 translational equivalent is early acquired. To sum up, it seems that AoA effects in late language processing are language-exclusive, but cross-lingual L1 AoA effects show op for longer L2 words that take longer to process.

In the L2 AoA or OoA literature, the semantic and mapping hypotheses are often portrayed as opposites, with specific predictions, that usually result in support of the mapping hypothesis. In an attempt to reconcile this with the current findings, we suggest an integration between the mapping and semantic hypotheses. In AoA/OoA research, there seems to be a general principle of "first learned, faster processed". The mapping hypothesis provides an excellent and parsimonious explanation for this finding. However, we have to keep in mind that we are studying language. It is not unreasonable that words in different languages, but with the same meaning, share semantic representations. These representations are more easily accessed when learned earlier, whether it is through the L1 or L2. It can indeed be the case that early learned words can alter a network's weights in its advantage more than late learned words, but at the same time it may also be that the semantic representation of the early learned word takes up a more central place in the network. Both of these AoA mechanisms then may influence the processing speed of words independently and simultaneously.

This brings us to two additional related topics: the organization of the (bilingual) lexicon and the critical acquisition period. From the interpretation of our results, we can conclude that AoA heavily influences the organization of the lexicon: the age at which you learn a word has a large impact on the position it will take up in the lexicon, and how easily accessible it will be. In the specific case of bilinguals, our results also point towards a shared lexicon for the two languages (see Dijkstra & van Heuven, 2002), as direct influences of L1 word characteristics on their L2 counterpart seem to take place. Furthermore, as AoA effects also emerge in late-learned L2, Izura & Ellis (2002) argued that the AoA effect is probably not due to some kind of critical period of 'easy' language acquisition (see Marinova-Todd, Marshall, & Snow, 2000) that would only apply to the period of L1 learning. There is indeed evidence that the onset of learning a new language has a later limited influence on word recognition processes (if controlling for proficiency; e.g., Cardimona, Smith & Roberts, 2015; Foote, 2014; Montrul & Foote, 2014). It seems that there however is a “relative” critical period in language acquisition: irrespective of the language or the age at which you start learning it, the order in which you learn the words will have an impact on their processing, with an advantage for what was acquired first.

Conclusion

In this eye tracking mega study of bilingual reading, we confirmed that L2 AoA also influences L2 natural reading. The AoA effect is however not only determined by the age at which the word was learned in L2, but also to a lesser extent by the age at which its translational equivalent was learned in L1. As the semantic and mapping hypotheses are not mutually exclusive, we propose an integration between these two to account for these results.

REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. http://doi.org/10.1016/j.jml.2007.12.005

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. http://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. http://doi.org/10.3758/BF03193014

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Barry, C., Johnston, R. A., & Wood, R. F. (2006). Effects of age of acquisition, age, and repetition priming on object naming. *Visual Cognition*, *13*(7–8), 911–927. http://doi.org/10.1080/13506280544000101

Belke, E., Brysbaert, M., Meyer, A. S., & Ghyselinck, M. (2005). Age of acquisition effects in picture naming: evidence for a lexical-semantic competition hypothesis. *Cognition*, *96*, B45-54. http://doi.org/10.1016/j.cognition.2004.11.006

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2001). Age-of acquisition and word frequency in the lexical decision task: Further evidence from the French language. *Cahiers de Psychologie Cognitive - Current Psychology of Cognition*, *20*(6), 401–443.

Brysbaert, M. (n.d.). Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings “corrected” for frequency. *Quarterly Journal of Experimental Psychology*.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related , partly frequency independent. *Visual Cognition*, *13*, 992–1011. http://doi.org/10.1080/13506280544000165

Brysbaert, M., Lange, M., & Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*(1), 65–86.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. http://doi.org/10.1016/j.actpsy.2014.04.010

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*, 215–226.

Cardimona, K., Smith, P., & Roberts, L. S. (2015). Lexical Organization in Second Language Acquisition: Does the Critical Period Matter? *TESOL Journal*, *7*(3), 540–565. doi:10.1002/tesj.219

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, *25*(1), 85–95. http://doi.org/10.1080/14640747308400325

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (n.d.). Presenting GECO : An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*. http://doi.org/10.3758/s13428-016-0734-0

Cop, U., Dirix, N., Van Assche, E., Drieghe, D., & Duyck, W. (n.d.). Bilinguals reading a novel: An eye movement study of cognate facilitation in L1 and L2 natural reading. *Bilingualism: Language and Cognition*. http://doi.org/10.1017/S1366728916000213

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. http://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. http://doi.org/10.3758/s13423-015-0819-2

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *The Quarterly Journal of Experimental Psychology*, *60*(8), 1072–1082.

http://doi.org/10.1080/17470210701315467

Cortese, M. J., & Schock, J. (2012). Imageability and age of acquisition effects in disyllabic word recognition. *The Quarterly Journal of Experimental Psychology*, *66*(5), 946–972. http://doi.org/10.1080/17470218.2012.722660

CRR. (2014). wordORnot. Retrieved September 1, 2015, from http://crr.ugent.be/archives/1533

De Moor, W., Ghyselinck, M., & Brysbaert, M. (2000). A validation study of the age-of-acquisition norms collected by Ghyselinck, De Moor, & Brysbaert. *Psychologica Belgica*, *40*(2), 99–114.

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. http://doi.org/10.1017/S1366728902273010

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103–1123. http://doi.org/10.1037/0278-7393.26.5.I103

Foote, R. (2014). Age of Acquisition and Sensitivity to Gender in Spanish Word Recognition. *Language Acquisition, 21*(4), 365–385. doi:10.1080/10489223.2014.892948

Gerhand, S., & Barry, C. (1999a). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, *27*(4), 592–602. http://doi.org/10.3758/BF03211553

Gerhand, S., & Barry, C. (1999b). Age-of-acquisition and frequency effects in speeded word naming. *Cognition*, *73*(2), B27–B36. http://doi.org/10.1016/S0010-0277(99)00052-9

Ghyselinck, M., Custers, R., & Brysbaert, M. (2004). The Effect of Age of Acquisition in Visual Word Processing: Further Evidence for the Semantic Hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 550–554. http://doi.org/10.1037/0278-7393.30.2.550

Hirsh, K. W., Morrison, C. M., Gaset, S., & Carnicer, E. (2003). Age of acquisition and speech production in L2. *Bilingualism: Language and Cognition*, *6*, 117–128. http://doi.org/10.1017/S136672890300107X

Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica*, *23*, 245–281.

Izura, C., & Ellis, A. W. (2004). Age of acquisition effects in translation judgement tasks. *Journal of Memory and Language*, *50*, 165–181. Retrieved from http://ac.els-cdn.com/S0749596X03001165/1-s2.0-S0749596X03001165-main.pdf?_tid=19b3bb46-5789-11e5-9bbe-

00000aab0f26&acdnat=1441868385_cc6f17a2dd58a1c4b65b15075c5a815b

Izura, C., Pérez, M. a., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. W. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of Memory and Language*, *64*, 32–58. http://doi.org/10.1016/j.jml.2010.09.002

Johnston, R. a., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*, 789–845. http://doi.org/10.1080/13506280544000066

Joseph, H. S. S. L., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition*, *133*(1), 238–48. http://doi.org/10.1016/j.cognition.2014.06.015

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*(5), 684–712. http://doi.org/10.1037/0033-2909.131.5.684

Juhasz, B. J., Gullick, M. M., & Shesler, L. W. (2011). The Effects of Age-of-Acquisition on Ambiguity Resolution : Evidence from Eye Movements. *Journal of Eye Movement Research*, *4*(1), 1–14.

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1312–1318. http://doi.org/10.1037/0278-7393.29.6.1312

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7–8), 846–863. http://doi.org/10.1080/13506280544000075

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. http://doi.org/10.3758/BRM.42.3.643

Kroll, J. F., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections Between Bilingual Memory Representations. *Journal of Memory and Language*, *33*(2), 149–174. http://doi.org/10.1006/jmla.1994.1008

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. http://doi.org//10.1080/17470218.2012.658820

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990. http://doi.org/10.3758/s13428-012-0210-4

Lambon Ralph, M. a., & Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. *Visual Cognition*, *13*(7–8), 928–948. http://doi.org/10.1080/13506280544000110

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–43. http://doi.org/10.3758/s13428-011-0146-0

Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three Misconceptions about Age and L2 Learning. *TESOL Quarterly*, *34*, 9–34. http://doi.org/10.2307/3588095

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model of Context Effects in Letter Perception: Part I. An Account of Basic Findings. *Psychological Review*, *88*, 375–407. http://doi.org/10.1016/B978-1-4832-1446-7.50048-0

Menenti, L., & Burani, C. (2007). What causes the effect of age of acquisition in lexical processing? *The Quarterly Journal of Experimental Psychology*, *60*(5), 652–660. http://doi.org/10.1080/17470210601100126

Monaghan, P., & Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language*, *63*(4), 506–525. http://doi.org/10.1016/j.jml.2010.08.003

Montrul, S., & Foote, R. (2014). Age of acquisition interactions in bilingual lexical access: A study of the weaker language of L2 learners and heritage speakers. *International Journal of Bilingualism*, *18*(3), 274–303. doi:10.1177/1367006912443431

Morrison, C. M., & Ellis, A. W. (1995). Roles of Word Frequency and Age of Acquisition in Word Naming and Lexical Decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 116–133.

Morrison, C. M., Ellis, a W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, *20*(6), 705–714. http://doi.org/10.3758/BF03202720

Pérez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative word frequency and frequency trajectory are controlled. *The Quarterly Journal of Experimental Psychology*, *60*(1), 32–42. http://doi.org/10.1080/17470210600577423

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

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. http://doi.org/10.1080/17470210902816461

Stewart, N., & Ellis, A. W. (2008). Order of acquisition in learning perceptual categories: a laboratory analogue of the age-of-acquisition effect? *Psychonomic Bulletin & Review*, *15*(1), 70–4. http://doi.org/10.3758/PBR.15.1.70

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, *29*(1), 41–78. http://doi.org/10.1207/s15516709cog2901_3

Van Hell, J. G., & De Groot, A. M. B. (1998). Disentangling context availability and concreteness in lexical decision and word translation. *The Quarterly Journal of Experimental Psychology Section A, 51*(1), 41–63. http://doi.org/10.1080/713755752

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2013). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. http://doi.org/10.1080/17470218.2013.850521

Wilson, M. A., Cuetos, F., Davies, R., & Burani, C. (2013). Revisiting age-of-acquisition effects in Spanish visual word recognition: the role of item imageability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1842–1859. http://doi.org/10.1037/a0033090

CHAPTER 4

CROSS-LINGUAL NEIGHBORHOOD EFFECTS IN GENERALIZED LEXICAL DECISION AND NATURAL READING.[1]

The present study assessed intra- and cross-lingual neighborhood effects, using both a generalized lexical decision task and an analysis of a large-scale bilingual eye-tracking corpus (Cop, Dirix, Drieghe & Duyck, in press). Using new neighborhood density and frequency measures, the general lexical decision task yielded an inhibitory cross-lingual neighborhood density effect on reading times of second language words, replicating van Heuven, Dijkstra and Grainger (1998). Reaction times for native language words were not influenced by neighborhood density or frequency but error rates showed cross-lingual neighborhood effects depending on target word frequency.

The large-scale eye movement corpus confirmed effects of cross-lingual neighborhood on natural reading, even though participants were reading a novel in a unilingual context. Especially second language reading and to a lesser extent native language reading were influenced by lexical candidates from the non-target language, although these effects in natural reading were largely facilitatory.

These results offer strong and direct support for bilingual word recognition models that assume language-independent lexical access.

[1] 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. doi: 10.1037/xlm0000352.

INTRODUCTION

During written word recognition, we are faced with the complex task of activating and identifying the correct lexical representation among a large group of orthographically similar, but not identical, representations. The term orthographic *neighbor,* coined by Coltheart, Davelaar, Jonasson, and Besner (1977), is used to denote such a similar word. Coltheart et al.'s (1977) definition of such a neighbor is any word that can be created by changing one letter of the target word while preserving letter positions (example: *house* is a neighbor of the word *horse*; see also Landauer & Streeter, 1973). We will refer to this kind of neighbor as a s*ubstitution* neighbor from now on. Most studies examining neighbor effects used this definition. The number of neighbors of a particular target word is called the *neighborhood density* (N density).

In the word recognition literature, most models of (monolingual) word recognition hypothesize that a written word activates a set of possible lexical candidates. This means that at some point the correct target word has to be selected out of a number of neighbors. The search model (Forster, 1976) and the activation verification model (Paap, Newsome, McDonald, & Schvaneveldt, 1982) both predict that the neighborhood density will affect language performance because the actual decision is established by a frequency-ordered lexical search within those candidates. The longer the list of neighbors, the longer it would take to select the correct representation. Another influential model of word recognition, the interactive activation model (IA model, McClelland & Rumelhart, 1981), also makes the prediction that the number of activated candidates should affect lexical access, but proposes that the reason for this is lateral inhibition. In the IA model, word identification starts with letter identification. These letters feed forward activation to lexical candidates. Each of these representations has a resting level of activation, which is determined by the frequency of the word. The activated representations feed activation backwards to the letter level. Word recognition is the end result of a competitive process between the activated lexical candidates, each inhibiting the others activation. The representation whose activation level first rises significantly above the identification threshold, is selected.

The most intuitive hypothesis formed by the IA model is that words with more orthographically similar lexical items would receive more lateral inhibition from these neighbors and this would slow lexical access to the target word (e.g. Grainger & Jacobs, 1993). On the other hand, a facilitative effect of a larger neighborhood is also not impossible within the IA model. More neighbors could cause greater overall excitation in the lexicon, which could help in specific tasks like the lexical decision task (e.g. Andrews, 1997; Grainger & Jacobs, 1996). Also, the feedback activation of multiple lexical candidates to particular letters, again activating the target representation, could facilitate activation of the correct lexical representation, so that large neighborhoods could again speed up word recognition in some instances. In the IA model, word frequency determines the resting level activation of representations, and lateral inhibition between the activated lexical candidates belonging to the orthographic neighborhood is also a function of their frequency. For this reason, it could be expected that recognition of low frequent words would show larger effects of neighborhood density and frequency. This is because a low frequent representation will need more time to accumulate enough activation to significantly rise above the activation levels of the higher frequency neighbors, thus delaying lexical access to the target word.

Within the IA architecture, precise predictions about the time course of neighborhood effects and whether the combination of these counteracting effects would result in facilitation or inhibition of recognition are difficult to make. With its complex interactions between parallel activation of letters and words and lateral inhibition among words, the IA model can account for a lot of different effects. Indeed, model simulations of the IA model have shown that effects can be both inhibitory (Jacobs & Grainger, 1992) and facilitatory (Coltheart & Rastle, 1994; Pollatsek, Perea, & Binder, 1999) depending on stimulus materials and small adjustment to the parameters of the model. For a more detailed discussion on this matter, we refer to Grainger and Jacobs (1996). Their Multiple Read-Out model allowed simulations of both inhibitory and facilitatory effects, based on multiple response criteria. As we will see below, empirical investigations of neighborhood effects have also yielded a complicated mix of findings, with multiple moderating variables. This mimics the complicated pattern of neighborhood effect simulations that the computational models may exhibit.

Monolingual neighborhood effects

Isolated word studies. In the empirical search for neighborhood effects, mainly two variables have been manipulated. The first one is the neighborhood density. Coltheart et al. (1977) were the first to show neighborhood density effects for isolated word recognition. In a lexical decision task, they found inhibitory effects for non-words with increasing neighborhood density, but no effects for words (see also Holcomb, Grainger, & O’rourke, 2002). After this, multiple authors investigated the effects of neighborhood density on lexical decision word performance. As Andrews (1997) argued in a review paper, large neighborhoods are almost always associated with better performance in standard lexical decision tasks. Indeed, most of these experiments pointed towards a facilitatory effect of increasing neighborhood size, for both the speed and accuracy of lexical decision (Andrews, 1989, 1992; Carreiras, Perea, & Grainger, 1997; Forster & Shen, 1996; Huntsman & Lima, 2002; Johnson & Pugh, 1994; Laxon, Coltheart, & Keating, 1988; Perea & Rosa, 2000; Pollatsek et al., 1999; Sears, Campbell, &

Lupker, 2006; Sears, Hino, & Lupker, 1995; for an additional review see Mathey, 2001). Similar results were found for naming (Peereman & Content, 1995; Sears et al., 1995) and semantic categorization tasks (Carreiras et al., 1997; Forster & Shen, 1996). Perceptual identification tasks have shown mixed results. Carreiras et al. (1997) reported slower reaction times for words with a large neighborhood density, whereas Snodgrass and Mintzer (1993) found a null effect in their Experiment 1, facilitation in Experiment 2 and inhibition in Experiment 3, 4 and 5. Andrews (1997) concluded that inhibitory effects of large neighborhoods observed for perceptual identification tasks are the result of unusual stimulus environments or elaborate guessing strategies.

Another neighborhood measure that is used regularly is whether the target word has a more frequent neighbor or not. We will refer to this factor as *neighborhood frequency* (N frequency). In lexical decision tasks it is usually found that reaction times are longer and accuracy is lower when a more frequent neighbor is present (Carreiras et al., 1997; Davis & Taft, 2005; Grainger, 1990; Grainger & Jacobs, 1996; Grainger, Oregan, Jacobs, & Segui, 1992; Grainger, O'regan, Jacobs, & Segui, 1989; Grainger & Segui, 1990; Huntsman & Lima, 1996; Perea & Pollatsek, 1998). This effect is also present for perceptual identification tasks (Carreiras et al., 1997; Grainger & Jacobs, 1996; Grainger & Segui, 1990).

Although research on neighborhood effects has predominantly used isolated word tasks, such as lexical decision tasks and naming tasks, there is some debate as to whether these tasks capture the cognitive processes underlying lexical access (e.g., Balota & Chumbley, 1984; Rayner & Pollatsek, 1989). The most important argument is that isolated word tasks entail a decision component or behavioral response, decreasing the validity of the measure (e.g., Paap & Johansen, 1994; Rayner & Liversedge, 2011, Snodgrass & Mintzer, 1993). Because of this decision component, and specifically in the case of neighborhood effects, the lexical decision task is for instance insensitive to the cases where the participant makes a response to the more frequent neighbor of the target word instead of the target itself and still responds with a correct "Yes" answer.

Kuperman, Drieghe, Keuleers and Brysbaert (2013) indeed showed that the lexical decision task and a more natural reading method, i.e. sentence reading in context are distinguishable and measure, to a large extent at least, different language processes. They found that lexical decision reaction times only explained 5-17% of the variance in gaze durations on target words embedded in sentences after partialling out the effects of word frequency and word length. This dropped to 0.2% of the variance in fixation durations in natural reading when not only the target words, but all words in the sentences are analyzed.

Eye tracking studies. It becomes clear that a more natural reading task, like sentence reading monitored by an eye tracker, could produce measures that are a closer approximation of natural language processes. Eye tracking can be used to assess the time that the eyes remain fixated on a word and thus provide more direct evidence for the existence of neighborhood influence on lexical access. In the case of neighborhood effects, eye tracking can be especially useful because it has a very high temporal resolution. This allows a specific investigation of the time course of potential N effects. Indeed, some eye movement measures (such as single fixation durations) reflect early stages in visual word recognition, whereas others like total reading time reflect higher-order language processes such as semantic integration. Eye tracking thus allows the study of language processing through multiple dependent variables reflecting several stages of word recognition, whereas the lexical decision task only allows investigation of reaction times and accuracy scores. Eye tracking during natural reading should therefore contribute to the study of cross-lingual neighborhood effects above and beyond lexical decision results.

So far, only a handful of studies investigated neighborhood effects in sentence reading using eye tracking. Only one of those investigated the effect of neighborhood density (Pollatsek, Perea, & Binder, 1999). In Experiment 2 of Pollatsek et al. (1999), English monolinguals read target embedded sentences for comprehension. Half of the target words had a lot of neighbors (average= 8.5), the other half few (average=2.2). All of these targets had at least one more frequent neighbor. Their first analysis showed an inhibitory effect of neighborhood density for gaze duration and total reading time. Because in this analysis the number of neighbors was confounded with the number of more frequent neighbors, Pollatsek et al. conducted another analysis, in which they held the number of more frequent neighbors constant. Under these conditions, they found that words with more low frequent neighbors were skipped more often, but these words were also regressed to more often. The authors noted that the facilitatory effect on skipping rates might be due to initial misidentification of the target word. However, they did find a facilitatory effect in gaze durations that could not be due to such misidentification because it was stronger in the sentences where the highest frequent neighbor was implausible in the sentence context.

Perea and Pollatsek (1998) conducted another reading study, this time investigating the effect of neighborhood frequency. In their Experiment 2 they instructed English monolingual participants to read sentences for comprehension. The embedded target words in these sentences were matched on number of neighbors. Half of the target words had an orthographic neighbor with a higher word frequency and the other half did not. The results showed more regressions towards the target word when it had a higher frequency neighbor than when it did not. Also, spillover effects were larger when the target word had a more frequent neighbor. These effects were larger for low frequent target words. Davis, Perea, and Acha (2009) and Slattery (2009) conducted similar reading studies and confirmed that inhibitory effects of neighborhood frequency might occur late in the reading process. Davis et al. (2009) found an inhibitory effect of neighbor frequency for gaze durations and total reading time. Although there were also more regressions toward words with a more frequent neighbor, this effect was not significant. Slattery (2009) found an inhibitory effect of the presence of a more frequent neighbor in a sentence-reading task. More regressions were made and the total reading time was longer when the target word had a more frequent neighbor. He pinpointed this effect on the initial misidentification of the target word, by showing that these effects are no longer present when the more frequent word is not compatible with the prior sentence context. However, Sears et al., (2006) failed to find similar neighborhood frequency effects in an extensive set of reading experiments. They concluded that, at least in English, neighborhood frequency has no direct effect on reading times and has little to no effect on post-identification processes.

It becomes clear that all previous experiments examining neighborhood effects, either in isolated word studies or eye tracking studies, have focused on one of the two neighborhood variables, density or frequency, while holding the other one constant. It is not clear what the net result would be of either variable in natural reading when both vary simultaneously.

Bilingual Neighborhood Effects

In the field of bilingualism, one of the most important questions has been whether word recognition involves activation of lexical candidates from the non-target language. This question is tied in with the architecture of the bilingual lexical models, which may have one integrated, or two separate lexicons. Some have argued that lexical access for bilinguals is language-selective, meaning that when reading one language, only representations of that language are activated (Caramazza & Brones, 1979; Scarborough, Gerard, & Cortese, 1984). More recently however, a consensus has evolved in the literature that word recognition involves cross-lingual activation for bilinguals (for an overview see Dijkstra, 2007). The evidence for this mechanism comes mostly from studies using words that share features across two languages, such as inter-lingual homographs (words sharing orthography but not meaning across languages) and cognates. The latter are translation equivalent words that not only overlap in meaning but also in orthography (example of an identical cognate is the word “piano” in English and in Dutch). Cognates are recognized faster and more accurately than control words in behavioral studies that present words in isolation, such as lexical decision tasks (Bultena, Dijkstra, & van Hell, 2013; Dijkstra, Grainger, & van Heuven, 1999; Dijkstra, Miwa, Brummelhuis, Sappelli, & Baayen, 2010; Duyck, Van Assche, Drieghe, & Hartsuiker, 2007;

Peeters, Dijkstra, & Grainger, 2013; Van Assche, Drieghe, Duyck, Welvaert, & Hartsuiker, 2011). Similarly, cognate effects have been observed when bilinguals read text (Duyck et al, 2007; Van Assche et al. 2011; Van Assche, Duyck, & Brysbaert, 2013), even in the native language (Van Assche, Duyck, Hartsuiker, & Diependaele, 2009). This is remarkable because the language of a running text might serve as a useful cue in restricting access to the target language and therefore could speed up word recognition in this way (Altarriba, Kroll, Sholl, & Rayner, 1996; Van Assche, Duyck, & Hartsuiker, 2012).

Generally, these cognate effects are attributed to spreading activation between representations of both languages. Alternatively, because cognates share the exact same orthography and almost exact phonology and semantics, it has been argued that identical cognates could have a single representation across languages (see Dijkstra et al., 2010). This is important, because there is only very indirect evidence that cognates would actually be represented separately for each language, which is necessary for an explanation in terms of cross-lingual lexical activation. So, a more conservative test of cross-lingual lexical activation would be one in which representations that are certainly language-specific, such as neighbors, interact with other language-specific representations of the other language.

This is why the most compelling evidence for cross-lingual lexical access would come from cross-lingual neighborhood effects in bilingual reading. However, there is only study so far providing such evidence (van Heuven, Dijkstra, and Grainger, 1998). In this study, Dutch-English bilinguals performed a blocked and mixed-progressive demasking task, a generalized lexical decision task and an English lexical decision task. Four item conditions were constructed by orthogonally manipulating the number of English and Dutch substitution neighbors in the CELEX database of the target words. In the progressive demasking task, participants had to identify four-letter words that gradually appeared on a screen as fast as possible. In the blocked version of the task, the experiment consisted of two blocks, one containing only L1 words, the other containing only L2 words. Both in the English and Dutch block of the progressive demasking task, van Heuven et al. found an inhibitory effect of non-target N density, but this effect only reached full significance in the L2 block. In the mixed progressive demasking task, L1 and L2 words were presented in a random order. Here the authors expected to find larger effects, because in a mixed language setting, both languages have to be active to perform the task. In this experiment, inhibition from the non-target neighbors was found for English and Dutch items. In the generalized lexical decision task, participants had to decide as fast and accurately as possible whether the target stimulus was a word (Dutch or English) or not. For the generalized lexical decision task, van Heuven et al. again found inhibition of Dutch N and facilitation for English N for reaction times to the English items. No neighborhood effects were found for the Dutch items. In the English lexical decision task, monolingual and bilingual participants had to decide whether the presented stimulus was an English word or not. Here, again an inhibitory effect of Dutch N was found, showing that cross-lingual activation is not limited to mixed language contexts. All of these results were taken as evidence that words automatically activate substitution neighbors both pertaining to the target and non-target language. Although van Heuven et al.'s (1998) results (nor design) were never directly replicated, two ERP studies supported the existence of cross-lingual N density effects, by showing a more negative N400 ERP component for words with more cross-lingual neighbors (Grossi, Savill, Thomas, & Thierry, 2012; Midgley, Holcomb, van Heuven, & Grainger, 2008). It is interesting to note that van Heuven et al. (1998) did not find any effects of cross-lingual N density in a blocked or selective L1 setting. Because this is the only study reporting cross-lingual N effects, so far there has been no direct evidence of cross-lingual activation of neighbors in L1 reading in a purely unilingual context. The present study will assess such an effect in bilingual natural reading.

Also, the effect of cross-lingual N frequency has never been investigated. In the monolingual literature, it is clear that the presence of a more frequent neighbor influences reaction times and error rates in lexical decision tasks (e.g. Carreiras et al., 1997; Davis & Taft, 2005; Grainger & Jacobs, 1996; Grainger, 1990; Grainger, O'regan, Jacobs, & Segui, 1989; Perea & Pollatsek, 1998). Also, several studies provided evidence for an important role of this factor in N density effects (Carreiras et al., 1997; Davis & Taft, 2005; Grainger & Jacobs, 1996; Perea & Pollatsek, 1998). We will address this issue in a bilingual context.

BIA+ model

The findings on cross-lingual activation in bilingual reading described above have led to the development of the BIA+ model (Dijkstra & van Heuven, 2002) This model is the successor of the original BIA model (Dijkstra & van Heuven, 1998), which is a bilingual adaptation of the Interactive Activation model (McClelland & Rumelhart, 1981). The BIA+ model is a language non-selective model of lexical access, which entails an integrated bilingual lexicon (see Figure 1). To account for differences in word recognition depending on tasks and other non-linguistic variables (for example instructions and expectations of the participants) the BIA+ model consists of a word identification system and a task/decision system. Like in the (B)IA model, a set of orthographic candidates is activated through bottom-up activation when a written word is encountered. Depending on their similarity to the printed word and their resting-level activation, determined by the word frequency, these representations are partly activated. As L2 items tend to be lower in subjective frequency for unbalanced bilinguals, their representations are activated somewhat slower than L1 items. The activation spreads from the orthographic candidates to the connected phonological and semantic representations. Every word in the lexicon is connected to one of the available language nodes, representing the language membership of that word. In BIA+, these nodes also represent the global lexical activity of a language. These nodes do not feed activation back to the orthographic or phonological level so they cannot function as a language selection mechanism. This architecture for the bilingual lexicon implies that both intra- and cross-lingual orthographic neighbors should prominently influence lexical activation during visual word recognition.

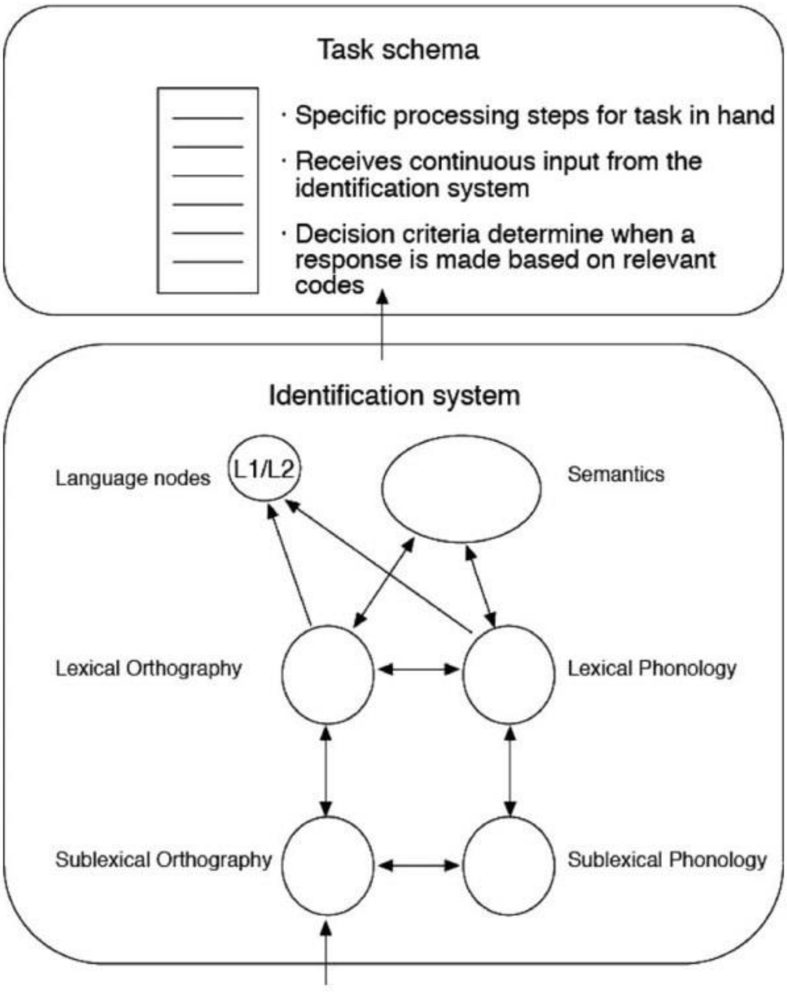

Figure 1. Schematic of the BIA+ model (taken from Dijkstra & van Heuven, 2002)

Unfortunately, there are no simulations of neighborhood effects within the BIA+ model. However, because the BIA+ was then not formulated yet, Dijkstra, van Heuven, and Grainger (1998) and van Heuven et al. (1998) explained cross-lingual neighborhood effects using simulations of their results in the BIA model. Because BIA+ is basically the combination of the orthographic system of the BIA model with new (non-implemented) task-scheme, phonology and semantic systems, and because neighborhood effects mainly rely on orthographic representations, these BIA simulations remain very informative about how BIA+ would model such effects. Dijkstra et al. (1998) operationalized simulated data as the amount of cycles the model needed to run for each item. The average amount of cycles for each condition (target language * N language * N density) could then be compared to the reaction time means of experimental data. Dijkstra et al. determined the degree of the correspondence between simulated and experimental data by qualitative (visual inspection of the response patterns) and quantitative (chi-square tests) measures. For each of the experimental conditions, the authors reported a reliable model fit on each measure. In the BIA simulations, inhibition from L1 neighbors for L1 word recognition is explained by the mechanism of lateral inhibition on the lexical level. Words with more neighbors suffer from the inhibition of their co-activated neighbors, thus taking longer to reach the identification threshold. The facilitation of within-language (L2) N density for bilinguals in English is explained by the relative activation of the two languages depending on word frequency in combination with asymmetric top-down inhibition from the language nodes implemented in the BIA-model. More specifically, the co-activated Dutch neighbors of the English word will exert inhibition on the target word through the Dutch language node. van Heuven et al. argue that inhibition will be larger towards words with a small compared to a larger N density, creating a relative facilitation effect for words with a larger within-language N density. Finally, inhibition from non-target language neighbors (both in L1 and L2) is again explained by lateral inhibition. As words from different languages are integrated in one lexicon in the BIA model, the co-activated cross-lingual neighbors also inhibit the target word. The similarity between simulation outcomes and experimental data proved to be quite high, as these authors reported that there was no difference between the two on statistical tests.

The facilitatory effect of within-language N density on L2 English word recognition in bilinguals was also found by van Heuven et al. (1998) for L1 English monolinguals. To further explain these results, they refer to Grainger and Jacobs (1996), who showed that facilitatory effects of large N could be simulated with the IA model with the help of read-out criteria. This means that Dijkstra et al. (1998) explained the monolingual and bilingual facilitatory N density effects in English in two different ways, which is not very parsimonious. Another challenge for Dijkstra et al.'s interpretation is that the top down activation from language nodes is not implemented in the BIA+ model. As such, it is unclear how the authors would explain the facilitatory effects of target and non-target N density within the BIA+ model.

Another complicating factor is that it has become clear that defining N densities by only including substitution neighbors is insufficient. For example, Davis et al. (2009) found an additional effect of *addition neighbors* (by adding a letter to a word, e.g., *frog* is an addition neighbor of *fog*) and *deletion neighbors* (by deleting a letter from a word, e.g., *rash* is a deletion neighbor of *trash*) above and beyond the effect of substitution neighbors. Word recognition models with fixed letter positions such as the IA and BIA+ have problems explaining these effects, because in these models lexical competition only occurs between representations of identical word length (see Davis & Bowers (2006) for an overview). Alternatively, there are monolingual models of word recognition with a relative positional nature that can account for effects of addition and deletion neighbors (e.g., the SOLAR model, Davis & Bowers, 2004; the SERIOL model, Whitney, 2001; the Overlap model, Gomez, Ratcliff, & Perea, 2008). In the study by van Heuven et al. (1998), the N densities were calculated by counting the number of Dutch and English substitution neighbors of the target word using the CELEX database (Baayen et al., 1993). We might get a more accurate picture of cross-lingual N effects when we include addition and deletion neighbors in the N density measure. This new measure might be more sensitive in detecting cross-lingual effects in L1, which did not show very strong effects in van Heuven et al.

To conclude, despite the development of the BIA+ model and the abundance of papers addressing other effects of cross-lingual activation in visual word recognition, such as cognate effects (e.g. Bultena et al., 2013; Dijkstra et al., 1999; Duyck et al., 2007, Peeters et al., 2013; Van Assche, et al. 2011), there has been only one behavioral study that has provided direct evidence for parallel activation of lexical representations in bilingual word recognition by showing neighborhood effects. This study by van Heuven et al. (1998) used lexical decision instead of natural reading, and found no indications of L2 activation during pure L1 reading. We will therefore begin by attempting to replicate van Heuven et al.'s generalized lexical decision task, both using their categorization of stimuli and an optimized N density measure, including addition and deletion neighbors. Next, we will investigate whether these cross-lingual N effects are present in a large database of bilingual eye movements of natural reading (Cop et al., in press) of parallel access to target language and non-target language representations of the bilingual lexicon. This conservative test, in which unilingual running text is read, assess the generalizability of the cross-lingual effects obtained in experimental conditions with isolated words.

EXPERIMENT 1

In Experiment 1 we attempted to replicate the generalized lexical decision task of van Heuven et al. (1998), using the exact same stimuli as them to investigate cross-lingual N density effects in a new group of bilingual Dutch-English participants. Based on their findings, we expect within-language facilitation and cross-language inhibition for L2 reading and only a small within-language inhibitory effect and no cross-lingual effect for L1 reading.

We will present linear mixed effects analyses including English and Dutch N frequency variables. By using a more inclusive measure of N density, we expect to find stronger effects of N density for L2 words and we might detect cross-lingual effects for L1 words.

Because of the architecture of the BIA+ model we expect larger effects of N density and frequency for low frequent target words (Perea & Pollatsek, 1998; Williams, Perea, Pollatsek, & Rayner, 2006). Low frequent representations in general need more time to accumulate sufficient activation to rise above the threshold of activation than high frequent ones, so that they can benefit more from (or be hindered by) their neighbors.

Method

Participants. Thirty undergraduates received course credit for their participation in this experiment (19 females, 27 right-handed, M_{age} = 19.07 [2.08]). All students were unbalanced Dutch-English bilinguals. Participants were tested for language proficiency with the Dutch and English version of the LexTALE (Lexical Test for Advanced learners of English, Lemhöfer & Broersma, 2012) and a self-report questionnaire (see Table B1 in Appendix 4B for detailed proficiency scores). For the questionnaire, participants rated how good they were at listening, speaking, reading and writing in both languages on a 5-point Likert scale.

Materials. The 160 words (80 Dutch and 80 English) and 160 nonwords were identical to those of Experiment 3 of van Heuven et al. (1998) (see Table 1 for word characteristics; see Appendix 4C for all stimuli). Importantly, we updated the N densities of the stimuli. It has become clear that N densities are inconsistently identified in the literature (Marian et al., 2012), so that researchers use different language databases to determine how large the neighborhoods of their stimuli are. This makes it difficult to compare results across experiments. To overcome this problem, Marian et al. (2012) developed the CLEARPOND database (Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities), which provides N densities and also allows comparing N densities across languages by including comparable corpora of multiple languages. When using this database to calculate N densities, we may replace the dichotomous neighborhood density classification that van Heuven et al. (1998) made with a more sensitive measure. In the current study, we used CLEARPOND (Marian et al., 2012) to determine a more accurate N density and frequency value, including within and cross-language substitution, addition and deletion neighbors. Furthermore, we calculated some additional word characteristics because they were not provided in the original study (e.g., bigram frequency) or because more up-to-date, and improved, measures exist nowadays (e.g., SUBTLEX frequencies, (SUBTLEX-NL, Keuleers, Brysbaert, & New, 2010;

Table 1. *Descriptive Statistics for the Stimuli Used in Experiment 1 by language and neighborhood density (standard deviations between parentheses).*

	Neighbors[a]		Number of neighbors[b]		Higher frequent neighbor[c]		Word Frequency[d]	Average Bigram Frequency[e]	CLD[f]
	Dutch	English	Dutch	English	Dutch	English			
Dutch	Large	Large	7 (2.49)	7.1 (4.17)	.85 (.36)	.90 (.30)	2.257 (0.61)	1828.47 (854.07)	0.30 (0.24)
	Large	Small	7.95 (2.94)	3.95 (2.8)	.90 (.30)	.75 (.43)	2.457 (0.86)	2533.67 (1665.93)	0.25 (0.24)
	Small	Large	4.05 (2.5)	6.6 (2.78)	.85 (.36)	.90 (.30)	2.364 (0.92)	1947.32 (961.72)	0.38 (0.27)
	Small	Small	3.45 (2.27)	4.6 (4.46)	.65 (.48)	.75 (.43)	2.368 (0.45)	2194.75 (1227.3)	0.26 (0.25)
English	Large	Large	5.35 (2.8)	8.15 (3.44)	.70 (.46)	.90 (.30)	3.576 (0.57)	1370.93 (541.67)	0.36 (0.23)
	Large	Small	2.15 (1.71)	8.3 (3.69)	.50 (.50)	.80 (.40)	3.758 (0.35)	1300 (608.32)	0.29 (0.27)
	Small	Large	5.9 (6.2)	5.4 (2.58)	.70 (.46)	.55 (.50)	3.434 (0.65)	1324.74 (668.89)	0.30 (0.33)
	Small	Small	1.9 (1.7)	4.15 (2.85)	.30 (.46)	.50 (.50)	3.505 (0.62)	1282.15 (653.78)	0.26 (0.27)
Nonwords	Large	Large	5.675 (2.59)	6.2 (2.94)					
	Large	Small	4.975 (2.19)	3.675 (2.41)					
	Small	Large	3.125 (2.27)	6.475 (3.14)					
	Small	Small	2.35 (1.92)	3.375 (1.84)					

[a]N densities as defined by van Heuven et al. (1998); [b]Total CLEARPOND N densities (Marian, Bartolotti, Chabal, & Shook, 2012); [c] The proportion of words with a higher frequent Neighbor, [d] Log10 Subtlex frequencies: SUBTLEX-NL for Dutch words (Keuleers et al., 2010), SUBTLEX-UK for English words (van Heuven et al., 2013); [e] Summated bigram frequencies (calculated using WordGen, (Duyck, Desmet, Verbeke, & Brysbaert, 2004) were normalized for corpus size and then divided by word length to obtain average bigram frequencies. Bigram frequencies could not be calculated for the nonwords: since van Heuven et al. (1998) didn't specify which of the nonwords were matched with which language, we couldn't determine which language corpus to use to calculate bigram frequencies; [f] Corrected Levenshtein distance was calculated as a measure of orthographic overlap with the formula in Appendix 4A by comparing the word with its closest translation in NIM (Guasch, Boada, Ferré, & Sánchez-Casas, 2013).

SUBTLEX-UK, van Heuven, Mandera, Keuleers, & Brysbaert, 2013) instead of CELEX (Baayen, Piepenbrock, & Van Rijn, 1993) word frequencies).

Each participant saw each stimulus once, which resulted in 320 trials. All stimuli were presented in black against a white background. The font was Courier New, size 18 bold. Instruction language (Dutch or English) and response mapping (pressing the left button for a word, right for a nonword or vice-versa) were counterbalanced across participants.

The experiment was programmed in E-Prime 2.0 software (Schneider, Eschman, & Zuccolotto, 2012). Stimuli were presented on a Benq XL2411Z 24 inch LED monitor. The computer used for the experiment was a Dell Optiplex 3020 mini-tower with a 3.2GHz Intel Core i5-4570 processor. Participants had to respond by pressing left and right buttons on a RB-730 Cedrus responsebox.

Procedure. The procedure as was based on Experiment 3 of van Heuven et al. (1998). Participants were seated in a comfortable chair at approximately 45-60cm from the screen. All instructions were presented on the screen. They were told they had to judge whether a presented letter string was either a word (in Dutch or English) or a non-word by pressing the according button. They were instructed to decide as quickly and accurately as possible. After the instructions, participants had to perform a practice block with 10 trials (five words and non-words each), which was repeated if their accuracy was below 80%. None of the stimuli used in the practice block were used in the experimental block. Afterwards the experimental block followed, with a presentation of the stimuli in a pseudo-random order (no more than four consecutive words or nonwords were presented). Halfway the experiment participants could take a short break.

A trial started with the presentation of a fixation cross (800ms), followed by a blank screen of 300ms. The stimulus was presented until the participant responded or for a maximum duration of 2500ms. The inter-trial interval was kept constant at 700ms.

After finishing the experiment, participants were presented with the English and Dutch version of the LexTALE and the self-reported questionnaire. The entire session lasted about 45 minutes.

Results

All analyses were performed in R version 3.1.2 (R Core Team, 2014). Models were fitted using the lme4-package in R (Bates, Mächler, Bolker, & Walker, 2014)[2]. Several predictors were included in the analysis. Word frequency was included because of its indisputable role in (bilingual) visual word recognition (Baayen et al., 2006; Keuleers et al., 2010). Bigram frequency was added because word characteristics showed that there was a lot of variation for this variable between conditions. We also added a measure of orthographic overlap (i.e., corrected Levenshtein distance, the distance between the target word and its translation; Schepens, Dijkstra, & Grootjen, 2012) to the analysis (See Appendix 4A for the formula). We included this predictor because Van Assche et al., (2011) showed that an increased amount of cross-lingual overlap causes a continuous facilitatory effect in word recognition.

For all analyses, RTs, word frequencies and average bigram frequencies were log transformed with base 10 to normalize their distribution. All continuous variables were centered to reduce collinearity between main effects and interactions. For the analysis, stimuli that did not reach 70% accuracy were excluded (5.31% of the data). Furthermore, responses that were more than 2.5 standard deviations above or below participants' mean RT (4.99% of the data) were excluded. Additionally, for the RTs analysis incorrect responses were excluded (4.96% of the data). Separate analyses were carried out for each language (Dutch and English) and for the nonwords, both for RTs and error rates. The fixed factors in the models were Dutch N density (continuous), English N density (continuous), Dutch N Frequency ("Yes" indicated that the word had a more frequent neighbor in Dutch, "No" if it did not), English N Frequency ("Yes" indicated that the word had a more frequent neighbor in English, "No" if it did not), word frequency (continuous), average bigram frequency (continuous) and orthographic overlap (continuous). We included a random intercept per subject in all initial models. This ensured that differences between subjects concerning genetic, developmental or social factors were modeled. We also included a random intercept per word, to be able to generalize to other nouns, because our stimuli sample is not an exhaustive list of all nouns in a language. First a full model, including the two random clusters and all of the 2-way interactions between the neighborhood variables and word frequency, word length and bigram frequency, was fitted. The optimal

[2] In an additional analysis, we analyzed the data by means of F1 (by participant) and F2 (by item) ANOVA's according to the procedure of van Heuven et al. (1998). By doing so we were able to directly compare our results to those of the original study. This analysis yielded no significant within- nor between- language effects, both in RTs and Error rates.

model was discovered by backward fitting of the fixed effects, then forward fitting of the random effects and again backward fitting of the fixed effects (Barr, Levy, Scheepers, & Tily, 2013). Finally, the condition number or κ was calculated for each model to check if collinearity was an issue. According to Belsley et al. (1980), condition indexes around 5 to 10 are associated with weak dependencies between predictors; values of 30 and higher indicate moderate to strong collinearity.

We report the analysis of the Dutch and English words below. The analysis of the nonwords is reported in the Appendix 4F.

Results Dutch words. Results of the analysis of RTs and error rates are presented in Table D1 and D2 of Appendix 4D. The condition indexes for the final models were 4.463 for RTs and 4.255 for error rates. We did not find any main effects of within- or cross-lingual neighborhood density or neighborhood frequency on reaction times or error rates.

However, for error rates the interaction between cross-lingual N density and word frequency approached significance ($\beta = 0.13$, SE = 0.07, t = 1.93, $p = .053$, see Figure D1 in Appendix 4D). Post-hoc contrasts revealed that cross-lingual N density had a facilitatory effect for low frequent words (< 1.73 log word frequency, $\chi^2 = 3.84$, df = 1, $p < .05$) and a small inhibitory effect for high frequent words (> 4.19 log word frequency, $\chi^2 = 2.71$, df = 1, $p < .1$).

Results English words. Results of the analysis of RTs and error rates are presented in Table D3 and D4 of Appendix 4D. For the RT model, $\kappa = 8.501$; for the error rates model, $\kappa = 5.725$

For reaction times, again no main effect of any neighborhood variable was found. Nevertheless, there was a significant interaction between cross-lingual N density and bigram frequency ($\beta = -0.013$, se = 0.0066, t = -2.04, $p < .05$, see Figure 2). Post-hoc contrasts revealed an inhibitory effect of Dutch N density for words with a low bigram frequency (< 2.953 log average bigram frequency, $\chi^2 = 3.85$, df = 1, $p < .05$). Reaction times for English words with a low bigram frequency were slower with increasing Dutch N density.

The interaction between the presence of a more frequent English neighbor and word frequency was significant ($\beta = -0.040$, se = 0.016, t = -2.51, $p < .05$). There was inhibition of a more frequent neighbor for low frequent words (< 3.29 log word frequency, $\chi^2 = 3.84$, df = 1, $p < .05$) and a trend towards a facilitatory effect for high frequent words (> 3.87 log word frequency, $\chi^2 = 2.71$, df = 1, $p < .1$). The contrasts of the marginally significant interactions between English N density and word frequency ($\beta = 0.0044$, se = 0.0025, t = 1.75, $p = .86$) and English N frequency and bigram frequency ($\beta = -0.056$, se = 0.032, t = -1.79, $p = .78$) did not yield significant effects.

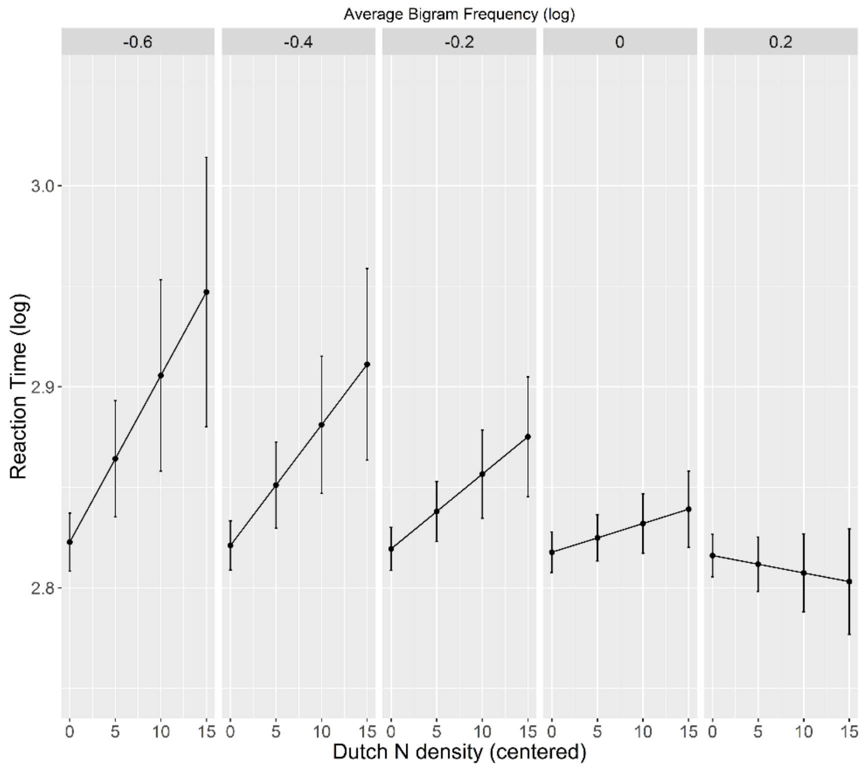

Figure 2. Reaction times (log transformed on the y-axis) for English words by Dutch N density (on the x-axis) and bigram frequency of the word (panels) for a generalized lexical decision task.

For error rates, the main effect of cross-lingual N density was significant ($\beta = 0.10$, se = 0.040, t = 2.32, $p < .05$, See Figure D2 in appendix 4D). More errors were made when the English noun had more Dutch neighbors. No other main effects of neighborhood were significant.

The marginal interaction between English N density and bigram frequency ($\beta = -0.29$, se = 0.17, t = -1.65, $p = .099$) showed significant facilitation for English N density, but only for low bigram frequency words (> 3.1055 log average bigram frequency, $\chi^2 = 3.84$, df = 1, $p < .05$). Contrasts for the marginally significant interaction between English N frequency and word frequency ($\beta = -1.01$, se = 0.55, t = -1.84, $p = .065$) showed that there was inhibition for words with a more frequent neighbor, but only for low frequent words (< 3.665 log frequency, $\chi^2 = 3.84$, df = 1, $p < .05$).

Discussion

In the present experiment, we attempted to replicate van Heuven et al.'s (1998) findings of cross-lingual effects of N density in a generalized lexical decision task for English words. A detailed pattern of neighborhood effects was discovered by analyzing the data by means of linear mixed models.

Cross-lingual neighborhood effects. ***L1 lexical decision.*** For L1 (Dutch) reading we found a near-significant effect on error rates: a lower error rate with an increased cross-lingual N density for low frequent words, but a trend for a reversed pattern for high frequent words.

L2 lexical decision. In L2 (English) reading, the cross-lingual N effects were all inhibitory: we found slower reaction times for low bigram frequency words and more errors for all L2 words when cross-lingual N density increased.

Within-language neighborhood effects. ***L1 lexical decision.*** We found no L1 (Dutch) within-language effect of N density or frequency in the current study.

L2 lexical decision. When within-language L2 (English) N density increased, fewer errors were made towards words with a low bigram frequency. We also found slower reaction times and more errors for low frequent words when the noun had a within-language more frequent neighbor.

Concerning cross-lingual N effects, van Heuven et al. (1998) found an inhibitory effect of L1 N density for reaction times on L2 words in a generalized lexical decision task. For L1 words the effect of L2 N density did not reach significance. In our lmer analyses of the replication, we found similar results for L2 and L1 words: inhibition with an increasing L1 N density for reaction times (for words with a low bigram frequency) and for error rates. For L1 words, the effect of L2 N density did not reach significance for reaction times, in the error rates there was only a trend.

Our RTs were in general slower than those of van Heuven et al. (1998). Instruction format can make a difference in lexical decision tasks when dealing with N density effects (Sears et al., 2006), but we emphasized both speed and accuracy (as van Heuven et al. did) so this is an unlikely cause of the slower reaction times. Furthermore, the language proficiency of our participants could be different from those of van Heuven et al. (1998), causing the difference in RTs. Unfortunately, van Heuven et al. (1998) did not provide proficiency scores for their participants so we cannot make a comparison. There was however a small procedural differences between our generalized lexical decision task and van Heuven et al.'s (1998). Our participants were allowed more time to answer, which indirectly might have slowed down the responses of our participants.

The dominant finding in the monolingual literature is facilitation of N density in lexical decision tasks (e.g., Andrews, 1989,1992; Carreiras et al., 1997; Pollatsek et al., 1999). We observed this in L2 but not in L1. The inhibitory within-language effect of a more frequent neighbor (e.g., Carreiras et al., 1997; Davis & Taft, 2005; Grainger & Jacobs, 1996; Perea & Pollatsek, 1998) usually found in monolingual lexical decision was also only present in L2. Taking in account all of these results, we can conclude that the results for the English L2 words are fairly comparable to the existing monolingual literature on neighborhood effects. The discrepancy for L1 words could be explained by the fact that a generalized lexical decision task was used, which creates a bilingual context that is different from a normal unilingual lexical decision task (e.g. van Heuven et al.'s (1998) English lexical decision task also yielded no L2 within-language effect for bilingual participants, whereas this effect was present in the generalized lexical decision task).

EXPERIMENT 2

In Experiment 2, we investigated N density and N frequency effects in a large publicly available database of natural reading (Cop et al., in press). We analyzed the eye movements of late unbalanced Dutch-English bilinguals when reading L1 and L2 nouns. Because of the discussion on lexical decision tasks as a marker for lexical access (Balota & Chumbley, 1984; Rayner & Pollatsek, 1989), it is very interesting to assess whether cross-lingual N effects obtained with isolated word stimuli generalize to natural text reading. Because cross-lingual neighborhood effects have only been investigated in lexical decision (van Heuven et al. (1998) and our Experiment 1), if we find cross-lingual neighborhood effects in these analyses, this would provide the first direct evidence in a completely unilingual context for the existence of activation of non-target language lexical representations.

Because of the low correlations between reaction times on lexical decision times and eye movements (Kuperman et al., 2013) and because it has been shown that neighborhood effects are very task dependent (e.g., Andrews, 1997; Carreiras et al., 1997) it is also difficult to make predictions based upon the results of the cross-lingual N effects found in lexical decision tasks. Some of the previous monolingual reading research has corroborated the idea that inhibition from neighbors might arise later in the reading process than facilitation (Perea & Pollatsek, 1998; Pollatsek et al., 1999.

We do expect that cross-lingual neighborhood effects should perfectly parallel within-language neighborhood effects, because in the BIA+ model lexical representations from both languages are included in the same integration lexical system, without distinction between both. Because top down inhibition from the language nodes is also absent (Dijkstra & van Heuven, 2002), cross-lingual neighbors should therefore behave exactly as intra-lingual neighbors. We also expect that for Dutch L1 reading, the cross-lingual effects will be smaller than for English L2 reading. This because for unbalanced bilinguals, most L2 representations are expected to be of lower frequency, thus having lower resting level activation.

Because our materials constitute an entire, long text, we analyze words varying in both neighborhood density and frequency. This means that we can examine the two effects at the same time. This will be very informative about the net effect of the neighborhood variables in bilingual natural language reading. The fact that our materials are not selected on certain lexical variables, also means that we will investigate a database of nouns from a full range of word frequency, word length and bigram frequency. Since some results have shown that such lexical variables can modulate the neighborhood effects, we do expect to identify some important conditions in which neighborhood effects are stronger.

Method

Participants and Materials. We selected all nouns (1 745 unique English and 1 777 unique Dutch nouns) from the GECO eye-tracking corpus of Cop et al., (in press). This corpus consists of eye movements recorded from nineteen unbalanced Dutch-English bilingual (seventeen female, M_{age} = 21.2 [2.2]) and thirteen English monolingual undergraduates (seven female, M_{age}=21.8 [5.6]) who read the entire novel “The mysterious affair at Styles” by Agatha Christie (Title in Dutch: “De zaak Styles”). Participants’ proficiency was tested with a proficiency battery including the LexTALE (Lemhöfer & Broersma, 2012), a lexical decision task and spelling tests (GL&SCHR for Dutch, De Pessemier & Andries (2009); WRAT4 for English, Wilkinson & Robertson (2006)). See Table B2 in Appendix 4B for detailed proficiency scores. All nouns that had an identical cognate in the other language were excluded from the dataset (8% for Dutch, 9.1% for English). The final dataset consisted of 1 576 unique Dutch and 1 447 unique English nouns. See Table 2 for characteristics of these nouns.

Procedure. Each participant read the entire novel silently in a self-paced reading task over four separate sessions. They read half of the novel in Dutch, the other half in English. The order was counterbalanced. After each chapter, multiple-choice questions were asked to check whether participants were reading for comprehension. For further details on the procedure, see Cop et al., (in press) or Cop, Keuleers, Drieghe, & Duyck (2015).

Analyses Eye movements. We analyzed two eye movement measures that reflect early language processes: Probability of first pass skipping of a word and single fixation duration, the first fixation duration on a word that is fixated exactly once. We analyzed a measure reflecting intermediate language processing: Gaze duration, the sum of all fixation durations during first passage before the eyes move out of the word.

Table 2. *Descriptive Statistics for the nouns analyzed in Experiment 2, averaged over stimuli per language (standard deviations between parentheses).*

	Neighborhood density[a]		Neighborhood Frequency[b]		Word Frequency[c]	Log Average Bigram Frequency[d]	Average Word Length	CLD[e]	Rank of Occurrence
	Dutch	English	Dutch	English					
Dutch	4.17 (5.35)	2.40 (5.16)	.30 (.44)	.18 (.35)	3.19 (0.97)	3.47 (0.23)	6.69 (2.65)	0.32 (0.26)	15.87 (30.42)
English	2.65 (4.60)	6.56 (7.44)	.25 (.42)	.53 (.50)	3.98 (0.91)	3.22 (0.24)	5.92 (2.19)	0.35 (0.29)	13.92 (20.13)

[a]Total CLEARPOND N densities (Marian et al., 2012); [b]The proportion of words with a higher frequent Neighbor; [c]Log10 Subtlex frequencies: SUBTLEX-NL for Dutch words (Keuleers et al., 2010), SUBTLEX-US for English words (Brysbaert & New, 2009); [d] Log10 summated bigram frequencies (calculated using WordGen, (Duyck et al., 2004) were normalized for corpus size and then divided by word length to obtain average bigram frequencies. [e]Corrected Levenshtein distance was calculated as a measure of orthographic overlap with the formula in Appendix 4A by manually comparing the word with its closest translation.

Finally, we analyzed two measures that reflect later, higher-order, language processes such as semantic integration: total reading time, the sum of all fixation durations on the target word, including refixations and finally regression probability, the probability of making a regression back towards the target word.

Reading time measures and skipping probabilities were fitted in (general) linear mixed models using the lme4 (version 1.1-7) and the lmerTest package (version 2.2-20) of R (version 3.1.2) (R Core Team, 2014). All of the initial models contained the fixed factors of English N Density (continuous), English N Frequency (Yes or No), Dutch N Density (continuous) and Dutch N Frequency (Yes or No). As in Experiment 1, Word Frequency (continuous), Bigram Frequency (continuous) and Orthographic Overlap (continuous) were included as predictors. Here, also Word Length (continuous) was included because this variable was not constant, as it was in Experiment 1. Furthermore, we included “rank of occurrence” as a predictor because some of the nouns occurred more than once in the novel, which could of course gradually facilitate their recognition. This factor simply consisted of the specific number of the presentation of the noun throughout the novel (i.e., “1” for the first occurrence, “2” for the second,…). All predictors were calculated the same way as in Experiment 1. Model fitting was done in the same way as in Experiment 1.

Results

Dutch L1 reading. ***Early measures.*** The outcome of the final model for skipping probabilities and single fixation durations is presented in Table E1 and E2 in Appendix 4E. For skipping probability, a logistic linear mixed model was fitted. For the single fixation analyses, only the nouns that received one fixation were selected (56.1%). Single fixation durations that differed more than 2.5 standard deviations from the subject means were excluded (2.20%). The condition index for the final skipping probability model was 10.708, for single fixation duration it was 5.169.

Cross-lingual N effects. We found no main effects of cross-lingual neighborhood density or neighborhood frequency for the early measures. The interaction between English N frequency and word frequency was marginally significant for skipping rates (β = 0.078, se = 0.043, z = -1.790, $p < .1$). The probability of skipping a word was higher when this noun had a more frequent English neighbor, but only when the noun was high frequent (>3.89 log word frequency, χ^2=3.85, df = 1, $p < .05$). For single fixation durations we found no cross-lingual neighborhood effects.

Within-language N effects. For skipping rates, we found a significant interactions of Dutch neighborhood density with word frequency (β = -0.011, se = 0.003, z = -3.266, $p <.01$) and also with word length (β = -0.007, se = 0.002, z = -2.918, $p < .01$). Post hoc contrasts showed that when nouns were low frequent (<1.90 log word frequency, χ^2=3.84, df = 1, *p*<.05) or 5 characters or less (χ^2=10.48, df = 1, *p* < .01), a larger amount of Dutch neighbors makes it more likely that the noun is skipped. For long words (14 characters or more, χ^2=3.96, df = 1, *p* < .05) a larger neighborhood density made it less likely the noun was skipped.

For single fixation durations, we found an interaction of Dutch neighborhood density with word frequency (β = 0.001, se = 0.0002, t = 3.595, *p* < .001). As the number of Dutch neighbors increased, single fixations became shorter for words with a log word frequency lower than 2.53 (χ^2=3.86, df = 1, *p* < .05) and longer for high frequent nouns (>4.23 log word frequency, χ^2=3.86, df = 1, *p* < .05).

To sum up, in L1 reading we only observed a trend for cross-lingual N effects in skipping rates, an indicator of early language processing. The presence of a more frequent cross-lingual L2 neighbor yielded skipping of high frequent L1 nouns. There was also within-language N density facilitation for low frequent and short words, and inhibition for long words early in the word recognition process.

Intermediate measures. The outcome of the final model for gaze durations is presented in Table E3 in Appendix 4E. Gaze durations that differed more than 2.5 standard deviations from the subject means were excluded (2.55%). The condition index for the final model was 6.844.

Cross-lingual N effects. None of the main or interaction effects including cross-lingual neighborhood variables reached significance.

Within-language N effects. There were no main effects of within-language N density or N frequency. Again, the interaction between Dutch N density and word frequency was significant (β = 0.001, se = 0.0003, t = 3.662, *p* < .001). Post hoc contrasts showed that for high frequent nouns (>4.39 log word frequency, χ^2=3.86, df = 1, *p* < .05), the effect was inhibitory whereas the effect was facilitatory for words with a log word frequency lower than 2.90 (χ^2=3.86, df = 1, *p* < .05). The interaction between Dutch N frequency and word frequency was also significant (β = 0.006, se = 0.003, t = 2.017, *p* < .05). Post hoc contrasts showed that fixations were shorter if a noun had a more frequent neighbor, but only when it had a log frequency lower than 3.05 (χ^2=4.02, df = 1, *p* < .05).

Late measures. The outcome of the final model for total reading times and regression rates is presented in Table E4 and E5 in Appendix 4E. Total reading times that differed more than 2.5 standard deviations from the subject means were excluded (2.90%). For regression rate a logistic linear mixed model was fitted. For the total reading time model, κ = 6.561; for the regression rate model, κ = 4.194.

Cross-lingual N effects. Participants were marginally less likely to make a regression if a Dutch noun had a more frequent English neighbor (β = -0.169, se = 0.087, z = -1.915, $p < .1$). Furthermore, for total reading times there was a marginally significant interaction between English N frequency and bigram frequency (β = -0.030, se = 0.017, t = -1.754, $p < .1$). Post hoc contrasts for this interaction did not result in any significant effects.

Within-language N effects. There was a main effect of Dutch N density for regressions (β = 0.019, se = 0.008, z = 2.384, $p < .05$): participants were more likely to make a regression to a word with an increasing number of neighbors. For total reading times, again the interaction between Dutch N density and word frequency was significant (β = 0.001, se = 0.0004, t = 3.281, $p < .01$). Dutch N density had a facilitatory effect for low frequent nouns (<2.64 log word frequency, χ^2=3.86, df = 1, $p < .05$) and an inhibitory effect for high frequent nouns (>4.34 log word frequency, χ^2=3.85, df = 1, $p < .05$). We also found a significant interaction between Dutch N frequency and word frequency (β = 0.009, se = 0.004, t = 2.394, $p < .05$) and a marginal significant one with bigram frequency (β = 0.027, se = 0.015, t = 1.859, $p < .1$). For words with a high word frequency there was an inhibitory effect of having a more frequent neighbor (>4.02 log word frequency, χ^2=3.85, df = 1, $p < .05$), but there was facilitation for words with a low word frequency (<1.08 log word frequency, χ^2=3.84, df = 1, $p < .05$). Contrasts for the interaction between Dutch N frequency and bigram frequency showed that there was inhibition of having a more frequent neighbor, but only for words with a high average log bigram frequency (>3.80, χ^2=3.84, df = 1, $p < .05$).

In sum, for L1 reading, having a more frequent L2 neighbor makes it marginally less likely that a regression will be made to the target word. Again, we found a facilitatory effect of within-language N density for low frequent words and an inhibitory effect for high frequent words. There was also an effect of within-language N frequency on total reading times (inhibitory for high frequent words, facilitatory for low frequent).

English L2 reading. ***Early measures.*** The outcome of the final model for skipping probabilities and single fixation durations is presented in Table E6 and E7 in Appendix 4E. We fitted a logistic linear mixed model for skipping probability**.** For the single fixation analyses, only the nouns that received one fixation were selected (53.7%). Single fixation durations that differed more than 2.5 standard deviations from the subject means were excluded (2.14%). For the final skipping probability and single fixation models, κ = 4.999 and κ = 8.350, respectively.

Cross-lingual N effects. For skipping probabilities, there was a significant interaction between Dutch N frequency and average bigram frequency (β = 0.256, se = 0.127, z = 2.022, $p < .05$). Post hoc contrasts for this interaction did not result in any significant effects. The main effect of cross-lingual N density was significant for single fixation durations (β = -0.002, se =

0.001, t = -2.508, $p < .05$). The interaction of Dutch N density and word length was also significant for single fixation durations (β = -0.001, se = 0.0004, t = -2.736, $p < .01$, see Figure 3). This interaction showed that there was a facilitatory effect of N density for words 5 characters long or longer (χ^2=4.72, df = 1, $p < .05$).

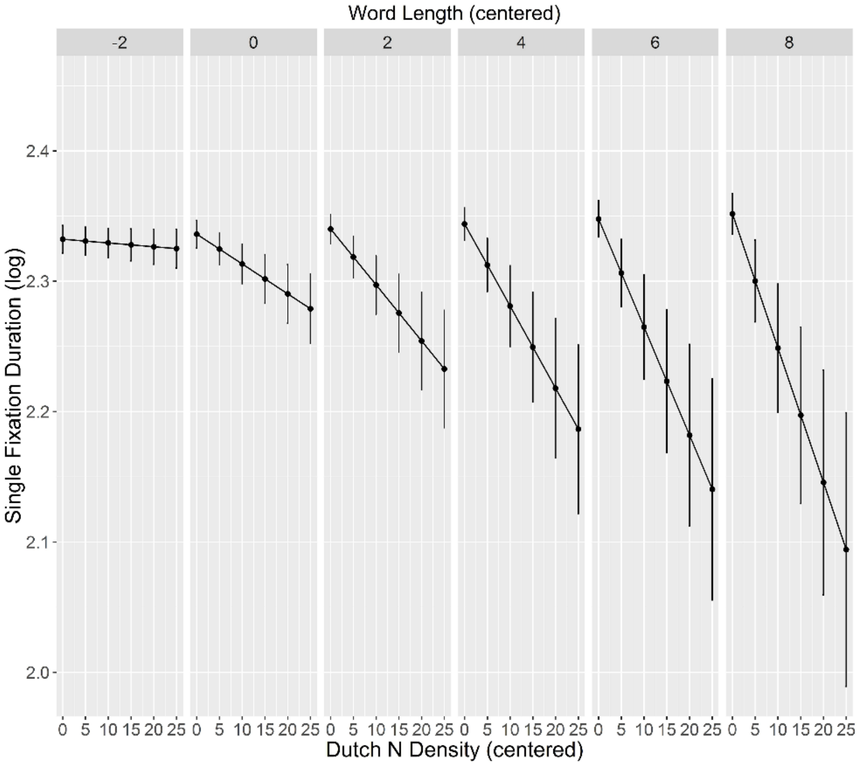

Figure 3. Single Fixation Durations (log transformed on the y-axis) for nouns dependent on Dutch cross-lingual N density (centered, on the x-axis) and target word length (panels) for English L2 reading.

Within-language N effects. The main effect of within-language N density was significant for skipping rates (β = 0.009, se = 0.003, z = 2.730, $p < .01$). Targets with more neighbors were more likely to be skipped. Furthermore, there were significant interactions between English N density and average bigram frequency (β = 0.018, se = 0.009, z = 1.986, $p < .05$), and between English N frequency and average bigram frequency (β = -0.300, se = 0.134, z = -2.239, $p < .01$). Post-hoc contrasts revealed that having a larger N density resulted in a higher skipping probability for nouns with a log average bigram frequency of 3.10 or more (χ^2=3.89, df = 1, $p < .05$). The effect of N frequency was also facilitatory, but only for nouns with a bigram frequency lower than 3.13 (χ^2=3.90, df = 1, $p < .05$). There was no effect of within-language neighborhood measures for single fixation durations.

In sum, for L2 reading, we found facilitatory effects of cross-lingual L1 N density on early language processing in single fixation duration. Within-lingual N density and N frequency also had a facilitatory effect in L2 reading, depending on the bigram frequency of the nouns.

Intermediate measures. The outcome of the final model for gaze durations is presented in Table E.8. Gaze durations that differed more than 2.5 standard deviations from the subject means were excluded (2.55%). The condition number for the final model was 8.845.

Cross-lingual N effects. We found a marginally significant main effect of cross-lingual N density on gaze durations, which was facilitatory (β = -0.002, se = 0.001, t = -1.871, $p < .1$). This measure interacted significantly with word length (β = -0.001, se = 0.0005, t = -2.174, $p < .05$, see Figure 4). For nouns with a length of 7 characters or more there was facilitation with an increasing cross-lingual N density (χ^2=4.12, df = 1, $p < .05$).

Within-language N effects. There were no significant effects of within-language N measures for gaze durations.

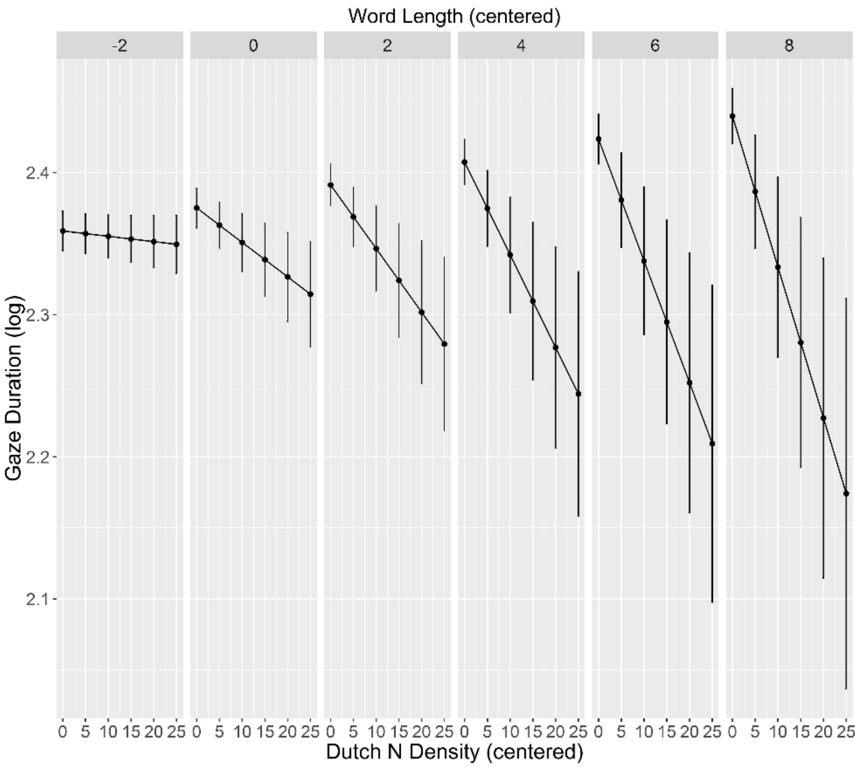

Figure 4. Gaze Durations (log transformed on the y-axis) for nouns dependent on cross-lingual Dutch N density (centered, on the x-axis), and target word length (panels) for English L2 reading.

Late measures. The outcome of the final model for total reading times and regression rates is presented in Table E9 and E10 in Appendix 4E. Total reading times that differed more than 2.5 standard deviations from the subject means were excluded (2.84%). For regression rate a logistic linear mixed model was fitted. For the total reading times model, κ = 5.898; for the regression rates model, κ = 4.954.

Cross-lingual N effects. We found a significant facilitatory main effect of cross-lingual N density on total reading times (β = -0.003, se = 0.002, t = -2.066, $p < .05$). This variable interacted significantly with word length (β = -0.001, se = 0.001, t = -1.984, $p < .05$, see Figure 5). Post hoc contrasts showed that the effect of cross-lingual neighborhood density was significantly facilitatory for words with 6 characters or more (χ^2=4.36, df = 1, $p < .05$). We failed to find any effects of cross-lingual neighborhood measures on regression rates.

Within-language N effects. There were no significant effects of any within-language N variables for regressions or total reading times.

In sum, for L2 reading, we found L1 N density facilitation for words of 6 letters and longer in late recognition processes, whereas there were no effects of L2 N density or N frequency.

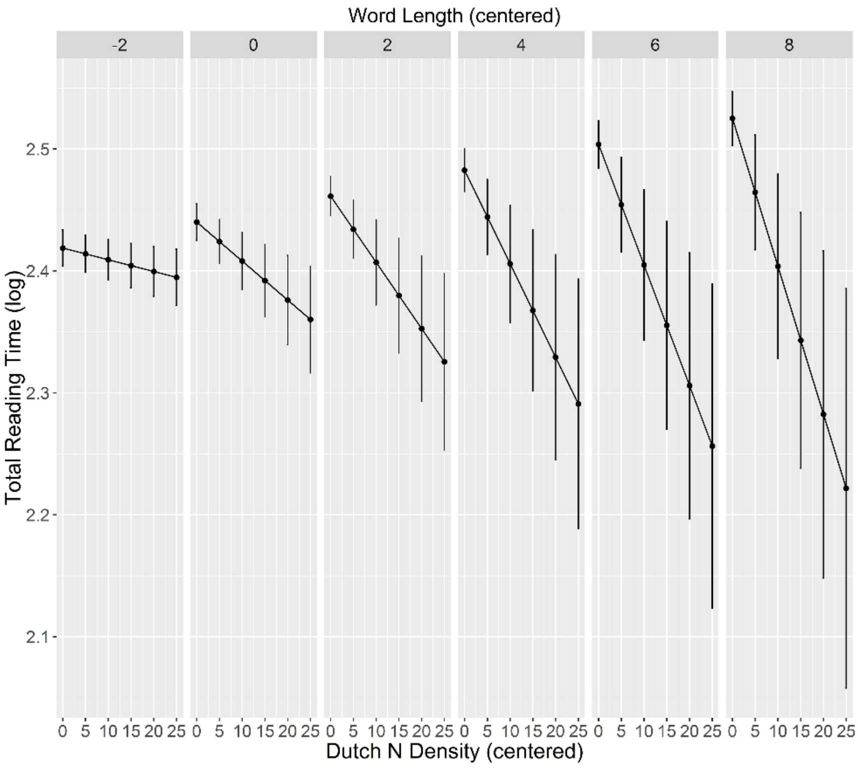

Figure 5. Total Reading Times (log transformed on the y-axis) for nouns dependent on cross-lingual Dutch N density (centered, on the x-axis) and target word length (panels) for English L2 Reading.

English Monolingual reading. To validate our neighborhood variables, we analyzed the eye movement towards nouns of monolinguals reading the same novel. These monolinguals were specifically selected as having no knowledge of any other language than English. None of the eye movement measures showed significant or marginally significant main effects of Dutch neighborhood density or frequency. Neither did any of the interactions between these measures and word frequency, word length or bigram frequency. We did find early and late facilitatory effects of English neighborhood density. For English neighborhood frequency, there was only a significant interaction with word length on skipping probability. For full analyses see Appendix 4G.

Discussion Experiment 2

Cross-lingual neighborhood effects. ***L1 reading.*** For L1 (Dutch) reading, effects of cross-lingual neighbors were rather limited. Only marginally significant effects showed up in the analysis of skipping probabilities and regression rates. In both these measures of early and late language processing, a trend towards facilitation of cross-lingual N emerged. None of the timed measures showed effects of cross-lingual N density of N frequency.

L2 reading. For L2 (English) reading, we found early facilitatory effects of cross-lingual N density: when nouns were fixated only once, these fixations were shorter. This facilitatory effect was also found for gaze durations. The fact that the effects on single fixation duration and gaze duration were stronger for long words, might be an indication that lexical access was indeed facilitated by feedback from activated neighbors to letter representations, thus speeding up the identification especially for longer words. For total reading times we also find also facilitation for nouns with increasing L1 N density. Again this effect was again stronger for longer words. This could also be due to feedback towards letter representations.

In summary the most important finding is that even when reading natural text, cross-lingual effects of neighbors were present, which is an indication of non-selective lexical activation. These effects were especially clear in L2 reading, whereas they were less convincing or absent in L1 reading. This was in line with our expectations, as the lower resting level of L2 representations could experience a larger influence of their L1 neighbors. The cross-lingual effects found in our L2 bilingual reading data were facilitatory, leading to shorter reading times for nouns with more L1 neighbors.

Importantly, the absence of an effect of cross-lingual neighbors in the monolingual data show that these effects are not confounds, but due to the knowledge of the second language of the participants.

Within-language neighborhood effects. ***L1 reading.*** For L1 reading we found within language effects of N density for early (skipping rates and single fixation durations), intermediate (gaze durations) and late (total reading times) eye movement measures. The direction of these effects was largely determined by the word frequency of the target noun. For low frequent words, a larger N density seemed to facilitate the processing of that word. For high frequent words the opposite was the case: an increasing neighborhood density slowed down the reading of the target word. Also, short words were skipped more often with increasing N density, whereas long words were skipped less. Words with a more frequent neighbor received longer total fixation times when they were high frequent, bot shorter fixation times when they were low frequent.

L2 reading. For English L2 reading, we only found an early facilitatory effect of N density. Nouns with a high bigram frequency were skipped more when they had a larger N density. For N frequency, again only an early effect was found: less skips were made of nouns with a more frequent within-language neighbor, except for nouns with a high bigram frequency.

Monolingual reading. The analysis of English monolingual reading showed facilitatory effects of N density for early measures (skipping probability and single fixation durations). For late measures, there was a facilitatory effect of N density on total reading times, as well as an inhibitory effect of N density for regression rate. For N frequency, we found more skipping with a more frequent neighbor for short words.

Our results for within-language neighborhood density are largely consistent with the results reported by Pollatsek et al. (1999). After controlling for the number of more frequent neighbors, they found early facilitatory effects of neighborhood density. Our early effects of neighborhood density were facilitatory for low frequent and short words, but inhibitory for high frequent, long words. Pollatsek et al.’s target nouns had a rather low word frequency (2.60 average log word frequency) and were rather short (average 4.5 characters). For the nouns with similar characteristics we also found facilitatory effects in our data.

To some extent we did replicate the late inhibitory effects of neighborhood frequency found in Davis et al. (2009), Perea and Pollatsek (1998), or Slattery (2009) in our bilingual reading data. In the Dutch L1 reading data total reading times were longer for words with a more frequent neighbor, but this was only true for high frequency words. For our English monolinguals, we did not find an inhibitory effect of neighbor frequency for regression rates. These monolingual English data support the hypothesis, brought forward by Andrews (1997) and Sears et al. (2006), that there would be no inhibition from neighborhood frequency for English thus separating it from other alphabetic languages, like Spanish and Dutch.

Many of the N effects are situated in the skipping rates. Facilitatory effects in skipping rates of neighborhood density or frequency have been explained by misidentification of the target word with its more frequent neighbor (Pollatsek et al., 1999; Slattery, 2009) instead of as a real reflection of faster lexical access. When we look at our Dutch L1 reading results, we observe similar effects in single fixation durations, gaze durations and total reading times. We indeed find a higher correlation between skips and regressions ($r = .55$) for nouns with a more frequent neighbor than we do for nouns without one ($r = .45$; $z = 11.16$, $p < .001$). But we did not find a positive correlation between the skipping rate for nouns with a more frequent neighbor and the total reading time for these nouns ($r = -0.043$, $t = -4.12$, $df = 9252$, $p = 1$). These results show that it might be the case that a fraction of nouns was misidentified but these misidentifications do not have a significant effect on the total time spent on nouns with a high frequent neighbor.

In general, our bilingual and our monolingual within-language reading data show, in accordance with Pollatsek et al. (1999) that there might be early facilitation from activation of letters/bigrams of lexical candidates and to some extent late inhibition in the later word selection phase, although the facilitation/inhibition mechanism seems to interact strongly with word frequency of the target.

GENERAL DISCUSSION

In this paper we investigated the effects of cross-lingual orthographic neighbors on bilingual language processing in two experiments. In Experiment 1, word recognition by Dutch-English bilinguals in a generalized lexical decision task was investigated, replicating van Heuven et al. (1998). In Experiment 2, a large database of eye movements during natural reading of a similar group (Cop et al., in press) was analyzed.

For the data of Experiment 1, using LMM's and updated measures for neighborhood density and frequency (Marian et al., 2008), we did find longer reaction times and more errors for L2 (English) words with increasing cross-lingual neighborhood density. For L1 (Dutch) words, error rates were higher for low frequent words with increasing cross-lingual N density, but there was a trend in the opposite direction (lower error rates) for high frequent. We can conclude that only with this more refined analysis did we replicate the most important result of van Heuven et al. (1998), namely the cross-lingual effect of neighbor density in a generalized lexical decision task for L2 words. We additionally found a trend towards a cross-lingual N density effect on L1 words in the error rates. This suggests that activation of cross-lingual lexical candidates may not be confined to the processing of L2 words, although this effect was not statistically reliable. Therefore, just as van Heuven et al., the present isolated word experiment offers strong evidence for an L1 influence on L2 processing, but not vice versa.

Despite these cross-lingual effects in the generalized lexical decision task, we also investigated whether these effects would be found in a more unilingual context, because such a context might provide a cue to restrict lexical search and access to the target language (e.g. Van

Assche et al., 2012), similar to the way in which readers use syntactic and semantic constraints in order to facilitate processing of upcoming words. In the current study, we therefore assessed the neighborhood effect with (a) words embedded in a completely unilingual language context and (b) a new paradigm, using eye tracking during natural language reading. In Experiment 2, a large database of bilingual eye movements (Cop et al., in press) was analyzed to find evidence for activation of cross-lingual representations. The eye movements showed effects of cross-lingual neighborhood in early and late eye movement measures for L2 reading and trends for L1 reading. The pattern of results provides strong evidence that during natural reading, both in the early phase of lexical access as in the later language processes, written words activate not only orthographically similar words belonging to the target language, but also representations belonging to the non-target language. The absence of any cross-lingual neighborhood effects for English monolinguals strongly suggests that it was indeed the knowledge of a second language that produced these cross-lingual neighborhood effects in the bilingual participants, and not some unknown lexical variable we failed to control. In summary, both the results of Experiment 1 and 2 provide evidence for parallel activation of lexical representations in bilingual word recognition and add strength to the argument of the existence of an integrated bilingual lexicon with language independent lexical access implemented in the BIA+ model (Dijkstra & van Heuven, 2002). We expected to see an asymmetry in cross-lingual effects (stronger effects in L2 than in L1 processing) because within the framework of the BIA+ model, L2 words should have a lower resting level of activation than L1 words, at least in our population of unbalanced bilinguals. This should imply that L2 words need more time to be activated, which makes them more sensitive to influences of other activated lexical candidates (Dijkstra & van Heuven, 2002). And indeed, both in the lexical decision and the eye movement results, the cross-lingual N effects were more pervasive in L2. In the lexical decision task, L1 words only showed a marginally significant cross-lingual effect in error rates, while L2 words showed effects in both error rates and reaction times. For the eye movements we see that cross-lingual N only marginally influenced skipping rates and regression rates for L1 reading, whereas for L2 reading cross-lingual N significant effects were also present in reading tomes.

Next to effects of N density, we investigated the role of N frequency (i.e., was there an effect of having a more frequent neighbor). In Experiment 1, for L2 words the effect of a more frequent within-language neighbor was inhibitory for low frequent words and facilitatory for high frequent words, but we found the reverse pattern in Experiment 2. Apparently in natural reading, a more frequent within-language neighbor speeds up low frequent word processing, while it slows down high frequent word processing. This is in contrast to what was found in the monolingual reading studies of Davis et al. (2009), Perea and Pollatsek (1998) and Slattery

(2009). This was also the first study investigating the effect of a more frequent cross-lingual neighbor on word recognition. In Experiment 1 we found no effect of cross-lingual N frequency above and beyond N density. In Experiment 2, we only found a marginally significant effect of cross-lingual neighborhood frequency in our L1 reading data. In L2 reading we only found effects of cross-lingual neighborhood density, not of N frequency. For Dutch L1 reading, the L2 neighbors seem to have to be of higher objective frequency than the target word before they are even known to our unbalanced participants. For English L2 reading the neighbors should not have to be of high frequency to have an effect, since the L1 neighbors will already be on average of higher (subjective) frequency than the L2 target words (Dijkstra & van Heuven, 2002). To address this issue of subjective frequency, an idea for future research might be to include N frequency as a continuous variable instead of the dichotomous variable we included in the current study, as this might better capture the influence of N frequency across languages on the reading process.

Lexical decision compared to eye tracking

In Experiment 1, the cross-lingual N effects were mostly inhibitory: for L2 words reaction times were slower and error rates were higher with increasing cross-lingual N density. Only for the error rates for low frequent L1 words was this effect facilitatory. In contrast, the cross-lingual N effects in the reading data were facilitatory, even in late language processing. This indicates that while performing a generalized lexical decision task, the cross-lingual activation generated by activated non-target language neighbors, slows performance, whereas in general, natural language reading benefits from this cross-lingual activation.

When interpreting the difference between the results of our experiments we have to keep in mind that the lexical decision task entails a decision component that might provoke different kinds of strategies in participants, masking the real nature of lexical access (e.g. Paap & Johansen, 1994; Rayner & Liversedge, 2011). Lexical decision results have thus been shown to be very sensitive to blocking manipulations (e.g. van Heuven et al. 1998) and the selection of nonword stimuli (e.g. Lupker & Pexman, 2010; Stone & Van Orden, 1993). The fact that we found mostly inhibitory effects of neighborhood in Experiment 1, whereas the results of most lexical decision experiments have found facilitatory effects of N density for monolingual participants illustrates this sensitivity.

Supporting the possibility that the results of the generalized lexical decision task might be influenced by processes not directly related to lexical access alone, the precise direction and interactions of effects differed substantially between Experiments 1 and 2, for both within-language and cross-lingual neighborhood effects[3]. In Experiment 1 for example, we found no within-language N effects for Dutch words, whereas there was a marginally cross-lingual N effect in error rates. For English words, we found an inhibitory effect of cross-lingual N density for error rates while this was facilitatory for within-language N density. A language system with an integrated lexicon, such as the BIA+ model, does not make a qualitative distinction between L1 and L2 lexical representations (Dijkstra & van Heuven, 2002). The partly activated neighbors from the target and non-target language should then have similar effects on target language word recognition or reading. In line with these expectations, for natural reading most of the cross-lingual effects resemble closely, although not exactly, the effects of within-language neighborhood effects. Where the patterns do diverge we see that this difference is driven by word frequency. In the current setting, natural reading might be a better approximation of lexical access than lexical decision.

Neighborhood effects in the BIA(+) model

Within the BIA+ architecture, orthographic neighbors, both of the target and the non-target language, should influence lexical access to the target word by a complex interplay between inhibitory and excitatory connections at the word and letter level (Dijkstra & van Heuven, 2002). Simulations with the BIA model have confirmed that cross-lingual neighborhood density effects could be inhibitory (Dijkstra et al, 1998). Indeed, inhibition of neighbors of the non-target language could be achieved by means of lateral inhibition. Within the BIA+ framework lateral inhibition from neighbors might be hidden by excitatory activation between representations for letters and words (as shown for the IA framework, Coltheart & Rastle, 1994; Pollatsek et al., 1999). Our data indeed shows that both inhibitory and facilitatory effects from neighbors are at play at the same time during word recognition.

The BIA+ architecture further predicts effects of the frequency of the target word and the frequency of the neighbor words. Because the subjective frequency of representations determines the resting activation of these representations, this could change the complex interactions between excitatory and inhibitory effects of activated neighbors. In our analyses of L1 and L2 language processing, we used corpus word frequencies that are supposed to reflect the frequency of exposure to words for monolinguals (Kuperman & Van Dyke, 2013). We especially expected word frequency effects to turn up in Experiment 2, because we investigated natural reading including a large range of noun characteristics (such as word frequency) in Experiment 2. In classic experiment designs where stimuli are matched on these variables per condition, it is more difficult to investigate these effects. Nevertheless, in both experiments we

[3] We do note that there is little overlap in the stimuli: only 15 of the Dutch and 17 of the English nouns of Experiment 1 were also present in the more than 1700 stimuli of Experiment 2.

found that the frequency of the target word modulates the neighborhood effects. In Experiment 1, the effect of within-language N frequency on L2 reaction times was modulated by word frequency. In Experiment 2, the effects of within-language N density on early and late language processes in L1 reading are modulated by word frequency. In both experiments the effect of increasing N density was facilitatory for low frequent words and inhibitory for high frequent words.

Considering our own findings as well as other studies finding effects of addition, deletion and transposition neighbors (e.g. Blythe, Johnson, Liversedge, & Rayner, 2014; Davis et al.,2009), we believe it important that the BIA+ model should be modified to accommodate a more flexible letter position coding mechanism. A mechanism lending itself for this purpose is the one proposed in the *overlap model* of Gomez et al. (2008). This model proposes that the representation of a letter is distributed across ordinal positions in the letter string. Every letter position has a specific standard deviation as free parameter in the model. This model expressively only models the letter coding mechanism, not any other higher order word recognition processes. This makes the overlap model easy to implement in other models, such as the BIA+ model. The effects of average bigram frequency in our data might also suggest that some kind of *open bigram coding* (Grainger & van Heuven, 2003), also implemented in the SERIOL model (Whitney, 2001), might be a good fit for these effects. Here words are coded by all of the ordered letter pairs that occur in that word. For example, the word *hand* would be determined by the bigrams [ha, hn, hd, an, ad, nd]. In our opinion, the main architectural elements of the BIA+ model have promise in accommodating our most important results, namely the cross-lingual neighborhood effects found in natural reading, as long as a more flexible letter coding mechanism is implemented.

Neighborhood effects in models of eye movements

Following the large amount of eye tracking research in reading, several models of eye movements of reading have been proposed in the last decades. As N effects never have been considered by such models, our findings could be of interest here. A first example of a model of eye movement control is the E-Z reader model (Pollatsek, Reichle, & Rayner, 2006; Rayner, Ashby, Pollatsek, & Reichle, 2004; Reichle, Rayner, & Pollatsek, 1999; Reichle, Warren, & McConnel, 2009). Although E-Z reader was designed for monolingual reading, Cop, Drieghe, and Duyck (2015) showed that L2 reading resembled child-like reading, which has been successfully simulated with the model (Reichle et al., 2013), thus making it likely this model can be applied to bilingual reading. The E-Z reader model assumes that lexical processing of words occurs serially and in two stages. In the early stage, which is called the *familiarity check*, orthographic and phonological information of the word is processed and presumably the possible lexical candidates become active. When this stage is complete, the oculo-motor system starts programming a saccade towards the next word. After completing the second stage, the *verification stage* in which full lexical identification is accomplished, attention is shifted to the next word. The duration of the two stages is assumed to be determined by the frequency of the word and its predictability. However, it seems that the neighborhood of the word could be an additional factor of contribution to their duration (this was also hypothesized by Williams et al., 2006), given the role of N density and N frequency on changes in skipping probabilities and timed measures. For example, in L2 reading we find facilitation of the cross-lingual neighborhood in early and late measures of the reading process, meaning that the familiarity and verification stages are executed faster when L2 words have a larger L1 N density. Importantly, this means that not only characteristics of the target words, but also of their neighbors determine the duration of these stages. Indeed, the facilitation could for example be due to the higher subjective word frequencies of the L1 neighbors for L2 N density effects.

Another model of eye movements is SWIFT (Engbert, Lontin & Kliegl, 2002; Engbert, Nuthmann, Richter, & Kliegl, 2005). It also proposes two lexical processing stages (preprocessing and lexical completion). The largest difference with E-Z reader is that SWIFT assumes that parallel processing of target words is possible, whereas the former proposes serial processing. Again, N density or N frequency could influence fixation times through the duration of lexical processing stages. Indeed, in simulation studies of SWIFT Engbert et al. (2005) found for example a smaller frequency effect in simulated data than in experimental data. They suggest that certain variables that were not modeled, such as N frequency, are probably needed for a larger correspondence between their simulated data and experimental observations.

Conclusion

In conclusion, our lexical decision and natural reading data both provide convincing evidence for the existence of cross-lingual activation of lexical candidates during bilingual visual word recognition. Further research should focus on the lexical variables that modulate the size or the direction of these effects, such as the word frequency, both of the target word and its neighbors.

REFERENCES

Altarriba, J., Kroll, J. F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition*, 24, 477–492.

Andrews, S. (1989). Frequency and Neighborhood effects on Lexical Access: Activation or Search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 802–814.

Andrews, S. (1992). Frequency and Neighborhood Effects on Lexical Access or Orthographic Redundancy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 234–254. doi:10.1037//0278-7393.18.2.234

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical data base on CD_ROM*. Philadelphia: Linguistic Data Consortium.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access: The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 340–357.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [stat]*. Retrieved from http://arxiv.org/abs/1406.5823

Blythe, H. I., Johnson, R. L., Liversedge, S. P., & Rayner, K. (2014). Reading transposed text: effects of transposed letter distance and consonant-vowel status on eye movements. *Attention, Perception, & Psychophysics*, *76*(8), 2424–2440. doi:10.3758/s13414-014-0707-2

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.

Bultena, S., Dijkstra, T., & van Hell, J. G. (2013). Cognate and word class ambiguity effects in noun and verb processing. *Language and Cognitive Processes*, *28*(9), 1350–1377. doi:10.1080/01690965.2012.718353

Caramazza, A., & Brones, I. (1979). Lexical Access in Bilinguals. *Bulletin of the Psychonomic Society*, *13*(4), 212–214.

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning Memory and Cognition*, *23*(4), 857–871. doi:10.1037/0278-7393.23.4.857

Clark, H. H. (1973). The Language-as-Fixed-Effect :A Fallacy Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359. doi:10.1016/s0022-5371(73)80014-3

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). London: Academic Press.

Coltheart, M., & Rastle, K. (1994). Serial Processing in Reading Aloud - Evidence for Dual-Route Models of Reading. *Journal of Experimental Psychology:Human Perception and Performance*, *20*(6), 1197–1211. doi:10.1037/0096-1523.20.6.1197

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (in press). Presenting GECO: An Eye-tracking Corpus of Monolingual and Bilingual Sentence Reading. *Behavioral Research Methods.*

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*. doi:10.1371/ journal.pone.0134008

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency Effects in Monolingual and Bilingual Natural Reading. *Psychonomic Bulletin and Review*, *20*, 1216-1234.

Davis, C. J., & Bowers, J. S. (2004). What Do Letter Migration Errors Reveal About Letter Position Coding in Visual Word Recognition? *Journal of Experimental Psychology: Human Perception and Performance, 30*(5), 923–941. doi:10.1037/0096-1523.30.5.923

Davis, C. J., & Bowers, J. S. (2006). Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. *Journal of Experimental Psychology:Human Perception and Performance*, *32*(3), 535–557. doi:10.1037/0096-1523.32.3.535

Davis, C. J., Perea, M., & Acha, J. (2009). Re(de)fining the orthographic neighborhood: The role of addition and deletion neighbors in lexical decision and reading. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1550–1570. doi:10.1037/a0014253

Davis, C. J., & Taft, M. (2005). More words in the neighborhood: Interference in lexical decision due to deletion neighbors. *Psychonomic Bulletin & Review*, *12*(5), 904–910.

De Pessemier, P., & Andries, C. (2009). *GL&SCHR*. Leuven/Apeldoorn: Garant.

Dijkstra, T. (2007). The multilingual lexicon. In M. Gaskell (Ed.), *Handbook of Psycholinguistics* (pp. 251–265). Oxford University Press.

Dijkstra, T., Grainger, J., & van Heuven, W. (1999). Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology. *Journal of Memory and Language*, *41*(4), 496–518.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, *62*(3), 284–301. doi:10.1016/j.jml.2009.12.003

Dijkstra, T., & van Heuven, W. J. B. (1998). The BIA model and bilingual word recognition. In J. Grainger & A. Jacobs (Eds.), *Localist Connectionist Approaches to Human Cognition* (pp. 189–225). New York: Mahwah.

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(03). doi:10.1017/S1366728902003012

Dijkstra, T., van Heuven, W. J. B., & Grainger, J. (1998). Simulating cross-language competition with the bilingual interactive activation model. *Psychologica Belgica*, *38*(3-4), 177–196.

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods Instruments & Computers*, *36*(3), 488–499. doi:10.3758/BF03195595

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. http://doi.org/10.1037/0278-7393.33.4.663

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42*, 621-636. doi: 10.1016/s0042-6989(01)00301-7

Engbert, R., Nuthmann, E., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review, 112,* 777-813. doi: 10.1037/0033-295X.112.4.777

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New Approaches to Language Mechanisms* (pp. 257–287). Amsterdam.

Forster, K. I., & Shen, D. (1996). No enemies in the neighborhood: Absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 696.

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review*, *115*(3), 577–600. doi:10.1037/a0012667

Grainger, J. (1990). Word-Frequency and Neighborhood Frequency-Effects in Lexical Decision and Naming. *Journal of Memory and Language*, *29*(2), 228–244. doi:10.1016/0749-596X(90)90074-A

Grainger, J., & Jacobs, A. M. (1996). Orthographic Processing in Visual Word Recognition: A Multiple Read-out Model. *Psychological Review*, *103*(3), 518–565. doi:10.1037/0033-295X.103.3.518

Grainger, J., Oregan, J., Jacobs, A., & Segui, J. (1992). Neighborhood Frequency-Effects and Letter Visibility in Visual Word Recognition. *Perception & Psychophysics*, *51*(1), 49–56. doi:10.3758/BF03205073

Grainger, J., O'regan, J. K., Jacobs, A. M., & Segui, J. (1989). On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Perception & Psychophysics*, *45*(3), 189–195.

Grainger, J., & Segui, J. (1990). Neighborhood Frequency-Effects in Visual Word Recognition - a Comparison. *Perception & Psychophysics*, *47*(2), 191–198. doi:10.3758/BF03205983

Grainger, J., & van Heuven, W. J. B. (2003). Modeling letter position coding in printed word perception. In P. Bonin (Ed.), *Mental lexicon: Some words to talk about words* (pp. 1–23). Hauppauge, NY: NOVA Science Publishers.

Grossi, G., Savill, N., Thomas, E., & Thierry, G. (2012). Electrophysiological Cross-Language Neighborhood Density Effects in Late and Early English-Welsh Bilinguals. *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00408

Guasch, M., Boada, R., Ferré, P., & Sánchez-Casas, R. (2013). NIM: A Web-based Swiss army knife to select stimuli for psycholinguistic studies. *Behavior Research Methods*, *45*(3), 765–71. doi:10.3758/s13428-012-0296-8

Holcomb, P., Grainger, J., & O'rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Cognitive Neuroscience, Journal of*, *14*(6), 938–950.

Huntsman, L. A., & Lima, S. D. (1996). Orthographic neighborhood structure and lexical access. *Journal of Psycholinguistic Research*, *25*(3), 417–429.

doi:10.1007/BF01727000

Huntsman, L. A., & Lima, S. D. (2002). Orthographic neighbors and visual word recognition. *Journal of Psycholinguistic Research, 31*(3), 289–306. doi:10.1023/A:1015544213366

Jacobs, A., & Grainger, J. (1992). Testing a semistochastic variant of the interactive activation model in different word recognition experiments. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 1174–1188.

Johnson, N., & Pugh, K. (1994). A Cohort Model of Visual Word Recognition. *Cognitive Psychology, 26*(3), 240–346. doi:10.1006/cogp.1994.1008

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. doi:10.3758/BRM.42.3.643

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. & Van Dyke, J.A. (2013). Reassessing word frequency as a Determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology, 39* (3), 802-823.

Laxon, V., Coltheart, V., & Keating, C. (1988). Children Find Friendly Words Friendly Too - Words with Many Orthographic Neighbors Are Easier to Read and Spell. *British Journal of Educational Psychology*, *58*, 103–119.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12,* 119-131. doi:10.1016/S0022-5371(73)80001-5

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*(2), 325–343. doi:10.3758/s13428-011-0146

Lupker, S.J., & Pexman, P.M. (2010). Making things difficult in lexical decision: The impact of pseudohomophones and transposed-letter nonwords on frequency and semantic priming effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35* (5), 1267-1289. doi:10.1037/a0020125

Mathey, S. (2001). Influence of orthographic neighborhood in word recognition tasks. *Canadian Journal of Experimental Psychology*, *55*, 1-23. Doi:10.1037/h0087349

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLOS ONE*, *7*(8), e43230. doi:10.1371/journal.pone.0043230

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception. An account of basic findings. *Psychological Review*, *88,* 375–407.

Midgley, K. J., Holcomb, P. J., van Heuven, W. J. B., & Grainger, J. (2008). An electrophysiological investigation of cross-language effects of orthographic neighborhood. *Brain Research*, *1246*, 123–135. doi:10.1016/j.brainres.2008.09.078

Paap, K., & Johansen, L. (1994). The Case of the Vanishing Frequency Effect - a Retest of the Verification Model. *Journal of Experimental Psychology-Human Perception and Performance*, *20*(6), 1129–1157. doi:10.1037/0096-1523.20.6.1129

Paap, K., Newsome, S., McDonald, J., & Schvaneveldt, R. (1982). An Activation Verification Model for Letter and Word Recognition - The Word-Superiority Effect. *Psychological Review*, *89*(5), 573–594. doi:10.1037/0033-295X.89.5.573

Peereman, R., & Content, A. (1995). Neighborhood Size Effect in Naming – Lexical Activation or Sublexical Correspondences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *21*(2), 409–421. doi:10.1037/0278-7393.21.2.409

Peeters, D., Dijkstra, T., & Grainger, J. (2013). The representation and processing of identical cognates by late bilinguals: RT and ERP effects. *Journal of Memory and Language*, *68*(4), 315–332. doi:10.1016/j.jml.2012.12.003

Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 767.

Perea, M., & Rosa, E. (2000). Repetition and form priming interact with neighborhood density at a brief stimulus onset asynchrony. *Psychonomic Bulletin & Review, 7(4),* 668-677. doi:10.3758/bf03213005

Pollatsek, A., Perea, M., & Binder, K. S. (1999). The Effects of "Neighborhood Size" in Reading and Lexical Decision. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 1142–1158.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology, 52(1),* 1–56. doi:10.1016/j.cogpsych.2005.06.001

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The Effects of Frequency and Predictability on Eye Fixations in Reading: Implications for the E-Z Reader Model. *Journal of Experimental Psychology: Human Perception and Performance, 30(4),* 720–732. doi:10.1037/0096-1523.30.4.720

Rayner, K., & Liversedge, S. P. (2011). Linguistic and cognitive influences on eye movements during reading. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford Handbook of Eye movements* (pp. 751–766). UK: Oxford University Press.

Rayner, K., & Pollatsek, A. (1989). The psychology of reading. New York, NY: Prentice-Hall.

Reichle, E. D., Liversedge, S. P., Drieghe, D., Blythe, H. I., Joseph, H. S. S. L., White, S. J., & Rayner, K. (2013). Using E-Z Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review, 33(2),* 110–149. doi:10.1016/j.dr.2013.03.001

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105(1),* 125–157. doi:10.1037/0033-295x.105.1.125

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16(1),* 1–21. doi:10.3758/pbr.16.1.1

Scarborough, D., Gerard, L., & Cortese, C. (1984). Independence of Lexical Access in Bilingual Word Recognition. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 84–99. doi:10.1016/S0022-5371(84)90519-X

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, *15*(01), 157–166. doi:10.1017/S1366728910000623

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Pschology Software Tools, Inc.

Sears, C. R., Campbell, C. R., & Lupker, S. J. (2006). Is there a neighborhood frequency effect in English? Evidence from reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 1040–1062. doi:10.1037/0096-1523.32.4.1040

Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood size and neighborhood frequency effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 876.

Slattery, T. J. (2009). Word misperception, the neighbor frequency effect, and the role of sentence context: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1969–1975. doi:10.1037/a0016894

Snodgrass, J. G., & Mintzer, M. (1993). Neighborhood effects in visual word recognition: Facilitatory or inhibitory? *Memory & Cognition*, *21*, 247–266.

Stone, G.O, & Van Orden, G.C. (1993). Strategic control of processing in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 744-774.

Van Assche, E., Drieghe, D., Duyck, W., Welvaert, M., & Hartsuiker, R. J. (2011). The influence of semantic constraints on bilingual word recognition during sentence reading. *Journal of Memory and Language*, *64*(1), 88–107. doi:10.1016/j.jml.2010.08.006

Van Assche, E., Duyck, W., & Brysbaert, M. (2013). Verb Processing by Bilinguals in Sentence Contexts. *Studies in Second Language Acquisition*, *35*(02), 237–259. http://doi.org/10.1017/S0272263112000873

Van Assche, E., Duyck W., & Hartsuiker, R.J. (2012). Bilingual word recognition in a sentence context. *Frontiers in Psychology,3* ,174, doi:10.3389/fpsyg.2012.00174

Van Assche, E., Duyck, W., Hartsuiker, R. J., & Diependaele, K. (2009). Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological Science*, *20*(8), 923–927.

van Heuven, W. J. B., Dijkstra, T., & Grainger, J. (1998). Orthographic Neighborhood Effects in Bilingual Word Recognition. *Journal of Memory and Language*, *39*, 458–483.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2013). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. doi:10.1080/17470218.2013.850521

Wilkinson, G. S., & Robertson, G. . (2006). *Wide Range Achievement Test 4 professional manual*. Lutz, FL: Psychological Assessment Resources.

Williams, C. C., Perea, M., Pollatsek, A., & Rayner, K. (2006). Previewing the neighborhood: The role of orthographic neighbors as parafoveal previews in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 1072–1082. doi:10.1037/0096-1523.32.4.1072

CHAPTER 5

READING TEXT WHEN STUDYING IN A SECOND LANGUAGE: AN EYE-TRACKING STUDY.[1]

In this study, we investigated how the reading pattern is influenced by different text reading goals (informative reading vs. studying of a text) in the first (L1) and second (L2) language. Participants had to read or study multiple texts in different languages while their eye movements were recorded. Additionally, they had to complete true/false questionnaires about the content of the texts, as we also wanted to investigate whether recognition memory for texts is affected by the reading goals and the language of the material. In general, more time was spent on studying than on informative reading, which also resulted in higher test scores in the study condition. The results also showed that studying in a L2 takes about 20% longer, about 15% more fixations were made and saccades were smaller in comparison to L1 studying. Interestingly, the limited additional time for L2 studying seemed to compensate the impaired processing, as recognition test scores were similar across languages.

[1] Dirix, N., Vander Beken, H., de Bruyne, E., Brysbaert, M., & Duyck W. (2018). Reading text when studying in a second language: an eye-tracking study. *Manuscript submitted for publication.*

INTRODUCTION

In countries with a native language other than English, there is an increase in the use of English as a Medium of Instruction (EMI). This development is, amongst others, driven by a desire to increase international education and mobility or to adapt to the international domination of English as a lingua franca in science and education. One of the consequences in higher education is an increase in the use of English textbooks. As a consequence, students who have Dutch as their native language (L1) are expected to understand and remember the content of these textbooks in a second language (L2) in which they are less proficient.

Although it is not unreasonable to assume disadvantages for students who have to study in L2, the exact total cost is largely unknown. Besides studying and recall, at least for encoding (reading) of written study materials, there is some evidence that shows detrimental consequences of L2 use: L2 words are read slower than their L1 counterparts (Duyck, Vanderelst, Desmet, & Hartsuiker, 2008; Whitford & Titone, 2012). In a comparison of L1 and L2 reading of an entire book, Cop, Drieghe, and Duyck (2015) found similar results for sentence reading times: highly proficient, but unbalanced and late Dutch-English bilingual university students read 18% slower when reading in L2 relative to L1. For later stages of studying, in a paper on incidental learning Gablasova (2014) has shown that students' vocabulary retention is worse for L2 than L1 words. The question arises whether these L2 costs have an influence on the actual study of academic texts and how this affects the memory for the content of the text, relative to L1.

Studying in L1 and L2

Only a few studies exist on this issue: Chen and Donin (1997) investigated the studying of texts by Chinese (L1) – English (L2) bilingual subjects. In the experimental conditions, participants had to read short L1 and L2 texts and orally recall the content in the same language. They found that participants spent more time to read the texts in L2 in comparison to L1, but surprisingly no L1 – L2 difference was found on the recall test. In a similar study with English (L1) – French (L2) participants, longer reading times for L2 than L1 were again found, but this time an L2 recall cost appeared (Donin, Graves, & Goyette, 2004). Given the slower processing rate in L2 reading (e.g., Cop, Drieghe, et al., 2015), it is not surprising that L2 studying takes longer. It is however interesting to note that this impaired processing does not necessarily result in a diminished memory performance for the content of the texts. Additional research on the role of the encoding stage (and study time) in L1/L2 memory for texts seems necessary.

In a recent study, Vander Beken and Brysbaert (2017) further examined memory for short texts (approximately 250-300 words) in L1 and L2 by comparing different types of tests. Their Dutch (L1) – English (L2) bilingual university student participants were split in two groups: one group had to study two expository texts in L1, the other group in L2. All participants received a free recall test for one of the texts and a true/false judgement test for the other one. Their test scores revealed better performance for L1 than L2 on the recall test, but there was no difference between the groups on the true/false judgement test. Equal performance on recognition memory suggests that encoding of information is similar between L1 and L2, and that this information may be retrieved to a similar extent if cues are provided. In a follow-up study with the same materials, retaining only the recognition test, Vander Beken, Woumans, and Brysbaert (2017) found that recognition performance was equal for L1 and L2 even after a delayed period of up to 30 days. This suggests that the memory trace of information encoded in L2 also does not decay at a faster rate than in L1. However, the L1 – L2 difference on the recall test in the first study suggests there is however an L2 cost when access to the encoded information is not supported by cues. Vander Beken and Brysbaert provide two possible explanations for this recall cost. One possibility is that it might be more difficult for participants to reproduce their thoughts in writing in L2 than in L1 (e.g., see Joh, 2006). In language production tasks, it is indeed a common finding that bilinguals make fewer and slower correct responses and show delayed retrieval (e.g., Gollan, Fennema-Notestine, Montoya, & Jernigan, 2007; Sandoval, Gollan, Ferreira, & Salmon, 2010). Another explanation for the L2 recall cost involves the Landscape Model of van den Broek, Young, Tzeng, and Linderholm (1999). The assumption of this model is that while reading a text, concepts are activated and interconnected, resulting in a specific “landscape” of concept activation for the text. Vander Beken and Brysbaert (2017) propose that a difference in richness between mental models created in L1 and L2, caused by lesser (co-)activation of concepts in L2, may lead to the L2 recall cost. Whether one of these issues, or perhaps a combination of the two, is at the root of the L2 recall cost remains to be investigated.

To summarize, two main findings show up in studies directly comparing the study of L1 and L2 texts. First, performance on a recognition test is equal in L1 and L2, but an L2 cost appears in more demanding recall tests (Donin et al., 2004; Vander Beken & Brysbaert, 2017). Furthermore, the studies of Chen and Donin (1997) and Donin et al. (2004) suggest that participants need more time to study an L2 than an L1 text. Vander Beken and Brysbaert do not report longer L2 study times as their participants had a fixed interval to study the texts. It is however not clear whether studying in L2 actually took as long as studying in L1, because this was not explicitly measured.

On a more general note, although these studies have provided valuable insights in the memory for texts in L1 and L2, there still is little understanding about the studying process itself. For instance, does studying in L2 mainly take longer because of the slower processing rate of individual words, longer fixations on certain text information, or is the increase in study time due to increased repetition of the different parts of the text (for example because of a more difficult integration of sentences)?

A technique that could provide a detailed overview of the studying process in L1 and L2 is eye-tracking. The study of eye movements has indeed demonstrated its value in written language research by successful applications in visual word recognition and text processing.

Eye movements in L1 and L2

Eye-tracking is a non-invasive technique with a high spatial and temporal resolution: the position of the eyes is monitored at a rate of up to 2 000 times per second. If the eyes remain still over a period of time, this is called a fixation. A movement of the eyes between two fixation points is a saccade. An important advantage of this technique for reading research is that it allows to study natural reading without specific task demands or response strategies (e.g., in comparison to a lexical decision task, see Kuperman, Drieghe, Keuleers, & Brysbaert, 2013). It suffices to instruct participants to simply read what appears on the screen. Yet, a lot of valuable information can be derived from the particular eye movements (Rayner, 1998, 2009; Rayner, Chace, Slattery, & Ashby, 2006). For example, the duration of the first fixation on a word indicates how easily this word is retrieved from the mental lexicon and recognized. Some words are also fixated multiple times, so that the total reading time becomes longer. That may be because the word is long, or because refixations of the word are needed for verification of recognition or its integration in the semantic context. Eye movement research is well-established within the domain of psycholinguistic research, and has also found its way to bilingualism in the last two decades.

Reading. Eye-tracking has for example been applied in the discussion whether bilinguals have an integrated lexicon for both languages and, as a consequence, activate both languages when processing written language (Dijkstra & van Heuven, 2002; Van Assche, Duyck, & Hartsuiker, 2012). In single sentence reading, it has been demonstrated that fixation durations on individual words are influenced by knowledge of a task-irrelevant language, both in L1 and L2 (Duyck, Van Assche, Drieghe, & Hartsuiker, 2007; Van Assche, Duyck, Hartsuiker, & Diependaele, 2009).

In addition to single sentence reading, eye-tracking has also been applied to investigate word-level effects in larger chunks of text. For example, word frequency has a bigger effect on reading in L2, especially because low-frequent words are recognized much slower (Duyck et al., 2008), an effect that was also observed in paragraph reading (Whitford & Titone, 2012). This is relevant for the present study because academic textbooks often containing difficult, low-frequent words.

Looking at word-level effects in lager chunks of text, a study well worth mentioning is the Cop, Drieghe, et al.'s (2015) investigation of the Ghent Eye-tracking COrpus (GECO; Cop, Dirix, Drieghe, & Duyck, 2017)). This is a collection of eye movement data of bilinguals reading half a novel for comprehension/pleasure in L1 and the other half in L2. Participants showed inferior text processing in L2 in comparison to L1, making more and longer fixations, smaller saccades, and skipping fewer words in L2. They address these differences to slower activation and verification processes of word recognition in L2, and suggest that a general reduction in the rate of lexical processing could underlie this impaired L2 processing of text.

These eye-tracking studies of bilingual reading indeed show that a lot of information on underlying processes is contained within eye movement patterns. A next logical question is whether eye movement patterns can also provide information on reading differences while studying texts.

Studying. In a recent study, Yeari, van den Broek, and Oudega (2015) investigated the eye movements of participants who read short texts in Dutch for different reading goals (i.e., informative reading, reading to prepare a presentation, reading to prepare for a closed-question test and to prepare for an open-ended question test). They examined various eye movement measures of the processing of *information units* in the text (consisting of a "main predicate, it's arguments […] and the adjectives and/or adverbs of these arguments", p. 1076). Importantly, these units were also rated on their degree of *information centrality*: units or ideas that are *central* are of critical importance to the overall understanding of the text, whereas *peripheral* ideas are less crucial to achieve a good text comprehension (e.g., Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975; Miller & Keenan, 2009; Yeari, Oudega, & van den Broek, 2017). In previous research it was demonstrated that information centrality has an influence on the eye movement pattern while reading a text: it takes participants longer to read a sentence for the first time (i.e., the first-pass time) if it contains central information than when it contains peripheral information, and more regressions are made towards central sentences (Hyönä & Niemi, 1990). Yeari et al. (2015) also presented all participants with a multiple-choice question test, irrespective of their reading goal, to assess memory for the text in general and for central and peripheral information in particular. Their eye-tracking results showed that (a) participants had longer total reading times and made more fixations for studying purposes compared to informative reading; (b) the first pass time was longer for central than peripheral sentences for all reading goals, but for the total reading time, this difference disappeared for the two test purposes. The authors conclude that the eye movements reveal a different reading strategy based on the specific purpose: a text is more thoroughly read when a test is expected, and more attention is directed towards peripheral ideas when re-reading the text for a studying purpose.

The resulting test accuracy scores were higher for the two test purposes (closed- and open-ended questions) than for informative reading, and participants' accuracy on central information questions was higher than that of peripheral information for all reading goals. Finally, the authors tested the notion whether the time spent reading an information unit can be related to the test score for an item about that unit, or in other words, whether reading time can predict test accuracy. They reported relatively low correlations between the total reading times and response accuracy on the multiple choice tests, which led them to the conclusion that the strength of a memory representation is only influenced to a lesser degree by the time spent encoding it.

This study again demonstrates the value of eye-tracking research, this time more closely related to educational purposes. For the current study, we will extend the investigation of different reading goals to the bilingual domain, as currently no study exists on eye movements of studying in a L2.

Current study

We presented participants with four texts that are representative for the factual and academic texts students have to process in higher education, half of them in L1, the other half in L2. We manipulated the reading goal of these texts between-subjects: half of the participants studied the texts in order to prepare for true/false judgement tests, the other half were told to read for entertainment (informative reading). All participants received the true/false judgement test afterwards, which relates to typical recognition tests in a higher education setting. We monitored the eye movements of all participants during their processing of the texts (more specifically, the information units in the text, see below) in order to assess effects of language (L1 or L2) and reading goal (reading or studying) on the various common eye movement measures. For the current study, we investigated the first pass time (the time it took to read an information unit for the first time), the total reading time (the summed duration of all passages of an information unit), the fixation count (the total number of fixations made towards an information unit), the regression count (the number of regressions made towards an information unit, hence the number of times readers go back to a section of the text) and the saccadic amplitude (the length of the forward saccade departing from the unit, hence the size of the 'jumps' that readers make, progressing through the texts).

A first goal of this study was to compare the eye movement patterns of participants studying texts between L1 and L2. Vander Beken and Brysbaert (2017) propose that the encoding of L2 information in memory is not impaired (as recognition memory seems to be similar to that in L1), but studies on reading, rather than studying, such as Cop, Drieghe, et al.'s (2015), suggest that L2 text processing is somewhat impaired. Similar to Yeari et al. (2015), we also coded the information units in our texts according to information centrality to assess effects of central and peripheral information in the units. We expected an information centrality effect on the first pass time (a faster first passage for peripheral than central information), but not on total reading time. In accordance to Cop, Drieghe, et al. (2015), we expected that a general pattern of slower L2 processing would appear: longer and more fixations (both in first pass and total reading time) and smaller saccades for L2 studying, relative to L1. This would imply that EMI for non-native speakers of English would carry an encoding cost that is larger than that observed in plain reading studies (Cop, Drieghe, et al., 2015).

A second goal was to examine the effects of a different reading goal (informative reading or studying) within, and across languages. Our L1 results could serve as a replication of part of the findings of Yeari et al. (2015). Furthermore, the L2 pattern could reveal whether participants show the same adaptations of their eye movement pattern to the reading goal in a L2. In L1, we expected a similar pattern to that of Yeari et al. (2015): (a) longer reading times and more fixations for the study condition compared to reading, (b) a longer first pass time for central than peripheral information in both reading goals and (c) an interaction between reading goal and information centrality for the total reading time: the central-peripheral difference remains for reading, but disappears for studying. We expected a similar pattern for L2 reading vs. studying, although the longer L2 reading times could accumulate in the studying condition, resulting in even larger differences with the reading condition in comparison to L1 studying.

The third goal of the study was to investigate access to the memory trace of text content, by means of the accuracy scores on the true/false judgement tasks in all conditions. These judgements were related to specific central and peripheral units in the texts (see below for a detailed description of the test construction). We expected that the accuracy scores on this recognition test would not be different between the L1 and L2 study condition, following the results of Vander Beken and Brysbaert (2017). In combination with our expectations for the reading measures, this would mean that although L2 studying comes with an encoding cost, recognition processes are not harmed by the inferior primary encoding stage. Furthermore, we expected higher test accuracy in the studying condition than in reading, and better scores for central than for peripheral information (cf. Yeari et al., 2015).

A fourth additional and final goal was to examine, as Yeari et al. (2015) did, whether reading times could predict accuracy scores. In general, Yeari et al. found low correlations between these measures, but in a separate analysis of their closed-question condition (which is similar to our testing condition), the correlation between total reading time and accuracy score was significant. We expected to obtain a similar result.

We decided to include several covariates in both the analysis of the eye movement measures and the accuracy scores. In eye movement research, there are well-established findings of variables influencing reading times, such as word frequency (Cop, Keuleers, Drieghe, & Duyck, 2015; Duyck et al., 2008; Whitford & Titone, 2012), word length (Drieghe, Brysbaert, Desmet, & De Baecke, 2004; Rayner, Slattery, Drieghe, & Liversedge, 2011) the number of words in the unit (or unit length; Cop, Drieghe, et al., 2015), and the language proficiency of the participants (both L1 and L2; Cop, Keuleers, et al., 2015; Whitford & Titone, 2012). As we do not specifically control these variables in our experimental design or test material, we found it important to account for them in the analyses. Following the same reasoning, for the accuracy score analysis, we included factors that could influence text comprehension: text perception (Vander Beken & Brysbaert, 2017), language proficiency (Droop & Verhoeven, 2003), reading motivation (Andreassen & Bråten, 2009) and prior knowledge about the topic (Coiro, 2011).

METHOD

Participants

Eighty participants took part in this experiment (M_{age} = 19.39, SD_{age} = 4.66; 67 females). They were first year bachelor psychology students at the Faculty of Psychology and Educational Sciences of Ghent University. They were all Dutch native speakers, who received formal English education from age 14 on, and who were exposed to English regularly through (online) media. Hence, they were proficient, but unbalanced and late bilinguals (see Table 2 for language proficiency ratings). The participants were asked to sign an informed consent and received course credit as well as an additional payment of €5 for their participation. All participants had corrected or corrected-to-normal vision.

Materials

Texts. Four texts were used in the current experiment: two expository texts (“Sea otters”, “The Sun”) and two academic texts (“Metacognition”, “Problem solving”). We did not include the distinction of text type in our analyses, as this was not a topic of interest for the current study. The length of the texts varied between 248 – 432 words in Dutch, and between 285 – 421 words in English. All texts were presented in Arial, size 18 black letters on a white background with 1.5 line spacing. All texts fitted on one screen.

The expository texts about the Sun and sea otters were taken from a study into the test effect by Roediger and Karpicke (2006). Vander Beken & Brysbaert (2017) translated the English texts to Dutch, and matched the language versions on semantics, frequency and prevalence (in that order; see Vander Beken & Brysbaert, 2017 for a detailed description).

The academic texts were taken from research articles that are published in international peer-reviewed journals (metacognition: Efklides, 2006; problem-solving: Mayer, 1998), as these constitute typical learning materials for university students. The metacognition text discusses the role of metacognition in learning, and explains metacognitive skills and metacognitive experiences. The problem solving text includes an introduction about routine and nonroutine problem solving, and the components of successful problem solving. The original English texts were translated by a content expert to Dutch and checked by a native speaker. Both language versions were compatible in style, content, structure and length.

The four texts were divided into information units. These units were the specific interest areas for which the eye movements were monitored (see Table 1 for unit characteristics). Roediger and Karpicke (2006) divided the texts about the Sun and sea otters into 30 propositions or ideas. In a few cases, we further divided propositions that contained several units of factual information (such as a name and an event). Those units were then rated on a 5-point Likert scale. The instruction for this task was the following: “Indicate how important you think the content of every expression is in this text by circling the corresponding number from 1 (totally unimportant) up to 5 (very important)”. Since we aimed at a realistic measure of importance, predicting to some extent what the subjects of this study would consider important and what they ought to remember from this text in an academic context, the units were rated by 10 experts and 10 students. Half of the experts were academics (PhDs or PhD-students) in (marine) biology, the other half in astronomy. The students were psychology students at XXX University (similar to the Vander Beken and Brysbaert study and this study) and had not taken part in any study with these materials. Similar to the expository texts, the academic texts were divided in several units of textual information or ‘information units’. Twelve experts and six students rated the units. The experts were academics (PhDs or PhD-students) in educational sciences or psychology. The students were masters in educational sciences at XXX University and had not taken part in any study with these materials. The average of all ratings was taken per unit. Units with a rating higher than 4 were considered as central information, ratings lower than 3 were considered peripheral (note that the overall ratings were consequently higher than 2, possibly due to the density of information in the texts).

Table 1. *Descriptive Statistics of the texts of the current study, averaged over information units per language and information centrality type (standard deviations between parentheses).*

	Dutch (L1)			**English (L2)**		
	unit length[a]	average word length[b]	average word frequency[c]	unit length[a]	average word length[b]	average word frequency[c]
Central	11.55 (3.54)	6.09 (1.78)	5.30 (0.59)	11.86 (4.48)	5.23 (1.63)	5.51 (0.54)

Peripheral	9.9 (4.56)	5.48 (1.00)	5.38 (0.57)	9.68 (4.22)	5.09 (0.79)	5.37 (0.76)

[a]The number of words in the information unit; [b]The average word length of the words in the unit; [c]The average Zipf SUBTLEX frequency of the words in the unit: SUBTLEX-NL for Dutch words (Keuleers, Brysbaert, & New, 2010), SUBTLEX-UK for English words (van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

Tests. Ten true/false statements were created for each text, five of which related to the five most central units and five of which related to the five most peripheral units. Vander Beken and Brysbaert (2017) created true/false questions that corresponded to all 30 (or more) of Roediger and Karpicke's propositions in "The Sun" and "Sea otters", from which these ten were selected. For "Metacognition" and "Problem solving", De Bruyne, Aesaert, & Valcke (2017) developed knowledge mastery tests to measure domain-specific content knowledge after reading the research articles of Efklides (2006) and Mayer (1998). These include free recall questions and multiple-choice items. Based on the relatively short selection we made from those texts, true/false statements were created out of almost literal sentences from the knowledge tests.

Participants had to complete the digital tests on the open source survey software tool Limesurvey (version 2.05; http://www.limesurvey.org). Students made true/false judgements after the following instruction: "Are the following statements true according to the text? Answer with yes (true) or no (false)". The test administration is further described below under "procedure".

Additional questionnaires. The participants had to complete an additional questionnaire with questions regarding their perception of the text and their reading motivation in Dutch and English. For text perception, participants were asked a few questions about each of the texts they read: how interesting they found them (text interest), how difficult they found the content of the text (content difficulty), and to which degree they were already familiar with the content (prior knowledge). For reading motivation, participants were asked for each language how much they like to read (reading motivation), their personal judgment of their reading capability in that language (reading self-efficacy) and how important they thought it was to be able to understand texts in that language (perceived reading importance). All these questions were ratings on a 7-point Likert scale. Finally, participants completed the Dutch and English version of the LexTALE (a language proficiency test; Lemhöfer & Broersma, 2012), as well as self-ratings of their L1 and L2 proficiency on a 5-point Likert scale (see Table 2). The LexTALEs were programmed in C with Tscope (Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006), the additional questionnaires were presented on Limesurvey.

Table 2. *Descriptive statistics of the participants' language proficiency (standard deviations between parentheses)*

	Dutch (L1)	English (L2)
LexTALE score	88.92 (5.41)	72.89 (9.75)
Self-ratings		
Reading	4.91 (0.33)	3.48 (0.78)
Listening	4.96 (0.19)	3.79 (0.67)
Writing	4.79 (0.47)	4.79 (0.47)
Speaking	4.98 (0.16)	3.48 (0.78)

Apparatus. The eye movements were monitored with an Eyelink 1000+ (desktop mount version; SR Research). Only the dominant eye was recorded, at a sampling rate of 1000Hz. The texts were presented on a 24-inch screen (1920 x 1080). The participants were seated in a comfortable chair at approximately 95cm of the screen. During the reading or studying of the texts, their head was fixed by a chin- and headrest mounted to the table. The questionnaires were administered on a Dell Latitude E5550 laptop with a 15.6-inch monitor.

Procedure

The participants were welcomed by the experimenter and given oral instructions about the experiment. Participants in the studying condition were informed that, and how they would be tested about text content afterwards. They were asked to study the text in order to be optimally prepared to complete the tests, as they would do for a regular university exam. Participants in the informative reading condition were unaware about the subsequent tests, and they were asked to read the texts like they would encounter them in a magazine or on a website.

After these initial instructions, participants were installed in front of the eye tracker. Before the presentation of the first text, a 9-point calibration was performed. Participants were then given a maximum of 10 minutes to read or study the first text. The text was removed from the screen after the 10 minutes expired, or participants could press the spacebar when they finished earlier. Between texts, they were allowed to take a small break. Before the presentation of each following text, a new 9-point calibration procedure was executed. There was always a switch of language between texts: if the first text was an L2 text, the following would be an L1 text. Text and language order were counterbalanced across participants. The fixed interval of 10 minutes remained the same for all texts.

After studying or reading the four texts, participants had to complete the four comprehension tests with true/false questionnaires. Participants in the informative reading condition were only informed about the tests at this point of the experiment. The order and language of the tests was the same as the presentation of the texts in the studying / reading phase. After completing the tests, the participants were first presented with the additional questionnaires, then the Dutch and English LexTALEs (the order of these two was determined by the language of the first text they read: if this was an L1 (Dutch) text, the Dutch LexTALE was completed first) and finally the proficiency self-ratings. The entire session lasted about 60 minutes for reading and 90 minutes for studying.

Data Analysis

All data-analysis was performed in R (version 3.4.1, R Core Team, 2017). For the (generalized) linear mixed effects models ((G)LMMs), lme4 (1.1-13) and lmertest (for computation of p-values; 2.0-33) packages were used.

Eye movements. All eye movements were analyzed with LMMs. Only data of the 10 units in each text with the highest / lowest information centrality were included in the analysis. The dependent variables were first pass time (in seconds), total reading time (in seconds), fixation count, regression count and saccadic amplitude (in degrees of visual angle). The fixed effects structure consisted of our factorial design: language (Dutch/L1 or English/L2) x reading goal (studying or reading) x information centrality (central or peripheral). In addition, we included the covariates unit length (the number of words in the unit), average word frequency (the average Zipf word frequencies of the unit; Dutch frequencies from SUBTLEX-NL (Keuleers et al., 2010), English frequencies from SUBTLEX-UK (van Heuven et al., 2014)), average word length (the average number of letters of the words in the unit), and L1 and L2 proficiency of the participants (Dutch and English LexTALE scores). As language and reading goal were our main variables of interest, we also included three-way interactions between these variables and all of the covariates. All continuous predictors were centered to reduce correlations between main effects and interactions. Random effects per participant and information unit were included to ensure that (a) genetic, developmental or social differences between participants were represented in the model and (b) we were able to generalize to other information units and texts (Baayen, Davidson, & Bates, 2008). As each participant read texts in multiple languages, and each unit could be presented in two reading goals, language was fitted as a random slope for participants and reading goal as a random slope for unit, respectively.

Accuracy scores. Accuracy scores were analyzed with a GLMM. The dependent variable was the score for each true/false statement (1 for correct or 0 for incorrect). The factorial design was again included in the fixed effects (language x reading goal x information centrality), as well as covariates for reading motivation (self-efficacy, perceived reading importance, and reading motivation), text perception (prior knowledge, text interest, and content difficulty) and L1-L2 proficiency. A random factor for participant and question were included, the same random slopes were fitted as in the eye movement analysis.

To test whether the accuracy scores could be predicted by reading times, we first calculated the correlations between the accuracy score and the timed eye movement measures (first pass time and total reading time). We only formally investigated this relation (by running statistical models with the time measures included as predictor) for significant correlations.

RESULTS

The data of four participants (three students in the studying condition, one student who read informatively) could not be included in the final dataset due to recording issues (mainly caused by head movements during the recording), leaving us with 76 participants. We first present a between-group comparison of the variables of the additional questionnaires to check whether we can assume the groups were equal (see Table 3). The Dutch and English LexTALE were analyzed with two-sample t-tests, the text perception and motivation measures with Wilcoxon signed rank tests (with continuity correction). A Dunn-Šidák correction for multiple testing was applied to determine significant differences ($\alpha = .00465$).

Table 3. *Between-group comparison of the reading goal groups on proficiency, text perception and motivation measures (standard deviations between brackets).*

	Reading group	**Study group**	**Test statistic**	***p*-value**
Dutch LexTALE (max = 100)	88.72 (5.54)	89.49 (5.08)	$t = 0.633$	.529
English LexTALE (max = 100)	71.86 (9.73)	74.51 (9.66)	$t = 1.193$	.237
Text Interest (max = 7)	3.81 (1.00)	4.50 (0.91)	$W = 1027$	.002
Content Difficulty (max = 7)	3.99 (0.70)	3.94 (0.92)	$W = 695$	.787
Prior Knowledge (max = 7)	2.03 (0.73)	1.81 (0.49)	$W = 860$	.151
L1 reading Motivation (max = 7)	5.43 (1.61)	5.87 (1.30)	$W = 624$	.292
L2 reading Motivation (max = 7)	4.92 (1.52)	5.14 (1.64)	$W = 636$	.363
L1 reading Self-Efficacy (max = 7)	5.68 (1.16)	6.11 (0.88)	$W = 560$	.074
L2 reading Self-Efficacy (max = 7)	4.49 (1.24)	4.67 (1.36)	$W = 645$	.413
L1 Perceived reading Importance (max = 7)	6.65 (0.59)	6.87 (0.52)	$W = 566$	.017
L2 Perceived reading Importance (max = 7)	6.41 (0.80)	6.51 (0.91)	$W = 637$	.308

Note: the significance level is at $\alpha = .00465$

The study group rated the texts as more interesting in comparison to the reading group. It could be that the study group found the texts more interesting because they analyzed their content more thoroughly. Furthermore, and interestingly, reading motivation was higher for L1 (M = 5.66, SD = 1.47) than for L2 (M = 5.03, SD = 1.57) across groups, which is in accordance to the studies of Vander Beken and Brysbaert (2017) and Vander Beken, Woumans, and Brysbaert (2017) (Wilcoxon test: V = 1352, p <.001).

Below, the results of the analyses of each dependent variable are presented. Planned comparisons were conducted if there was a significant interaction between two of the main predictors (language, reading goal and information centrality). In the case of a significant three-way interaction of a control variable with language and reading goal, a separate analysis was performed for (a) each language and (b) each reading goal. For each of these four resulting models, contrasts were run if the relevant interaction was significant to identify its specific pattern. The tables with all the final models are presented in Appendix 5A.

First pass time

The model outcomes for the first pass time are presented in Table A1. The main effect of reading goal was significant (β = -0.331, se = 0.130, t = -2.541, p < .05), showing that the first pass time was shorter when participants studied the texts than when just reading them, which is surprising. There was also a main effect of unit length (β = 0.068, se = 0.020, t = 3.396, p < .01), with a shorter first pass time for shorter units. Reading goal had a significant interaction with unit length (β = -0.052, se = 0.018, t = -2.839, p < .01). Post-hoc contrasts revealed that the first pass time was shorter for studying than reading when units contained at least 10 words (χ = 5.18, df = 1, p < .05).

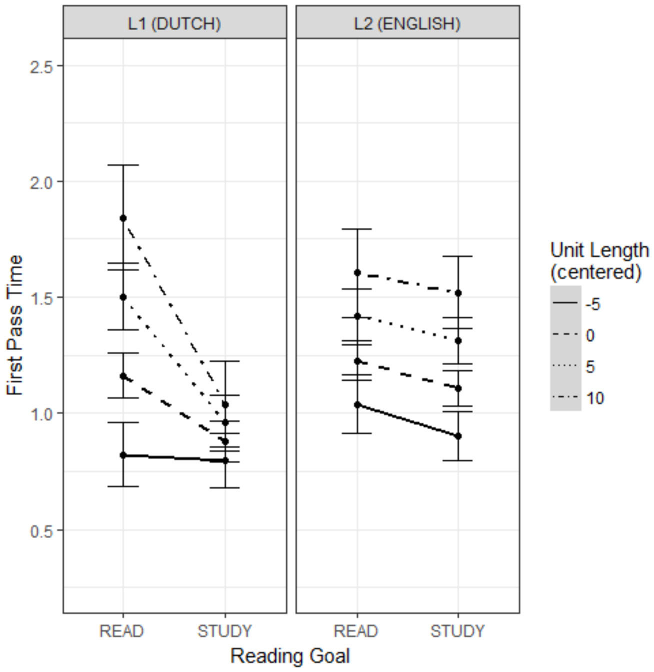

Figure 1. The interaction between reading goal (x-axis), language (panels) and unit length (lines) for first pass time (y-axis, in seconds). Error bars represent standard errors.

Finally, there was a significant three-way interaction between language, reading goal and unit length (β = 0.055, se = 0.024, t = 2.286, p < .05; see Figure 1). In the L1 analysis there was a significant interaction between reading goal and unit length (β = -0.052, se = 0.020, t = -2.605, p < .05): post hoc contrasts showed that the first pass time was shorter when studying, compared to reading in L1 if the unit contained at least 10 words (χ = 4.63, df = 1, p < .05). No significant interaction with unit length was found in the separate analysis for L2, reading or studying (all t < 1.199, p > .23).

Total reading time

The model outcomes for total reading time are presented in Table A2. The main effect of reading goal was significant (β = 6.765, se = 0.736, t = 9.190, p < .001), with a longer total reading time for studying compared to reading, as expected. There were also main effects of unit length (β = 0.230, se = 0.043, t = 5.355, p < .001) and average word length (β = 0.338, se = 0.155, t = 2.174, p < .05), showing that reading times became longer with an increasing unit length or average word length.

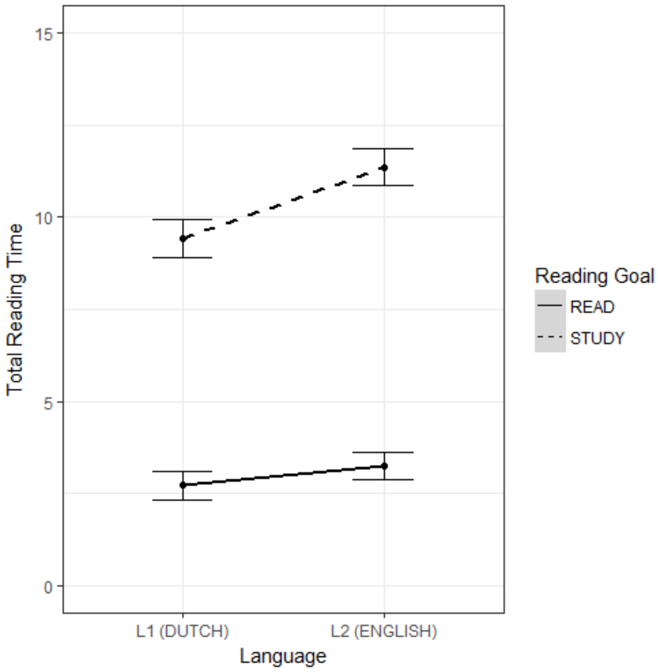

Figure 2. The interaction between language (x-axis) and reading goal (lines) for total reading time (y-axis, in seconds). Error bars represent standard errors.

The interaction between language and reading goal was significant (β = 1.638, se = 0.778, t = 2.105, p < .05; see Figure 2). Planned contrasts showed that (a) total reading times were somewhat faster for L1 reading than L2 reading (β = 0.541, se = 0.277, t = 1.954, p < .10), but this difference was larger between L1 and L2 studying (β = 1.912, se = 0.555, t = 3.443, p < .001); (b) total reading times were shorter for L1 reading than L1 studying (β = 6.711, se = 0.634, t = 10.590, p < .001), and again this effect was larger in L2 (β = 8.082, se = 0.614, t = 13.170, p < .001). Furthermore, reading goal interacted with average word length (β = 1.336, se = 0.388, t = 3.441, p < .001) and unit length (β = 0.439, se = 0.108, t = 4.062, p < .001). Post-hoc contrasts showed that total reading times were shorter for reading than studying when the average word length exceeded 3.222 letters (χ = 10.01, df = 1, p < .01; the difference between the reading goals became larger with an increasing average word length) or when the unit contained at least two words (χ = 5.89, df = 1, p < .05; the effect became larger with an increasing unit length).

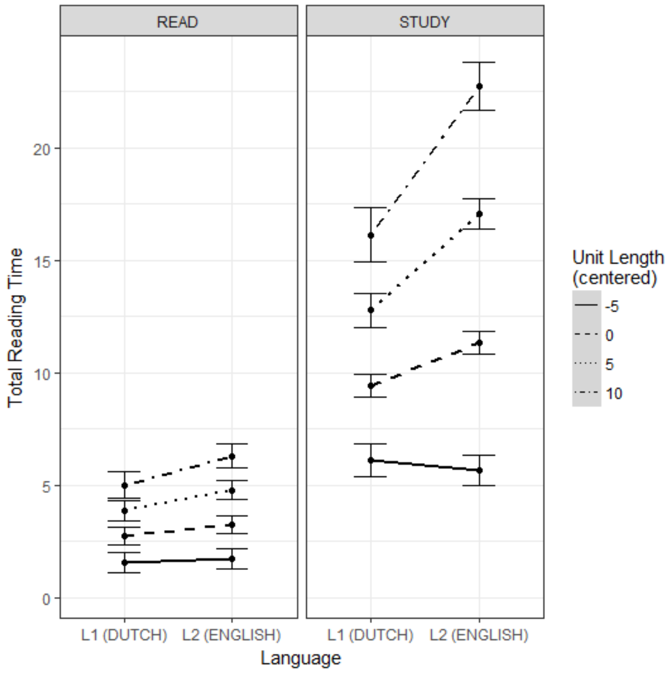

Figure 3. The interaction between language (x-axis), reading goal (panels) and unit length (lines) for total reading time (y-axis, in seconds). Error bars represent standard errors.

Finally, the three-way interaction between language, reading goal and unit length was significant (β = 0.395, se = 0.143, t = 2.757, p < .01; see Figure 3). In the separate analyses for reading goal, the interaction between language and unit length was significant for reading (β = 0.077, se = 0.030, t = 2.551, p < .05) and for studying (β = 0.472, se = 0.141, t = 3.333, p < .01). Contrasts revealed that total reading times were shorter for L1 than L2 if the unit contained at least 10 words, both for reading (χ = 5.64, df = 1, p < .05) and for studying (χ = 5.20, df = 1, p < .05). The interaction between reading goal and unit length was significant in both the L1 (β = 0.438, se = 0.062, t = 7.033, p < .001) and L2 analysis (β = 0.835, se = 0.052, t = 16.189, p < .001). The post-hoc contrasts showed that the total reading time was shorter for informative reading compared to studying if the unit contained at least 2 words in L1 (χ = 13.84, df = 1, p < .001) and 3 words in L2 (χ = 7.90, df = 1, p < .01). The difference between reading and studying became larger with an increasing unit length.

Fixation count

The model outcomes for fixation count are presented in Table A3. The main effect of reading goal was significant (β = 26.990, se = 0.2.819, t = 9.574, p < .001): units received more fixations for studying than for reading, as expected. The main effects of unit length (β = 0.946, se = 0.158, t = 5.970, p < .001) and average word length (β = 1.496, se = 0.577, t = 2.594, p <

.01) were also significant. The fixation count increased for longer units or a longer average word length. Reading goal interacted with average word length (β = 5.312, se = 1.432, *t* = 3.710, *p* < .001) and unit length (β = 1.898, se = 0.398, *t* = 4.771, *p* < .001). Post-hoc contrasts showed that the fixation count was higher for reading than studying when the average word length exceeded 3.222 letters (χ = 11.53, df = 1, *p* < .001; the magnitude of this difference increased when the average Length increased) or when the unit contained at least two words (χ = 5.48, df = 1, *p* < .05; the difference again became larger for longer unit lengths).

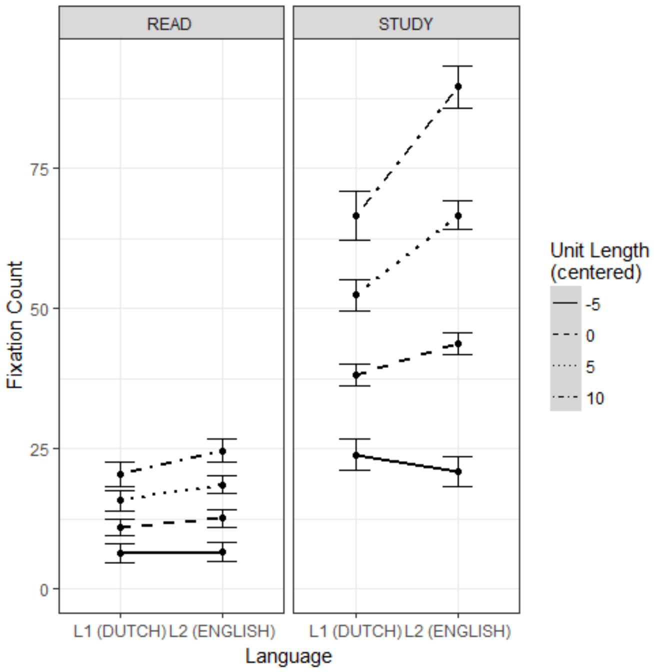

Figure 4. The interaction between language (x-axis), reading goal (panels) and unit length (lines) for fixation count (y-axis). Error bars represent standard errors.

Finally, the three-way interaction between language, reading goal and unit length was again significant (β = 1.470, se = 0.527, *t* = 2.788, *p* < .01; see Figure 4). In the separate analyses for reading goal, the interaction between language and unit length was significant for reading (β = 0.261, se = 0.114, *t* = 2.285, *p* < .05) and for studying (β = 1.734, se = 0.517, *t* = 3.352, *p* < .01). Contrasts revealed that the fixation count was lower for L1 than L2 if the unit contained at least 10 words for reading (χ = 5.79, df = 1, *p* < .05) and for studying (χ = 5.40, df = 1, *p* < .05). The interaction between reading goal and unit length was significant in both the L1 (β = 1.902, se = 0.234, *t* = 8.141, *p* < .001) and L2 analysis (β = 3.382, se = 0.189, *t* = 17.931, *p* < .001). The post-hoc contrasts showed that the more fixations were made while reading compared to studying if the unit contained at least 2 words in L1 (χ = 11.78, df = 1, *p* <

.001) and 3 words in L2 (χ = 5.50, df = 1, p < .05). The difference between reading and studying became larger with an increasing unit length.

Regression count

The model outcomes for regression count are presented in Table A4. For this model, reading goal could not be added as a random slope as this resulted in convergence errors. The main effect of reading goal was significant (β = 3.270, se = 0.290, t = 11.277, p < .001): units received more regressions for studying than for reading. Reading goal interacted significantly with average word length (β = 0.811, se = 0.132, t = 6.122, p < .001), average word frequency (β = 1.084, se = 0.334, t = 3.248, p < .01) and unit length (β = 0.234, se = 0.037, t = 6.354, p < .001). Post-hoc contrast showed that the regression count was higher for studying than reading if the average word length of the unit was 3.222 letters or more (χ = 11.60, df = 1, p < .001), and this difference increased for longer average lengths. More regressions for studying than reading were also made when the average Zipf word frequency was at least 5.431 (χ = 3.91, df = 1, p < .05), and this difference between studying and reading increased when word frequencies were higher. Finally, the regression count was higher for studying than reading when the unit length was at least 2 words (χ = 4.79, df = 1, p < .05), this effect became larger with an increasing unit length.

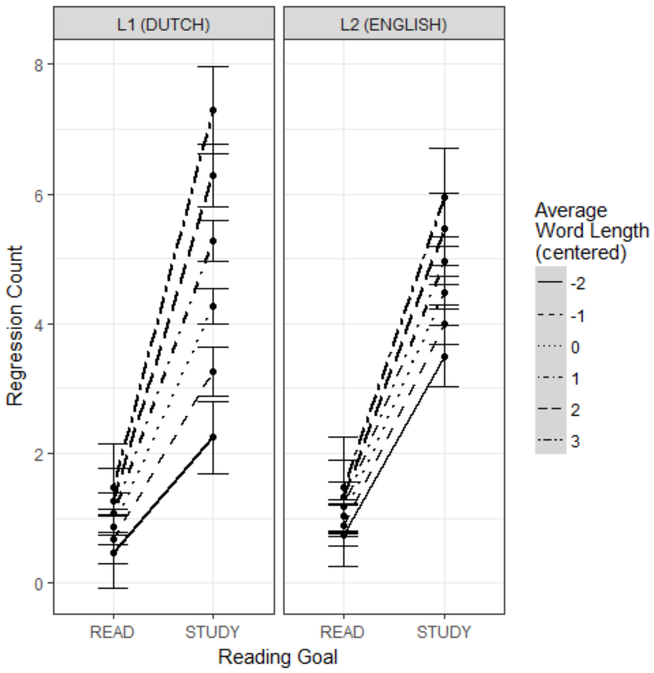

Figure 5. The interaction between reading goal (x-axis), language (panels) and average word length (lines) for regression count (y-axis). Error bars represent standard errors.

The three-way interaction between language, reading goal and average word length was significant (β = -0.469, se = 0.187, t = -2.501, p < .05; see Figure 5). In the separate analyses for L1, the interaction between reading goal and average word length was significant (β = 0.805, se = 0.142, t = 5.653, p < .001). Contrasts revealed that the regression count was lower for reading than studying if the unit had an average word length of 4.252 letters or more (χ = 31.52, df = 1, p < .001); the difference became larger with an increasing average word length. No significant interaction with average word length was found in the separate analysis for L2, reading or studying (all t < 1.157, p > .26).

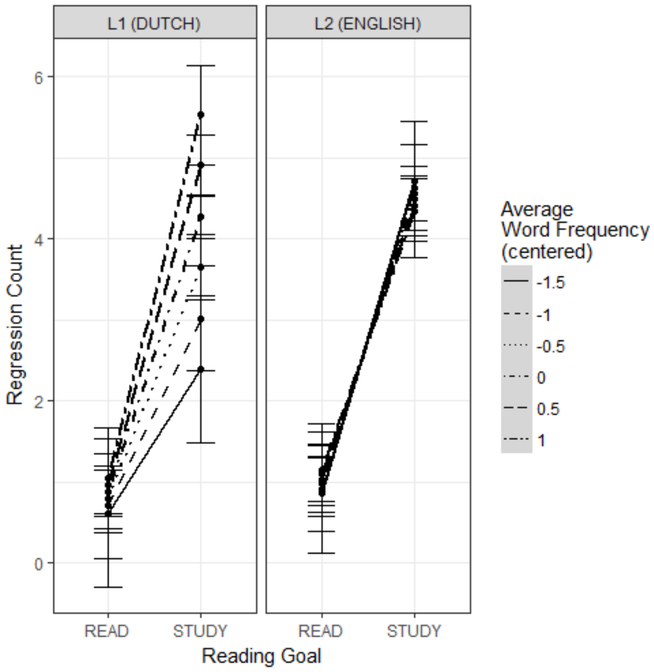

Figure 6. The interaction between reading goal (x-axis), language (panels) and average Zipf word frequency (lines) for regression count (y-axis). Error bars represent standard errors.

Finally, there was a significant three-way interaction between language, reading goal and average word frequency (β = -1.346, se = 0.441, t = -3.053, p < .01; see Figure 6). In the separate analyses for L1, the interaction between reading goal and average word frequency was significant (β = 1.087, se = 0.359, t = 3.032, p < .01). Contrasts revealed that the regression count was lower for reading than studying if the average Zipf word frequency of the unit was 4.639 or more (χ = 6.72, df = 1, p < .01). The difference between reading and studying became larger with increasing word frequency. No significant interaction with average word frequency was found in the separate analysis for L2, reading or studying (all t < 1.173, p > .24).

Saccadic amplitude

The model outcomes for saccadic amplitude are presented in Table A5. The main effects of reading goal (β = 0.307, se = 0.120, $t = 2.554$, $p < .05$) and language were significant (β = -0.267, se = 0.065, $t = -4.077$, $p < .001$; see Figure 7): Saccades were larger for studying than reading and smaller for L2 compared to L1. Furthermore, there was a significant main effect of L2 proficiency (β = 0.020, se = 0.008, $t = 2.320$, $p < .05$): the saccadic amplitude became larger when L2 proficiency of the participants.

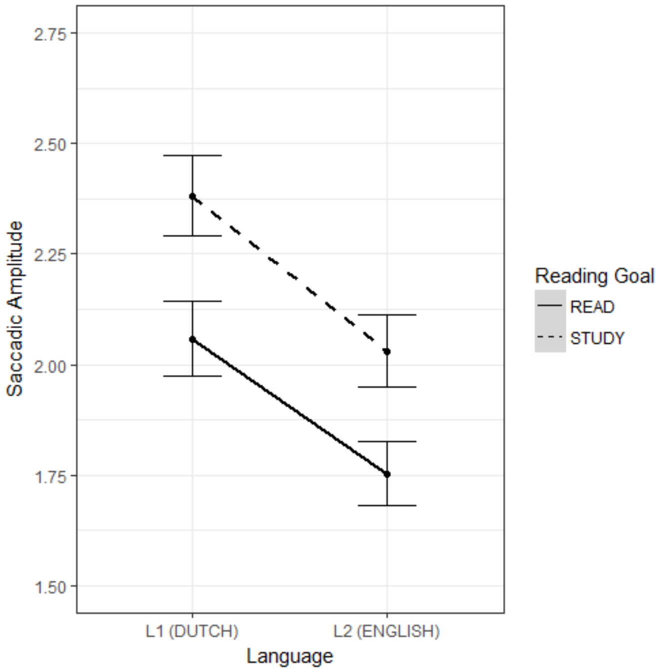

Figure 7. The interaction between language (x-axis) and reading goal (lines) for saccadic amplitude (y-axis, in degrees of visual angle). Error bars represent standard errors.

Reading goal interacted significantly with average word length (β = 0.092, se = 0.041, $t = 2.259$, $p < .05$) and average word frequency (β = 0.220, se = 0.108, $t = 2.033$, $p < .05$). Post-hoc contrasts showed that the saccadic amplitude was smaller for reading than studying when the average word length of the unit was at least 5.125 letters ($\chi = 3.85$, df = 1, $p < .05$) or when the average Zipf word frequency was at least 5.575 ($\chi = 3.84$, df = 1, $p < .05$). Finally, the three-way interaction between language, reading goal and L2 proficiency was significant (β = 0.013, se = 0.005, $t = 2.448$, $p < .05$). In the separate analysis for studying, the interaction between language and L2 proficiency was significant (β = 0.013, se = 0.003, t = 4.119, p < .001). Contrasts revealed that the saccadic amplitude was smaller for L2 compared to L1 if the LexTALE score was lower than 80.16 (χ = 3.85, df = 1, p < .05), which corresponds to an L2 proficiency effect on L2 studying, but not L1 studying. No significant interaction with L2 proficiency was found in the separate analysis for L1, L2, or reading (all t < 1.497, p > .14).

Accuracy score

The model outcomes for accuracy scores are presented in Table A6. No random slopes could be fitted in this model due to convergence errors. There was a main effect of reading goal (β = 0.935, se = 0.187, z = 5.005, p < .001; see Figure 8): accuracy scores were higher for studying than for reading. Furthermore, the main effects of content difficulty (β = -0.106, se = 0.038, z = -2.809, p < .01) and reading motivation (β = 0.087, se = 0.044, z = 2.010, p < .05) were significant. Accuracy scores were lower when participants perceived the texts as more difficult or when their motivation score for reading in the target language was lower.

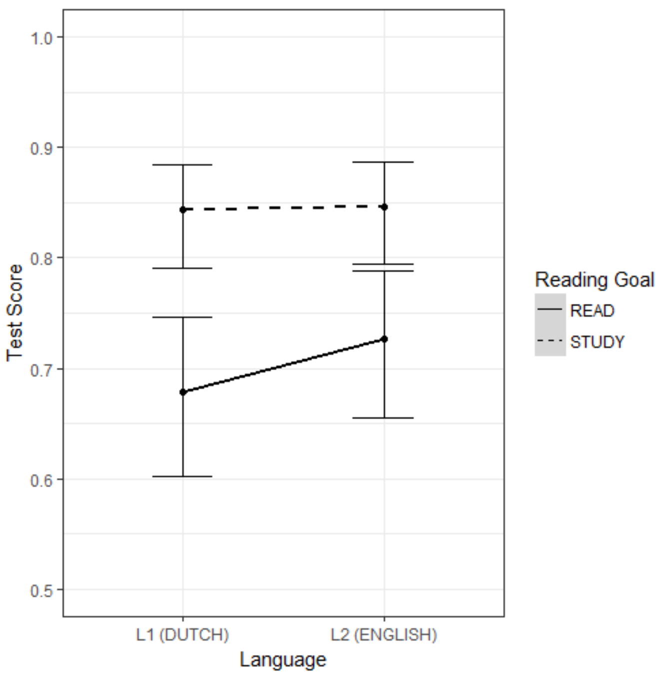

Figure 8. The interaction between language (x-axis) and reading goal (lines) for accuracy score (y-axis). Error bars represent standard errors.

We calculated the correlations between the two timed measures (first pass and total reading time) and accuracy scores in order to determine whether we would further investigate the relation between reading times and accuracy scores. The correlation of test accuracy with first pass time was not significant (r < .01, p = .866), whereas the correlation with total reading time was very small but significant (*r* = .07, *p* < .001). This association was however confounded by reading goal: none of the correlations between total reading times and accuracy score remained significant when we calculated them separately for each group (*r* = .02, *p* = .378 for reading; *r* = -.04, *p* = .125 for studying), resulting in our decision not to further formally investigate the influence of reading time on test scores.

DISCUSSION

In the current study, participants had to read or study four texts (two in L1, two in L2) while their eye movements were monitored. When they finished reading all four texts, all participants received tests with true/false judgements to examine their memory for the content of the texts. Applying this paradigm, we set out to answer four research goals: (a) to investigate whether there are differences in the eye movement patterns between L1 and L2 studying; (b) to examine the effects of different reading goals (informative reading vs. studying) on the eye movements within, and across each language; (c) to determine whether the memory trace for the text content is affected by reading goal, language or information centrality and (d) whether it is possible to predict accuracy scores of the test based on reading times. The results are highly relevant to EMI in higher education settings, as our participant sample, materials and recognition tests are representative to such a setting. We discuss the results in light of the research questions below.

The influence of language on eye movements while studying

In accordance with previous eye-tracking studies of written language processing in L1 and L2 (e.g., Cop, Drieghe, et al., 2015), we expected a general processing impairment in eye movement measures when studying in a L2. While no language effect was found in the first pass time of units, total reading times were 20% longer and 15% more fixations were made when studying in a L2 than when studying in L1. This effect interacted with unit length: if the number of words increased, the additional time needed to study the unit increased more in L2 than L1. Interestingly, in the sentence reading analysis of Cop, Drieghe, et al. (2015), an interaction between language and sentence length on reading times was also reported with a similar pattern. It seems that especially the processing of longer sentences, which are syntactically more complex, results in an additional difficulty for L2 studying. This is relevant for EMI in higher education, as academic textbooks often contain difficult texts, so that the cost of L2 use is large.

There was no effect of language on regression counts when studying, suggesting that the reading process is not more error-prone in L2. Indeed, the additional reading time seems to be sufficient for the verification of the content of the units and their integration in the context, without the need for further revisitations. The saccade length was influenced by L2 proficiency, but only in L2 studying: the average saccadic amplitude was larger (hence, bigger 'jumps' were made through the text) for participants with a higher L2 proficiency level.

We also expected an information centrality effect on the first pass time in the study condition in accordance to the results of Yeari et al., (2015), but the interaction between reading goal and information centrality was not significant here. In fact, there was no significant effect of information centrality on any measure (which is also discussed in more detail below).

To summarize, studying in L2 comes with an encoding cost, especially for complex text, as students specifically seem to have most difficulties processing longer sentences. The L2 – L1 differences in eye movement patterns closely resemble those of the previous sentence reading studies, showing more and longer fixations and smaller saccades in L2.

The influence of reading goal on eye movements

In the study of Yeari et al. (2015), differences were found between eye movement patterns of informative reading an L1 text and for preparation of a test. For our second research goal, we attempted to replicate these effects and wanted to investigate whether the influence of reading goal would be similar, or rather more pronounced, in a L2. The many significant differences in cognitive processes between the study and reading conditions confirm that our condition manipulation to induce studying was effective.

The first pass time was shorter for studying than reading; a higher-order interaction revealed that this effect was driven by L1, it was not significant in L2. Interestingly, Yeari et al. (2015) also reported a shorter first pass time when participants read the texts in preparation for a closed-question test compared to informative reading. This reading goal effect on L1 could be caused by specific reading strategies (e.g., participants who had to study the text perhaps browsed too quickly through the text to get familiar with the structure or content, after which they examined it more thoroughly in later re-readings. There were indeed more regressions made towards the texts for studying compared to reading in L1). In L2, participants either adopted another strategy, or they simply were not able to go faster through the text because of their slower processing capabilities.

In the total reading times, there was a large difference between reading goals: participants spent about 70% longer on the information units when they had to study the text than when they were just reading. The reading goal effect was larger for L2 than L1. As mentioned earlier, this effect was modulated by unit length: an increase in the number of words in the unit led to longer studying than reading times, and this additional processing time was larger for L2 than L1.

An almost identical pattern emerged in the fixation count: about 70% more fixations were made when participants studied the texts compared to reading, but again this reading goal effect was larger for L2 than L1. The unit length modulated this effect once more: when participants were studying the texts, the additional amount of fixations they made (compared to reading) towards more complex units increased more in L2 than L1.

More regressions were made while studying than reading the texts. This is consistent with the finding that longer reading times for studying than reading only emerged on total reading times, and not on first pass times. It informs us that participants who study, do so by reading the text multiple times, or look back in the text often, instead of spending a long time on each particular unit. The reading goal effect interacted with word length and frequency (in L1 only) and unit length (for both languages). A higher number of regressions were made towards longer, more complex units, but this effect was larger for studying than reading. Furthermore, in L1 studying the increase in regressions towards units containing longer and higher frequent words was higher than in L1 reading. Whitford and Titone (2012) also reported higher regression rates towards higher frequent words. They hypothesized that high frequency words are for example skipped more often, resulting in a higher need to revisit these words (although higher regression rates for low frequency words have also been demonstrated, see Dirix, Cop, Drieghe, & Duyck, 2017). As mentioned, the word length and frequency effect were not significant for L2. Usually, the magnitude of word-level effects is larger in L2 than L1 (e.g., Cop, Keuleers, et al., 2015; Whitford & Titone, 2012). It could be that in this particular studying context these variables have a lesser impact on L2 processing. Furthermore, the unit length already seems to be considerable source of influence on L2 eye movements. However, in general, there doesn't seem to be a large language difference on the number of regressions. Yeari et al. (2015) pointed out that regressions in sentence reading can for example be caused by a failure of or difficulties with sentence comprehension (Frazier & Rayner, 1982; Vauras, Hyona, & Niemi, 1992). This suggests that, after the initial longer processing time, comprehension is not more erroneous or more difficult in L2 compared to L1.[2]

The saccadic amplitude was smaller for L2 than L1 (cf. Cop, Drieghe, et al., 2015), but larger for studying than reading. We considered the possibility that this was caused by larger jumps through the text at a later stage in the study process due to a higher degree of familiarity with the material, or searching strategies for particular information units. An extra analysis of the saccades with the interaction between reading goal and an additional factor (saccade on the

[2] Note that there also was no significant difference on the subjective rating of the content difficulty of the texts between L1 (M = 3.79, SD = 1.18) and L2 (M = 4.16, SD = 1.18), Wilcoxon test: V = 1114, *p* >.05.

first pass vs. saccades on later passages) provided evidence for this hypothesis: on the first passage, there was no difference in saccadic amplitude between reading and studying, and there was indeed a significant difference on later passages ($\beta = 0.146$, se = 0.054, $t = 2.679$, $p < .01$). So, participants proceed through the text similarly when reading or studying at first, but because studying involves going through the same text repeatedly, further repetitions of the text imply larger jumps. Furthermore, the increase in saccadic amplitude when moving away from information units with higher word frequencies or longer words was larger for studying than reading.

As mentioned above, we found no effects of information centrality on any of the measures. We have three possible explanations to account for these null-effects in comparison to the studies of Yeari et al. (2015, 2017). First, we applied a different operationalization of 'information units' in comparison to Yeari et al.: whereas they included the main predicates and accompanying arguments, probably always resulting in a full sentence, we only included the smallest possible piece of unique information. For example, one of their central information units was: "Mount Vesuvius is a volcano located between the ancient Italian cities of Pompeii and Herculaneum". According to our definition of a unit, this sentence consists of two units: "Mount Vesuvius is a volcano"; "located between the ancient Italian cities of Pompeii and Herculaneum". It could be that readers rather pay more attention to whole sentences, which they judge to contain central information, instead of just looking for the most important parts within the sentence. This could have concealed information centrality effects in our study. Second, we applied a different approach in the rating of information centrality. Yeari et al. (2015) had three expert judges rate the centrality of each information unit, based on two criteria: the importance of this piece of information for the overall understanding of the text, and if the text understanding would be impaired if this piece of information would be missing. In our procedure, we combined the judgments of experts and our participant population, and we did not explicitly ask whether the understanding would be impaired if the unit was missing. Still, since we had a larger number of raters, our ratings ought to result in a more reliable estimation of what readers experience as important units. Finally, the difference of information centrality rating between our central and peripheral units was smaller than Yeari et al.'s (M = 4.12, SD = 0.26 vs. M = 4.7, SD = 0.3 for high centrality and M = 2.69, SD = 0.40 vs. M = 1.6, SD = 0.4 for low centrality in the current study and Yeari et al.'s, respectively). It could be that this difference was too small to result in significant differences on the eye movement measures.

In summary, we could only partially replicate the findings of Yeari et al. (2015) regarding L1 eye movement measures as a function of reading goal and information centrality. Differences between central and peripheral information units did not appear, but we did find some reading goal effects in various eye movement measures that are similar to those of Yeari et al.'s. In L1, the first pass time was shorter for studying than reading, but the total reading time was longer and more fixations were made, which indicates that students first go through the texts they need to study quickly and superficially, after which they begin studying the texts again. We expected a similar, yet somewhat inflated pattern for L2 studying vs. reading. L2 processing was indeed more effortful for studying than reading, and this difference was bigger than in L1 processing. Unit length had a large impact on this language difference, as more complex sentences required more additional effort to study in L2 than in L1. This shows that especially studying of long, difficult texts (like in EMI) comes with a cost in a L2, more than just reading in L2.

Memory for texts

Our third goal was to examine whether memory for texts was affected by reading goal, language, or information centrality. All participants received 40 true/false statements on the content of information units (the language was congruent with the texts they received). As expected, participants who studied the texts had a higher accuracy score on the tests than those who read them. There was again no effect of information centrality, but also no language effect: test performance was equal in L1 and L2. This suggests that with a cued recognition procedure, memory traces of studied texts are evenly accessible in both languages. These results are in line with those of Chen and Donin (1997), who reported longer study times for L2 but an equal performance on L1 and L2 tests. Although it is indeed more time-consuming and effortful to process texts in L2, the result is more or less the same, at least for this particular kind of information. Still, this means that the encoding process is hampered to some extent, and that readers use more time in L2 to compensate for that difficulty. These results in an EMI setting are a further confirmation of the findings of Vander Beken and Brysbaert (2017) and Vander Beken et al. (2017), who found no differences between L1 and L2 on recognition tests, even on a delayed test 30 days after studying the texts. So even the storage of the memory trace is not affected by the difficulties of L2 processing.

How is it possible that studying takes longer in L2, but that this compensatory strategy results in similar outcomes in both languages? There are two possible answers to this question. The first is related to the methodology of the study: similar to real-life studying, we did not impose a narrow time limit, so that sufficient time could be taken to compensate for L2 processing costs. At least this shows that when allowed, students take the necessary time, and they are good at guessing how much additional time is needed to reach a study goal that is similar to what they would achieve in their native language. Note that it was certainly not the case that no L2 effects on memory were observed because the test would be too easy: there is by far no ceiling effect in the scores. The second explanation can be found in the levels-of-processing framework (Craik & Lockhart, 1972). This framework assumes that the initial stage of encoding is aimed at surface form, while semantics are involved in the later stages. So the later stages, at which deeper encoding takes place, are responsible for the quality and strength of the memory trace, and for deeper understanding. At this point, the true/false statements come in: true/false judgements are basically a recognition task, which benefits from the knowledge of details rather than the ability to see the bigger picture (see also Vander Beken et al., 2017, for a discussion of this matter). Indeed, the levels-of-processing effect seems to be smaller in L2 compared to L1 (Francis & Gutiérrez, 2012): there is an L2 advantage for shallow processing tasks, such as word recognition, but this decreases for tasks that require deeper understanding. Or the other way around: L1 benefits more from deeper processing tasks than L2. As eye movements were recorded in this study, we can investigate whether any evidence is found for this theory. And indeed, the fact that there are more and longer fixations in L2 might indicate that a lot of attention is directed towards the lexical level, for a correct identification of the words (resulting in an equally strong memory trace for words and sentences in L2 as in L1), while the fact that saccades are smaller could indicate that the information is not integrated in the whole of text as much as in L1.

The relation between reading time and accuracy score

Finally, we were interested if it was possible to predict test scores of items related to an information unit by the reading times of that particular unit. As none of the correlations between the timed measures and accuracy scores were significant, we concur with Yeari et al. (2015) who state that "…attention allocation is only a minor factor in determining the memory strength of textual ideas." (p. 1088). Apparently, memory for academic studying is determined by factors that occur later than initial encoding.

Conclusion

We showed that participants need about 20% more time to study texts in a L2, compared to L1. Then, text content is equally well retained, resulting in similar scores for L1 and L2 recognition tests. The investigation of various eye movement measures has provided valuable insights in the strategies applied for different reading goals, and the similarities and differences between L1 and L2 studying. For example, when studying in L1, students seem to quickly scan the text on the first passage, but this does not seem to occur in L2 studying. The many interactions of language with unit length indicate that especially complex sentences are difficult to study in L2. This is highly relevant for EMI in higher education settings with non-native English bilinguals, where complex English academic texts are an important part of study material. Future research could investigate whether the L2 processing disadvantage could (partially) disappear if the same content is presented in shorter, more easily processed sentences.

REFERENCES

Andreassen, R., & Bråten, I. (2009). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading*, *33*(3), 263–283. https://doi.org/10.1111/j.1467-9817.2009.01413.x

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Chen, Q., & Donin, J. (1997). Discourse Processing of First and Second Language Biology Texts: Effects of Language Proficiency and Domain-Specific Knowledge. *The Modern Language Journal*, *81*(2), 209–227. https://doi.org/10.1111/j.1540-4781.1997.tb01176.x

Coiro, J. (2011). Predicting Reading Comprehension on the Internet. *Journal of Literacy Research*, *43*(4), 352–392. https://doi.org/10.1177/1086296X11421979

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

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

De Bruyne, E., Aesaert, K., & Valcke, M. (2017). THE DEVELOPMENT OF KNOWLEDGE TESTS WHEN STUDYING ACADEMIC TEXTS IN HIGHER EDUCATION. In *Edulearn17 9th International Conference on Education and New Learning Technologies* (pp. 10494–10507). Barcelona, Spain: IATED. https://doi.org/10.21125/edulearn.2017.0991

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. https://doi.org/10.1017/S1366728902273010

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

Donin, J., Graves, B., & Goyette, E. (2004). Second language text comprehension: Processing within a multilayered system. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, *61*(1), 53–76.

Drieghe, D., Brysbaert, M., Desmet, T., & De Baecke, C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology*, *16*(1–2), 79–103. https://doi.org/10.1080/09541440340000141

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, *38*(1), 78–103. https://doi.org/10.1598/RRQ.38.1.4

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

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, *1*(1), 3–14. https://doi.org/10.1016/j.edurev.2005.11.001

Francis, W. S., & Gutiérrez, M. (2012). Bilingual recognition memory: Stronger performance but weaker levels-of-processing effects in the less fluent language. *Memory & Cognition*, *40*(3), 496–503. https://doi.org/10.3758/s13421-011-0163-3

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

Gablasova, D. (2014). Learning and retaining specialized vocabulary from textbook reading: Comparison of learning outcomes through L1 and L2. *Modern Language Journal*, *98*(4), 976–991. https://doi.org/10.1111/modl.12150

Gollan, T. H., Fennema-Notestine, C., Montoya, R. I., & Jernigan, T. L. (2007). The bilingual effect on Boston Naming Test performance. *Journal of the International Neuropsychological Society*, *13*(2). https://doi.org/10.1017/S1355617707070038

Hyönä, J., & Niemi, P. (1990). EYE MOVEMENTS DURING REPEATED READING OF A

TEXT. *Acta Psychologica*, *73*, 259–280.

Joh, J. (2006). What Happens When L2 Readers Recall ? *Language Research*, *42*(1), 205–238.

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

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., & Keenan, J. M. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, *14*(2), 196–214. https://doi.org/10.1016/S0022-5371(75)80065-X

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

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–43. https://doi.org/10.3758/s13428-011-0146-0

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science*, *26*(1–2), 49–63.

Miller, A. C., & Keenan, J. M. (2009). How word decoding skill impacts text memory: The centrality deficit and how domain knowledge can compensate. *Annals of Dyslexia*, *59*(2), 99–113. https://doi.org/10.1007/s11881-009-0025-x

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. (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., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading*, *10*(3), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

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), 514–528. https://doi.org/10.1037/a0020990

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

https://doi.org/10.1111/j.1467-9280.2006.01693.x

Sandoval, T. C., Gollan, T. H., Ferreira, V. S., & Salmon, D. P. (2010). What causes the bilingual disadvantage in verbal fluency? The dual-task analogy. *Bilingualism: Language and Cognition*, *13*(2), 231. https://doi.org/10.1017/S1366728909990514

Stevens, M., Lammertyn, J., Verbruggen, F., & Vandierendonck, A. (2006). Tscope: A C library for programming cognitive experiments on the MS windows platform. *Behavior Research Methods, 38*(2), 280–286. https://doi.org/10.3758/BF03192779

Van Assche, E., Duyck, W., & Hartsuiker, R. J. (2012). Bilingual word recognition in a sentence context. *Frontiers in Psychology, 3,* 174. https://doi.org/10.3389/fpsyg.2012.00174

Van Assche, E., Duyck, W., Hartsuiker, R. J., & Diependaele, K. (2009). Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological Science*, *20*(8), 923–927. https://doi.org/10.1111/j.1467-9280.2009.02389.x

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The Landscape model of reading: Inferences and the online construction of memory representation. In H. van O. & S. R. Goldman (Ed.), *The construction of mental representations during reading.* (pp. 71–98). Lawrence Erlbaum Associates.

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

Vander Beken, H., & Brysbaert, M. (2017). Studying texts in a second language: the importance of test type. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728917000189

Vander Beken, H., Woumans, E., & Brysbaert, M. (2017). Studying texts in a second language: No disadvantage in long-term recognition memory. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728917000360

Vauras, M., Hyona, J., & Niemi, P. (1992). Comprehending coherent and incoherent texts: evidence from eye movement patterns and recall performance. *Journal of Research in Reading, 15*(1), 39–54. https://doi.org/10.1111/j.1467-9817.1992.tb00020.x

Whitford, V., & Titone, D. (2012). Second-language experience modulates first- and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review*, *19*(1), 73–80. https://doi.org/10.3758/s13423-011-0179-5

Yeari, M., Oudega, M., & van den Broek, P. (2017). The effect of highlighting on processing and memory of central and peripheral text information: evidence from eye movements. *Journal of Research in Reading*, *40*(4), 365–383. https://doi.org/10.1111/1467-9817.12072

Yeari, M., van den Broek, P., & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: evidence from eye-movements. *Reading and Writing*, *28*(8), 1071–1097. https://doi.org/10.1007/s11145-015-9561-4

CHAPTER 6

HOW WELL DO READING MEASURES CORRELATE? EFFECTS OF LANGUAGE CONTEXT AND REPEATED PRESENTATIONS.[1]

The present study assessed to what extent different reading time measures converge, using large databases of lexical decision times and eye tracking measures. We observed a low amount of shared variance between these measures, which limits the validity of lexical decision times for real-life reading. We also 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 accordance between eye tracking reading times of identical words in two different corpora is low, suggesting that the higher-order language context in which words are presented also plays a crucial role. Finally, our findings indicate that lexical decision times better resemble the average processing time of an increasing number of multiple presentations of the same word, across different language contexts.

[1] Dirix, N., Brysbaert, M., & Duyck, W. (2018). How well do reading measures correlate? Effects of language context and repeated presentations. *Manuscript submitted for publication.*

INTRODUCTION

In the domain of psycholinguistic research, and more specifically in the study of how we read words, 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 differences between word characteristics of the stimuli, such as the frequency or length of the words. The widespread use of this 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 processing, 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 sentence contexts increases correspondence to daily reading situations, where we read for meaning, rather than for lexicality (lexical decision). There are many timed dependent variables that can be derived from the eye movements: 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 only fixated 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 referred to as an "early" measure as it involves word identification. Total reading time is a "late" measure as it 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 in reading times between words. 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 early measures but not late measures or vice versa. In some rare cases however, completely opposite results were found between these 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).

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 in general, 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).

More recently, Kuperman, Drieghe, Keuleers, and Brysbaert (2013) built further upon Schilling et al.'s investigation by reanalyzing their dataset (with up-to-date word characteristics) and expanding their research to the analysis of three additional 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 for example 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 moderate amount of shared variance between lexical decision RTs and eye tracking reading times, ranging between 21% (additional dataset) and 45% (Schilling et al.’s data) for first fixation durations and 19% (additional dataset) and 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 amount of shared variance between lexical decision times and eye movement data to 1-15% for first fixation and 5-17% for gaze duration, indicating that word frequency and word length are the dominant factors in the correlations, but also that possibly very little common processes between lexical decision and eye tracking remain.

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 orthogonally manipulated 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 of 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 affects 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 on this topic).

Because of these issues, Kuperman et al. (2013) did not only assess the convergence between lexical decision and eye tracking in small-scale data (see above), but they also calculated correlations of reading times across lexical decision megastudies and eye tracking corpora. For lexical decision, data was obtained from the English Lexicon Project (ELP, Balota et al., 2007), in which RTs and accuracy scores for more than 40 000 words are collected. Eye tracking reading times were provided by the Dundee corpus (Kennedy & Pynte, 2005), an eye movement database of participants reading 20 newspaper articles (about 56 000 words in total). Kuperman et al. found a substantially lower correlation for the 6817 words common in these databases, compared to the correlations obtained in the factorial/single sentence experiments. The amount 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 the 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 findings: (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 in the contribution of the lexical decision task and eye tracking to study visual word recognition, they also identified some remaining concerns. For example, they 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 elaborated 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 another language, b) investigating convergence of paradigms in second-language (L2) reading for the first time 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 and lexical decision datasets.

First, 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. 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.

Second, we correlated 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 also very relevant to assess task convergence for L2 reading. If similar results are obtained in comparison to those in the L1 datasets, this would point towards similar general word recognition processes in L2 (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) can have 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 individual 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 had big effects in Kuperman et al. (2013), and in addition that 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 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 5012 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). 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 would be the first direct comparison of reliabilities of datasets of these paradigms; this could prove to be important as this could learn us whether low correlations can not only be explained in terms of little overlap in underlying processes, but also the potential low reliability of one of the measures. We estimated the reliabilities with the Intraclass Correlation Coefficient (ICC; McGraw & Wong, 1996; Shrout & Fleiss, 1979). As this coefficient is less sensitive to missing data (Courrieu, Brand-D'abrescia, Peereman, Spieler, & Rey, 2011), it seems to be perfectly suited for 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 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 within a total of approximately 2800 words (see Kennedy & 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 references 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 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 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, 2013) 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 2982 common words 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*

	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 multiple 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 late eye movement measures (TRT and GD) than for early measures (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.

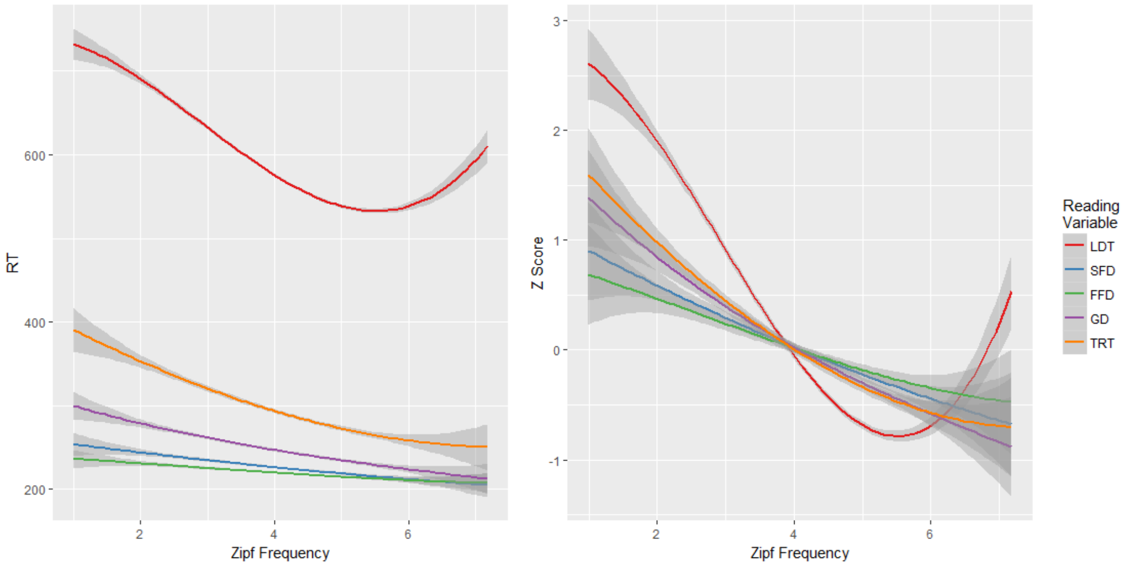

Figure 1. The effect of Zipf word frequency (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of BLP (LDT) and English monolingual GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

The effect of word length is plotted in Figure 2. Word length seems to have the largest impact on the late eye movement measures (TRT and GD), followed by LDT and the smallest effect is found in the early eye movement measures (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.

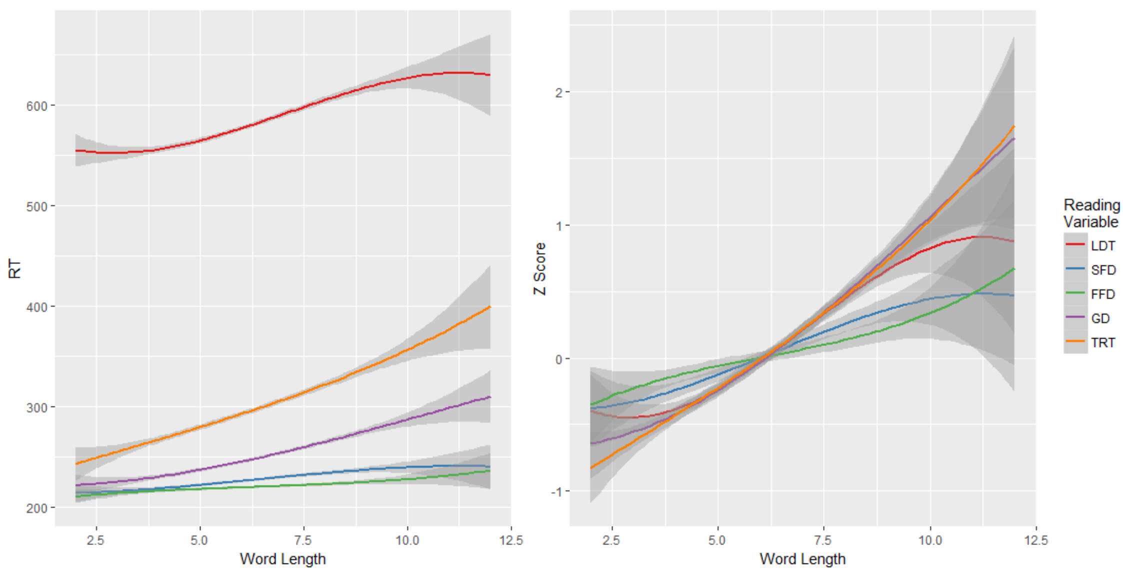

Figure 2. The effect of word length (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of BLP (LDT) and English monolingual GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

L1 Dutch reading. There were 3188 words 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*

	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 multiple 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 seems larger for lexical decision than for the eye tracking measures, and larger for the late eye movement measures (TRT and GD) than for early measures (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.

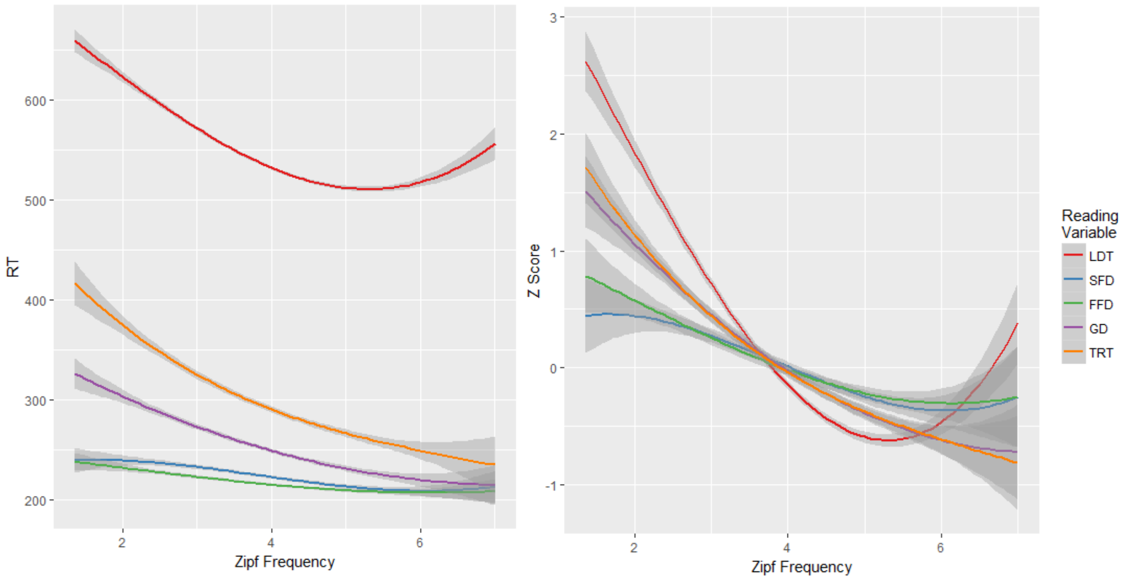

Figure 3. The effect of Zipf word frequency (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of DLP2 (LDT) and Dutch L1 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

The word length effect for the Dutch dataset is plotted in Figure 4. Word length again seems to have the largest impact on the late eye movement measures (TRT and GD), followed by LDT and the smallest effect is found in the early eye movement measures (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.

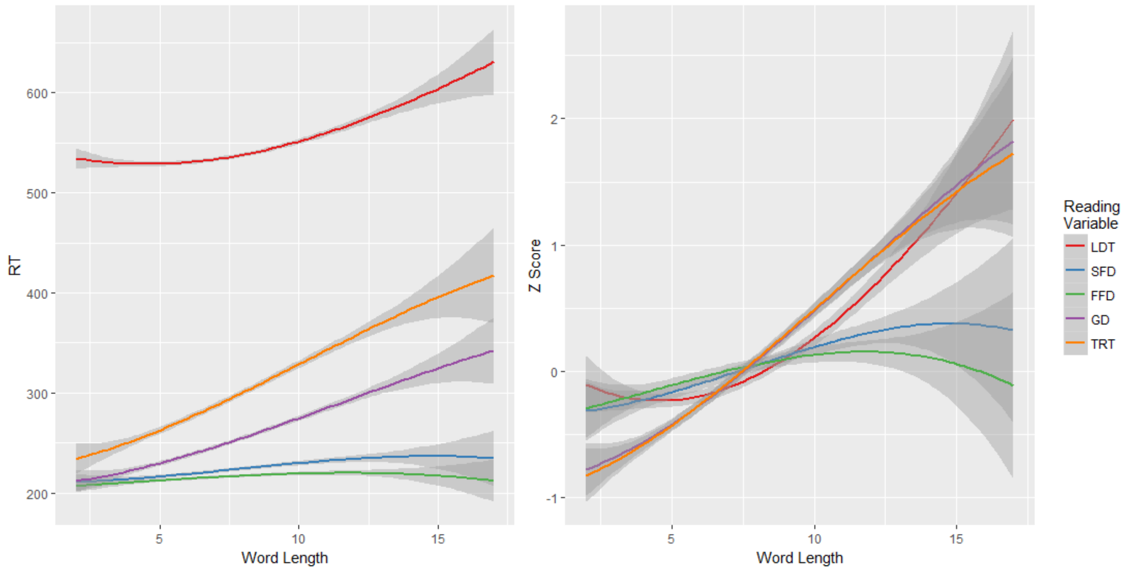

Figure 4. The effect of word length (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of DLP2 (LDT) and Dutch L1 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

L2 reading and lexical decision

There were 791 common words 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*

	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 multiple 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 the late eye movement measures (TRT and GD) than for early measures (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.

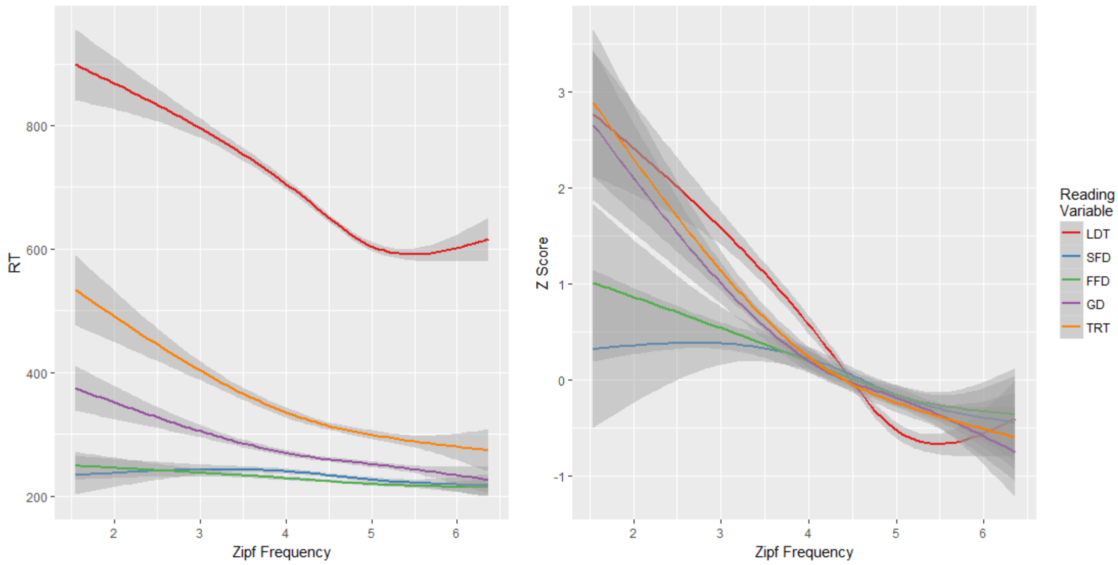

Figure 5. The effect of Zipf word frequency (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of the L2 lexical decision task (LDT) and English L2 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd 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 the late eye movement measures (TRT and GD), followed by LDT, and least on the early eye movement measures (SFD and FFD). The floor effect in LDT again appears for words up to 5 letters is present, 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.

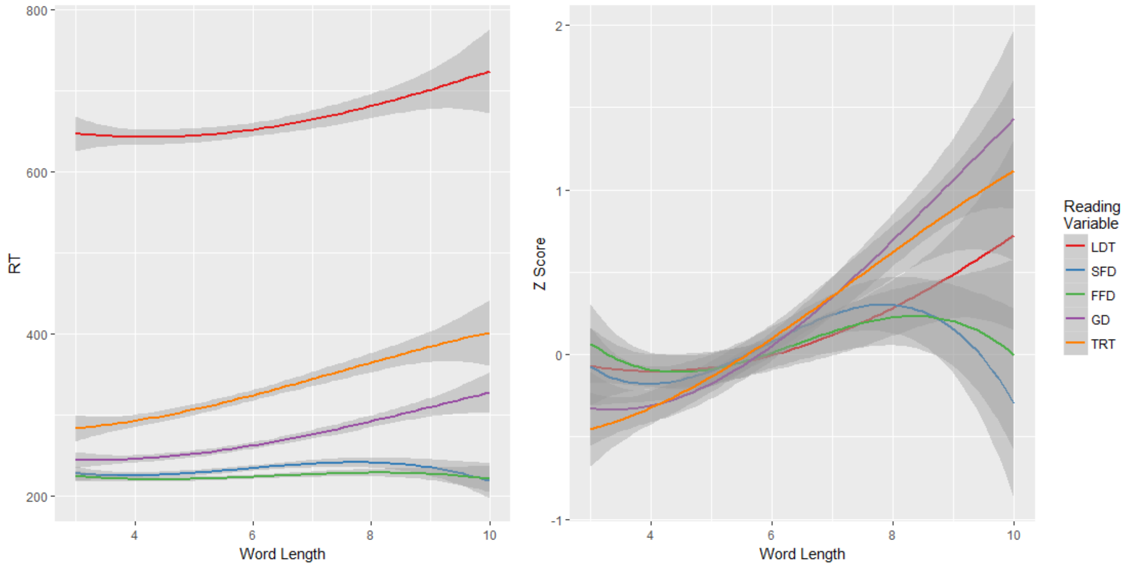

Figure 6. The effect of word length (x-axis) on raw data (in ms, left panel) and *z*-transformed data (right panel) of the dependent variables of the L2 lexical decision task (LDT) and English L2 GECO (SFD, FFD, GD, TRT). The grey bands indicate 95% confidence intervals. Polynomials of the 3rd degree.

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 1954 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 eye tracking corpora. Only the correlations for GD and TRT reached significance. The amount 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*

	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
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	—

Pearson correlations above the diagonal, *p* – values (Bonferroni-adjusted for multiple 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**.

First occurrence vs repeated presentations. We found 1915 words 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 early measures and higher correlations for late measures 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. Also 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 at approximately the same level for residualized values (26% - 32%).

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*

	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 multiple 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).

Reliabilities of the datasets

The low correlations between GECO and the lexical decision megastudies raises the question whether these are caused by the fact that they reveal different reading processes, or rather that (some of) the different measures may not very reliable. The ICC(C, 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 6. A consistent pattern turns up, indicating that the internal consistency of early eye movement measures (SFD and FFD) is 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 of the raw data [2]. This correction suggests that correlations are probably somewhat underestimated due to the internal consistencies of the datasets, although the correlations between LDT and the eye movement variables do not increase dramatically.

Table 6. *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	1.458	*.536*	.653	.476	.234	1.371	*.616*	.742	.542	.282	1.269	*.605*	.674	.441
GD	.381	.92	.954	*.875*	.779	.353	1.007	1.023	*.853*	.754	.337	.859	.927	*.874*	.743
TRT	.406	.690	.687	.880	*.894*	.406	.798	.731	.863	*.894*	.401	.677	.595	.835	*.906*

Pearson correlations above the diagonal, ICC(C, k) values on the diagonal, correlations corrected for attenuation below the diagonal. 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 languages, showing that the amount of shared variance between passage eye tracking reading times and lexical decision RTs is quite low, especially when controlling for the effects of word frequency and length. Second, we investigated L2 eye tracking reading times and RTs, to see whether similar results are found in L2 processing. The

[2] 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*.

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 of LDT with 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 in this case. 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 between .044 to .117. One possible reason might be the slightly better fit between databases because of the geographical correspondence of the experiments' 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 student in the Dundee study. This geographical correspondence has indeed been found earlier, as for example British SUBTLEX-UK (van Heuven et al., 2013) word frequencies could account 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 amount of word repetition in the texts of the corpora, which we discuss below.

In correspondence to 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) 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.

Convergence across reading tasks for L2 reading

Next to the L1 Dutch and English data, we also analyzed the convergence of 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 identical to that of L1 data, indicating that the influence of context and word characteristics manifest themselves in a similar way, 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). This further confirms the crucial role of words surrounding the target words, and predictability derived from the preceding words (spill-over effects) or the broader top-down language context of the narrative. In terms of eye movement control, it also points towards a parallel processing of words (Engbert et al., 2005; Kliegl et al., 2006).

Next, we found that averaging eye tracking reading times across repeated presentations, and hence across language contexts, increased correlations between LDT and GECO measures. As high correlations between these paradigms are only reported for studies with single sentence reading, where word-level characteristics are of more importance (e.g., Kuperman et al., 2013; Schilling et al., 1998) one could argue that eye movement research needs multiple presentations of target words in order to approximate effects of word-level variables like they are observed in lexical decision. However, just as well one could argue that the influence of such word-level variable effects is overestimated in lexical decision of isolated words, if eye tracking reveals that the top-down influences in natural text reading minimize such effects.

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 our English eye movement corpora contain approximately 56 000 words, but this corresponds to approximately 10 000 types for the Dundee corpus that Kuperman et al. analyzed, vs 5000 for our GECO. Indeed, in the subset of both datasets we analyzed in the current study, 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.

Word frequency and word length effects

The processing of words embedded in a sentence 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 previous studies, the word frequency effect for instance indeed is larger for lexical decision than for timed eye tracking measures (Kuperman et al., 2013), and also 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), 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). The word length effect seemed to have the largest and most linear effect in GD and TRT. In FFD and SFD the effect seemed to be even smaller than in LDT and also showed a ceiling effect, starting at around 9-10 letters. Word length has indeed been found to have a smaller impact on early than late measures (e.g., Kliegl et al., 2004), which can be related to reading strategy: longer words need more fixations to be processed entirely, and no additional information is gained while staying focused on the same position.

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.

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 rather whether some of the measures themselves may suffer from low psychometric reliability. To this end, we assessed and compared the reliability of the datasets analyzed in the current study. 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 than GD/TRT can probably be explained in terms of specific task demands of the lexical decision task (as discussed above). The high reliabilities in late eye 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 early eye movement measures. 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 due to task-specific processing demands and language context influences.

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.

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*(3), 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., 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

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

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

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

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. (2013). 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

CHAPTER 7

GENERAL DISCUSSION

Reading is a task we perform on a daily basis. We do it seemingly effortless, although the reading process is complex and can be influenced by multiple sources, such as the specific characteristics of the words or the linguistic context. In this dissertation, we aimed to answer several research questions related to (bilingual) reading. First, we investigated whether the well-established word-level AoA effect in psycholinguistics is also present eye movement pattern of participants reading a novel. From here on, the focus of the dissertation shifts to bilingual processing. Second, we wanted to investigate the AoA effect in bilingual reading to learn more about the origin of this effect: is it simply caused by the order in which we learn word forms, or by the organization of the underlying semantic network? Third, we studied cross-lingual neighborhood effects in lexical decision and natural reading, to investigate whether there is evidence for language non-selective access in bilinguals and also to investigate whether this would persist in a strict unilingual narrative context. Our fourth goal was to assess the impact of the previously reported inferior L2 processing on different purposes: informative reading vs. studying of texts. Finally, we wanted to investigate the comparability of two popular paradigms in the field of visual word recognition (the lexical decision task and eye-tracking), both in L1 and L2, but we also compared data of two eye-tracking corpora. By doing so, we wanted to assess the influence of (narrative) context on reading measures. We discuss these topics in more detail below.

CORPUS INVESTIGATIONS OF NATURAL READING

In four of the chapters of this dissertation, we used data of a corpus of eye movement data of participants reading an entire novel: the Ghent Eyetracking COrpus (GECO; Cop, Dirix, Drieghe, & Duyck, 2017). Mega studies and corpus linguistics seem to be increasingly popular in present-day research (see for example recent additions to the literature, such as the Dutch Lexicon Project 2 (DLP2), Brysbaert, Stevens, Mandera, & Keuleers, 2016; the eye-tracking Provo corpus, Luke & Christianson, 2017, etc.). The findings of CHAPTER 6 seem to at least suggest that an eye-tracking corpus can be a reliable dataset, and also show that this technique has certain advantages over paradigms such as the lexical decision task.

In CHAPTER 2, we investigated whether an AoA effect could be found in natural reading by investigating the monolingual part of GECO, and if so, how it would manifest itself in the various timed measures. We found a facilitating effect of AoA throughout the entire reading process, above and beyond the influences of other lexical variables, which was in accordance to previously reported single sentence studies (e.g., Juhasz & Rayner, 2006). Additionally, we found an interaction with word frequency in the later stages (the AoA effect in total reading times became larger with a decreasing word frequency). We suggested that these results could be taken into account in future versions of models of eye movement control such as E-Z reader (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Pollatsek, & Rayner, 2006). This model now mainly proposes word frequency and predictability as the main determiners of the duration of the familiarity and verification stages, but we believe that AoA could certainly be added to this list, as it has an undeniable impact on the reading process.

The monolingual AoA study is yet another example of an interesting and straightforward study that we can add to the expanding list of GECO studies (e.g., Cop, Dirix, Van Assche, Drieghe, & Duyck, 2017; Cop, Drieghe, & Duyck, 2015; Cop, Keuleers, Drieghe, & Duyck, 2015; Dotlačil, 2018; etc.), suggesting that this is a suitable, but also valuable instrument for exploratory or confirmatory analyses in reading research, of which we provide two more examples in the following paragraphs.

READING IN A SECOND LANGUAGE

The reading process in L2 has been shown to be quite different from that in L1 (e.g., longer and more fixations, smaller saccades and less skips, Cop, Drieghe, et al., 2015; larger frequency effects, Cop, Keuleers, et al., 2015; Duyck, Vanderelst, Desmet, & Hartsuiker, 2008). The data of GECO entails a unique part with data of bilingual participants reading half of the novel in their L1 and half in their L2. In CHAPTER 3, we presented a study using this part of the dataset, which includes the first investigation of L2 AoA in natural reading. As L2 learners learn new word forms for a semantic concept they already acquired in L1, this creates an interesting situation, where L2 words have an L2 AoA for their word form, but also an L1 AoA for the underlying semantic concept. In this study, we wanted to investigate how the AoA effect(s) would manifest themselves in L2 reading. Furthermore, this study contains valuable theoretical information on the origin of the AoA effect. The two major accounts that try to explain the cause of this effect are the mapping hypothesis (Ellis & Lambon Ralph, 2000) and the semantic hypothesis (Brysbaert, Van Wijnendaele, & De Deyne, 2000). The mapping hypothesis states that the AoA effect originates from the order in which the word forms are learned and entered into the network; if a new language is learned, a new network is installed, so no influence of the non-target language AoA should show up. The semantic hypothesis however states that the semantic concepts are the key in explaining the AoA effect: new concepts are linked to those we already know, so that earlier learned concepts have a more central place in the semantic network and are therefore easier accessible and activated.

The results of Experiment 2 indeed showed facilitatory within-language L2 AoA effects in L2 reading, and L1 AoA effects in L1 reading. However, we also found a facilitatory effect of

L1 AoA in early L2 reading processes, which was present for words that were more difficult to process (i.e., longer words). In terms of the two hypotheses, this could actually suggest that a combination of these accounts can explain the cause of the AoA effect: the order in which the word forms are learned is of a general importance to determine the degree to which it can exert its influence and take up a place in the network, but it is not unreasonable to think that the specific organization of the semantic network allows for earlier learned concepts to take up a more central place, also allowing them to be activated faster when encountered in a text.

In CHAPTER 4, we investigated cross-lingual neighborhood effects in an attempt to address an important theoretical issue related to the bilingual mind: do bilinguals have separate lexicons for each language, resulting in language selective lexical access, or do they have an integrated, shared lexicon for both their languages, accompanied by language non-selective lexical access? Previous research with cognates seems to point towards a non-selective access, as these identical words in two (or more) languages are read faster than equivalent non-cognate control words (e.g., Cop, Dirix, Van Assche, et al., 2017; Van Assche, Duyck, Hartsuiker, & Diependaele, 2009). However, there is still some discussion to the status of cognates: do they have a separate representation in each language, or do cognates have a shared representation for the two languages? A more conservative test of non-selective access would be by means of influences of cross-lingual, non-target language neighbors of target words on the reading times of these target words, but experimental behavioral evidence for such an effect has only been reported once in a series of lexical decision and progressive demasking tasks of van Heuven, Dijkstra, and Grainger (1998). In Experiment 1, we partially replicated their results with a generalized lexical decision task: inhibitory effects of cross-lingual, L1 neighbors were found on L2 lexical decision RTs and error rates. In Experiment 2, we investigated the bilingual GECO data to see whether cross-lingual neighborhood effects could be discovered in exclusive unilingual context of natural reading. Also here, in several of the L2 measures reflecting both early and late stage of the word recognition process, we discovered effects of the cross-lingual L1 neighborhood density, although they were of a facilitatory nature in the eye-tracking data. This further confirms the results of van Heuven et al. (1998), provides strong evidence for language non-selective access, and our results were also mostly compatible with an important model of bilingual word recognition, the BIA+ model (Dijkstra & van Heuven, 2002; although we suggested an adaptation in terms of the letter coding flexibility, given our results with deletion and transposition neighbors). There was however again a large contrast between the inhibitory effects found in lexical decision and the facilitatory effect in the eye-tracking results. Given the low convergence between these tasks reported in CHAPTER 6, and the possibility that effects of word characteristics are influenced to a different extent depending on the context, this did not entirely come as a surprise (in a recent study by Whitford and Titone, 2017a, facilitatory cross-lingual neighborhood effects were also reported in an eye-tracking study). In conclusion, it seems that even when a bilingual is reading a novel in L2, the unilingual context is not sufficient to counter cross-lingual L1 activation.

STUDYING IN A SECOND LANGUAGE

Finally, in CHAPTER 5 we investigated how the previously reported impaired L2 text reading processes impact studying in L2 in an EMI in higher education setting. In this more applied study of this dissertation, we asked participants to either informatively read or study four texts, half of them in L1 and half of them in L2, while their eye movements were monitored. This allowed us to estimate the effect of reading goal within each language on the reading pattern, but also to investigate the crucial L1 vs L2 studying conditions. An interesting set of results emerged from this study, showing quite substantial differences in L1 vs L2 studying: in L1, participants scan more quickly through the text on the first pass while studying in comparison to reading, but this difference is not present in L2; texts are fixates about 15% more and 20% longer in L2 studying than in L1 studying; and increasingly long, complex sentences need more additional time to process in L2 compared to L1, which is especially relevant for the complex material in academic handbooks. This pattern of results is in accordance with that of L1 vs L2 reading (e.g., Cop, Drieghe, et al., 2015).

We were also interested to see whether study language would influence the score on a recognition test on the content of the texts. There was however nog significant difference whatsoever between the L1 and L2 tests. This suggests that although the encoding stage does not run as efficiently in L2 as in L1, the (at least superficial) storage, retention and cued retrieval is quite similar in L1 and L2. These results are also similar to those of Vander Beken and Brysbaert (2017) and Vander Beken, Woumans, and Brysbaert (2017), who also didn't find L1 – L2 differences on recognition tests, even after a delayed test of up to 30 days. Furthermore, we tried to see whether we could predict the scores on questions about a specific passage by the fixation durations on this passage, but this does not seem to be the case.

In sum, we showed that the initial processing of a text when studying in L2 (specifically in an EMI context in higher education) is less efficient in comparison to studying in L1, but in the end, this does not seem to harm the academic performance on a recognition test. This was of course only a first study exploring the differences between studying in L1 or L2. Suggestions for future research, based on the current and previous chapters, are presented in the next part.

CORRESPONDENCE OF READING MEASURES

In accordance to previous research on the convergence between reading measures of different paradigms (Kuperman, Drieghe, Keuleers, & Brysbaert, 2013), in CHAPTER 6 we found that the correlations between lexical decision RTs of lexicon projects (Brysbaert et al., 2016; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and eye-tracking reading times of GECO (Cop, Dirix, Drieghe, et al., 2017) were quite low. The important influence of the language context cannot be captured by lexical decision, potentially leading into an overestimation of word characteristic effects in this paradigm. As context seems to have such a significant impact on eye-tracking reading measures, it also seems more plausible that words in a sentence or larger chunk of text are processed (at least partially) in parallel (Engbert, Nuthmann, Richter, & Kliegl, 2005; Kliegl, Nuthmann, & Engbert, 2006). The low correlations between lexical decision and eye-tracking data also suggests that these tasks do not necessarily (completely) tap into the same underlying processes. Results from studies of other chapters indeed do not always show a lot of correspondence between lexical decision RTs and eye-tracking reading times. However, the highest correlation, but also the highest degree of correspondence is usually found between RTs and total reading times (e.g., the lexical decision results of Izura & Ellis 2002, and the lexical decision task reported in Appendix 3B only show within-language AoA effects, as is the case in the L1 and L2 total reading times of CHAPTER 3, Experiment 1; the similar interaction between word frequency and AoA reported in the lexical decision tasks of Gerhand & Barry, 1999, and the monolingual total reading times in CHAPTER 2;etc.). This might suggest that common processes are possibly situated in the later stages of word recognition.

The late eye-tracking measures, total reading time and gaze duration also seem to be the most robust measures, as indicated by their high reliabilities. The results of our L2 analysis were very similar to those of L1. This suggests that in general, context and word characteristic effects manifest themselves in a comparable way, although L2 processing occurs at a slower rate and word-level effects tend to be more pronounced in comparison to L1 (Cop, Drieghe, et al., 2015; Cop, Keuleers, et al., 2015; Duyck et al., 2008).

The correlations of eye-tracking reading times of two corpora (GECO and Dundee, Kennedy & Pynte, 2005) were also extremely low, further conforming the importance of language context on the reading process. An important remark here is that we also found that repeated presentations of target words in the eye-tracking data increased the correlation with the lexical decision RTs, and that a single presentation is not a good predictor for multiple presentations. It seems important in eye-tracking studies to repeat the target words in different sentences, at different points in the study to get a good estimate of their true reading time.

GUIDELINES FOR FUTURE RESEARCH

First, we would like to suggest guidelines for future research related to GECO (Cop, Dirix, Drieghe, et al., 2017). This corpus of monolingual and bilingual reading has already proven its worth for the field of bilingual word recognition with studies on the cognate effect

(Cop, Dirix, Van Assche, et al., 2017), the word frequency effect (Cop, Keuleers, et al., 2015), a global comparison of sentence reading (Cop, Drieghe, et al., 2015) and the AoA and neighborhood studies reported in this dissertation. The attentive reader will have noticed that the majority of these studies are limited to the word level. Indeed, there are many more options to analyze GECO data on higher processing levels.

A first possible study is situated between the word- and sentence level, as it involves predictability. How predictable a word is, is a characteristic of the words itself, but depends on the context (i.e., the preceding words). It is defined as an important predictor of word recognition speed, which is reflected in shorter fixation times and a higher skipping probability for more predictable words (e.g., Kennedy, Pynte, Murray, & Paul, 2013; Rayner, Slattery, Drieghe, & Liversedge, 2011; Schuster, Hawelka, Hutzler, Kronbichler, & Richlan, 2016). A first objective could be to investigate the impact of the larger narrative context on predictability, in comparison to previous studies that often apply single sentence reading. In L2 reading, predictability is often operationalized as *semantic constraint* and is applied in the context of bilingual lexical access, for example to investigate whether a higher semantic constraint (or predictability) can modulate cognate effects (e.g., Titone, Libben, Mercier, Whitford, & Pivneva, 2011; Van Assche, Drieghe, Duyck, Welvaert, & Hartsuiker, 2011). Specific investigations into the difference in L1 – L2 predictability in a sentence context are scarce. A recent study of eye movement of paragraph reading showed no differences in predictability effects between L1 and L2 reading (Whitford & Titone, 2017b), but an EEG study suggest that L2 readers do not predict upcoming words as actively as L1 readers (Martin et al., 2013). Further research into this matter seems desirable, as there are indeed markers in previous research that could suggest different predictability effects in L1 and L2 (e.g., fewer skipping of words in L2 compared to L1, Cop, Drieghe, et al., 2015).

A study of predictability with GECO material would of course also entail collection of predictability ratings. A popular method to assess predictability of a word is collecting cloze ratings. It works as follows: participants are presented with (part of) the material, with words blanked out. They are required to make an estimated guess and fill in the blanks. The percentage of correct answers gives an estimation of how likely it seems that a certain word will follow, or in other words, how predictable it is (Taylor, 1953). This technique does however not seem to be suited to apply on the entire novel of GECO as it might be both time and money consuming (although the authors of the Provo corpus succeeded in collecting cloze ratings for all 2689 words in their eye-tracking corpus; Luke & Christianson, 2017). A collection of cloze ratings for a selection of target stimuli (such as nouns or verbs) or a single chapter of the novel might be feasible. An alternative option is to work with a computational approach, for example by training a language network, feeding the novel to the network, and calculating entropy (how uncertain is the model about the upcoming word?) and/or surprisal (how surprised is the model with the current word?) values as measures of predictability (e.g., Willems, Frank, Nijhof, Hagoort, & Van Den Bosch, 2016), based on *n* preceding words. A disadvantage would be that such a language model is probably not sensitive enough to pick up specific context information (e.g., it is a detective story, so certain unusual objects or characters might be more likely to appear in the story), but the advantage will be that predictability ratings are provided for each and every word in the corpus. The huge amount of data would also allow for a detailed analysis, including interactions with other word characteristics.

A second possible study could be on the sentence level, and investigate the way L1 and L2 readers parse sentences and deal with complex syntactic structures. Cop, Drieghe, et al. (2015) already showed that there are important differences in GECO on the sentence level between L1 and L2 reading in terms of fixation durations, fixation count, skips and saccadic amplitude. A lot of underlying word level characteristics that contribute to these differences have been identified, but the sentence level has not been studied extensively. Studies can for example be run to investigate if there are specific clauses that are more difficult to process in L2 than L1, to investigate whether reported difficulties in syntactic parsing for L2 readers of single sentences also apply to reading of a more extensive narrative (Frenck-Mestre, 2005; Papadopoulou & Clahsen, 2003), or to check if there are interactions between the lexical and syntactical level influencing the eye movement pattern (cf. Frenck-Mestre & Pynte, 1997).

A slightly ambitious plan would be to collect data for a new bilingual version of the corpus, with the same novel, but a different L1. English and Dutch are both Germanic languages and are quite close on the language family tree and in terms of lexical distance (Wilbert et al., 2013). A corpus with participants matched with the GECO participants on English as L2, but a more distant L1, perhaps even with a different script, could act as a database for replication of the currently reported effects, but it would also be interesting to see whether effects in L2 reading such as cross-lingual influences are modulated due to the lexical difference between languages

There are also possibilities to extend the research on studying texts in L1 and L2 as reported in CHAPTER 5. A striking result was the larger amount of additional time needed in for more complex sentences in L2 studying compared to L1 studying. This really seemed to have a large impact on the longer study times, so an intervention study could be set up to see if it would be possible to reduce some of the 20% additional study time needed in L2. This could for example be achieved by setting up a similar eye-tracking experiment: participants study texts in L1 and L2, but now there is an additional condition in which an L2 texts is provided with simplified sentences (i.e., cutting the long complex sentences in shorter pieces, while keeping the content identical). If sentence complexity puts additional strain on the L2 study process, the simplified L2 text should be processed faster than the original L2 text, and the eye movement pattern of studying the simplified L2 text should be more similar to that of L1 studying.

Another issue that could be investigated is the difference between L2 text processing of studying for open-ended questions (i.e., a recall test) and closed-ended questions (i.e., a recognition test). Indeed, Yeari, van den Broek, and Oudega (2015) reported differences in the L1 eye movement patterns for these different study goals, and Vander Beken and Brysbaert (2017) reported lower recall scores for L2 test than L1 test, but identical scores on recognition tests (cf. our results of CHAPTER 5). It could be worthwhile to identify potential processing difficulties specifically related to preparing for open-ended questions that could explain the recall cost. If no differences in the eye movement pattern between preparing for open- or closed ended questions should be found, it seems plausible that the cause of the recall cost is located at the retention or retrieval stage.

CONCLUSION

The studies of this dissertation contributed in several ways to the field of visual word recognition in psycholinguistics.

First, we showed that an eye movement corpus of book reading is suitable to run exploratory analyses, confirm established findings and extend them to the natural reading process or to second language processing. Furthermore, in our investigations of the GECO data, we were able to identify contributing factors to the ease or speed of processing, which could be implemented in models of eye movement control. Finally, we contributed to theoretical discussions on reading processes in general (the origin of the AoA effect) or bilingual word recognition specifically (language non-selective lexical access).

Second, we introduced eye-tracking in an investigation of studying in L2. We were able to pinpoint some important differences and disadvantages in comparison to L1 studying. Ultimately, these differences did not seem to harm the scores on the subsequent recognition tests, but they still might be part of the cause of a recall costs in L2.

Third, we showed that there is both a low convergence between lexical decision and eye movement measures, but also between eye-tracking corpora themselves, pointing out the crucial role of language context on the reading process. We also recommend to include repeated presentations of target stimuli throughout eye-tracking experiments.

REFERENCES

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

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*, 215–226.

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., Dirix, N., Van Assche, E., Drieghe, D., & Duyck, W. (2017). Reading a book in one or two languages? An eye movement study of cognate facilitation in L1 and L2 reading. *Bilingualism: Language and Cognition*, *20*(4), 747–769. https://doi.org/10.1017/S1366728916000213

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

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. https://doi.org/10.1017/S1366728902273010

Dotlačil, J. (2018). Building an ACT-R Reader for Eye-Tracking Corpus Data. *Topics in Cognitive Science*, *10*(1), 144–160. https://doi.org/10.1111/tops.12315

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

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103–1123. https://doi.org/10.1037/0278-7393.26.5.I103

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.

Frenck-Mestre, C. (2005). Eye-Movement Recording as a Tool for Studying Syntactic Processing in a Second Language: a Review of Methodologies and Experimental Findings. *Second Language Research*, *21*(2), 175–198. https://doi.org/Doi

10.1191/0267658305sr257oa

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic Ambiguity Resolution While Reading in Second and Native Languages. *The Quarterly Journal of Experimental Psychology A*, *50*(1), 119–148. https://doi.org/10.1080/027249897392251

Gerhand, S., & Barry, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, *27*(4), 592–602. https://doi.org/10.3758/BF03211553

Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica*, *23*, 245–281.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7–8), 846–863. https://doi.org/10.1080/13506280544000075

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

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. https://doi.org/10.1080/17470218.2012.676054

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., 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., 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

Luke, S. G., & Christianson, K. (2017). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, in press. https://doi.org/10.3758/s13428-017-0908-4

Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, *69*(4), 574–588. https://doi.org/10.1016/j.jml.2013.08.001

Papadopoulou, D., & Clahsen, H. (2003). Parsing Strategies in L1 and L2 Sentence Processing.

Studies in Second Language Acquisition, *25*(4), 501–528. https://doi.org/10.1017/S0272263103000214

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), 514–528. https://doi.org/10.1037/a0020990

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157. https://doi.org/10.1037/0033-295X.105.1.125

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4–22. https://doi.org/10.1016/j.cogsys.2005.07.002

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

Taylor, W. L. (1953). “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Titone, D., Libben, M., Mercier, J., Whitford, V., & Pivneva, I. (2011). Bilingual lexical access during L1 sentence reading: The effects of L2 knowledge, semantic constraint, and L1–L2 intermixing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1412–1431. https://doi.org/10.1037/a0024492

Van Assche, E., Drieghe, D., Duyck, W., Welvaert, M., & Hartsuiker, R. J. (2011). The influence of semantic constraints on bilingual word recognition during sentence reading. *Journal of Memory and Language*, *64*(1), 88–107. https://doi.org/10.1016/j.jml.2010.08.006

Van Assche, E., Duyck, W., Hartsuiker, R. J., & Diependaele, K. (2009). Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological Science*, *20*(8), 923–927. https://doi.org/10.1111/j.1467-9280.2009.02389.x

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

Vander Beken, H., & Brysbaert, M. (2017). Studying texts in a second language: the importance of test type. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728917000189

Vander Beken, H., Woumans, E., & Brysbaert, M. (2017). Studying texts in a second language:

No disadvantage in long-term recognition memory. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728917000360

Whitford, V., & Titone, D. (2017a). 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

Whitford, V., & Titone, D. (2017b). The effects of word frequency and word predictability during first- and second-language paragraph reading in bilingual older and younger adults. *Psychology and Aging*, *32*(2), 158–177. https://doi.org/10.1037/pag0000151

Wilbert, H., Jelena, G., Charlotte, G., Anja, S., Swarte, F., & Stefanie, V. (2013). Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In R. Gooskens, C., & van Bezooijen (Ed.), *Phonetics in Europe: Perception and Production* (Frankfurt, pp. 99–137). Peter Lang. https://doi.org/10.3726/978-3-653-03517-9/8

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van Den Bosch, A. (2016). Prediction during Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Yeari, M., van den Broek, P., & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: evidence from eye-movements. *Reading and Writing*, *28*(8), 1071–1097. https://doi.org/10.1007/s11145-015-9561-4

CHAPTER 8

ENGLISH SUMMARY

Reading is a task we perform on a daily basis: from the moment we start to learn to read, probably not a day goes by in our lifetime without reading. It seems to be an effortless task, although the reading process is complex, entails many sub-processes (from selecting the right letters to integrating words in a sentence) and can be influenced by multiple sources, such as specific word characteristics, the linguistic context, and the language we are reading in. In this dissertation, we aimed to answer several research questions related to (bilingual) reading processes. The attention for bilingual reading is highly relevant, given the large number of people who are bilingual nowadays (the European Union estimated that about 54% of the people in the EU are bi- or multilingual; European Union & European Commission for Education and Culture, 2012) and the increase of English as a Medium of Instruction (EMI) in (higher) education.

CHAPTER 2 was an investigation of the word-level age of acquisition (AoA) effect (earlier learned words are processed faster, e.g., Gerhand & Barry, 1999) in the Ghent Eyetracking COrpus (GECO; Cop, Dirix, Drieghe, & Duyck, 2017), an eye movement database of participants reading an entire novel. Part of the GECO participants were monolinguals who completed the novel in English, the other part were Dutch-English bilinguals who read half of the book in either language. As the purpose for the participants was to concentrate on reading the story, and nothing else, this type of research is referred to as "natural reading" for its close resemblance to daily life reading. We were curious whether we would be able to find a strong effect of AoA, a word characteristic, as the effect of context could heavily influence the reading times (cf. the results of CHAPTER 6). We found a facilitating effect of AoA throughout the entire reading process (i.e., shorter fixation durations in both early and late eye movement measures), above and beyond the influences of other lexical variables such as word frequency and length, which was in accordance to previously reported single sentence studies (e.g., Juhasz & Rayner, 2006). These results are also of importance to the development of models of eye movement control such as the E-Z reader model (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Pollatsek, & Rayner, 2006), as they currently do not take AoA explicitly into account as a predictor of the speed of fixations.

In CHAPTER 3 we continued our investigation of the AoA effect, but now shifted our attention to bilingual word processing. By learning L2 words, an interesting situation occurs: a bilingual learns a new word form for a concept that he or she already learned in L1. The result is that each L2 word has an L2 AoA for its word form, but also an L1 AoA for the underlying semantic concept. We compared the AoA effects in L1 and L2 reading measures in GECO, to assess whether it is the order of learning the word forms that causes the AoA effect (cf. the *mapping* hypothesis, Ellis & Lambon Ralph, 2000), because early learned words can benefit from the flexibility of an empty network. Another possibility is that the specific organization of the semantic network causes the AoA effect (cf. the *semantic hypothesis*, Brysbaert, Van Wijnendaele, & De Deyne, 2000): as we encounter new words, we will link the concept to the semantic knowledge we already possess, which causes earlier learned concepts to take up a more central place in the semantic network, making them more easily accessible. The mapping hypothesis would predict that only the L1 AoA would influence L1 reading, and L2 AoA would influence L2 reading, as the overall order of learning is most important. The semantic hypothesis however would predict an effect of L1 AoA on L2 reading, as the order in which the semantic concepts are learned in L1 is crucial. This issue was previously addressed by Izura and Ellis (2002) in a series of lexical decision tasks, who only found within-language AoA effects, which are in favor of the mapping hypothesis. We decided to run the first big data study of the L2 AoA effect by using the bilingual GECO data. Furthermore, we wanted to have the opportunity to study the influence of L1 and L2 AoA in a more detailed manner by investigating multiple eye-tracking measures. The results of this study showed a facilitating within-language L2 AoA effect in L2 reading throughout the entire reading process (shorter fixations for L2 words if they were learned earlier in L2), but there was also a facilitating effect of L1 AoA in early L2 reading processes. The early L1 AoA seemed to boost the recognition speed especially for words that are difficult to process, as the effect was only present for long words. In terms of the two hypotheses, this could suggest that a not one of the two, but possibly a combination of the discussed accounts can explain the cause of the AoA effect: the order in which the word forms are learned is of a general importance to determine the degree to which it can exert its influence and take up a place in the network, as we find distinct within-language AoA effects, but it is not unreasonable to think that the specific organization of the semantic network allows for earlier learned concepts to take up a more central place, also allowing them to be activated faster when encountered in a text and even boosting the activation of the corresponding L2 word form.

In CHAPTER 4, we addressed another interesting issue in the field of bilingualism: lexical access. The main question is whether bilinguals have a shared lexicon with a language non-selective access, or in other words: the language that is irrelevant at the moment of reading is still active, and can influence word recognition processes. The most conservative manner to investigate this, is by looking for influences of cross-lingual neighbors on the reading process. Orthographic neighbors are words that resemble each other closely, differing only in one letter.

Words can have both intra-lingual (e.g., *purse* and *nurse*), but here we are mainly interested in cross-lingual neighbors (e.g., *purse* and *puree*, mash in Dutch). In a generalized lexical decision (with both L1 and L2 stimuli in the same experiment), van Heuven, Dijkstra and Grainger (1998) found that Dutch – English bilinguals responded slower to English target words with many Dutch neighbors, compared to a small amount of cross-lingual neighbors, providing evidence for the non-selective lexical access and the shared lexicon. In Experiment 1 of CHAPTER 4, we partially replicated these result in an identical task, but more importantly, we also found facilitatory influences of L1 cross-lingual neighbors on L2 reading, both in early and late measures. It seems that even a strict unilingual narrative context cannot prevent the co-activation of the non-target language, providing even more convincing evidence for the notion of language non-selective access.

In CHAPTER 5 our goal was to assess the impact of previously reported inferior L2 processing in text reading (e.g., Cop, Drieghe, & Duyck, 2015) on different reading goals: informative reading vs studying texts in L2. As mentioned before, investigations like these are highly relevant given the increase of EMI in higher education, which for example has the consequence that students have to study more often in international English editions of handbooks. We asked participants to either informatively read or study four texts, half of them in L1 and half of them in L2, while their eye movements were monitored. This allowed us to estimate both the effect of reading goal within each language on the eye movement pattern, and the crucial L1 vs L2 studying conditions. The results showed quite substantial differences in L1 vs L2 studying: in L1, participants scanned more quickly through the text on the first passage while studying in comparison to reading, but this difference is not present in L2, possibly due to a ceiling effect for L2 reading (they simply could not go faster); texts are fixates about 15% more and 20% longer in L2 studying than in L1 studying; increasingly long, complex sentences need more additional time to study in L2 compared to L1, which is especially relevant for the complex material in academic handbooks. To make the context realistic, but also to be able to assess academic performance in L1 and L2, our participants received true/false recognition tests after studying the texts. Vander Beken and Brysbaert (2017) previously reported an L2 cost on a recall test, but equal scores for L1 and L2 recognition tests. Our results were similar, suggesting that the impairment of the initial processing stage does not harm the superficial storage, retention and retrieval processes required for a cued recognition.

Finally, in CHAPTER 6, we wanted to confirm the finding that there is only a little amount of shared variance between two popular paradigms in the field of visual word recognition: the lexical decision task and eye-tracking (Kuperman, Drieghe, Keuleers, & Brysbaert, 2013). The lexical decision task consists of the presentation of a strings of letters on a screen, after which the participant has to decide as fast and accurately as possible whether an existing word or a nonword is presented by pressing buttons. In eye-tracking, sentences or larger chunks of text are presented on a screen. Participants simply have to read; in the meantime, their eye movements are monitored. Multiple eye movement measures can be derived from the raw data, such as the duration of the first fixation on a word, the sum of all fixations, the number of fixations, etc. These measures are thought to reflect different processes in visual word recognition, such as early word recognition or late verification or integration (Boston, Hale, Kliegl, Patil, & Vasishth, 2008). We calculated correlations between reaction times (RTs) of lexical decision mega studies and reading times of an eye-tracking corpus, and indeed found that the convergence between these measures was dramatically low. This suggests that these task at least partially tap into different processes, and that eye-tracking reading measures are probably influenced less by word characteristics and more by context effects in comparison to lexical decision RTs. Similar results were obtained with data from first (L1) and second language (L2) datasets. We could exclude poor consistency of the datasets as a cause of the low correlations, as we calculated the reliability for each measure and these were fairly high. Furthermore, we compared data of two eye-tracking corpora to assess the impact of the context on reading measures. Surprisingly, the eye-tracking reading times of identical words, but embedded in sentences in a different general context do not correspond at all, pointing out a critical role of context in determining eye movement fixation times. Finally, the eye-tracking measures of a single presentation didn’t turn out to be good predictors of reading times of subsequent presentations, and repeated presentations raised the correlation with lexical decision, pointing out that it is important to repeatedly present target stimuli in different language contexts in eye-tracking experiments to enhance the quality of the measure.

The studies presented in this dissertation contributed in several ways to the field of visual word recognition in psycholinguistics. First, we showed that a corpus such as GECO, which entails eye movement data of participants reading an entire book, is suitable to run exploratory analyses, but also confirm established findings and even extend them to the natural reading and/or L2 processing. Furthermore, in our investigations of the GECO data, we could identify factors that clearly contribute to the duration of fixations, which could be implemented in models of eye movement control. We also contributed to theoretical discussions on reading processes in general (e.g., the origin of the AoA effect) or bilingual word recognition specifically (e.g., the shared lexicon and language non-selective lexical access). Second, we were the first to investigate the eye movement pattern of studying in L2. We were able to highlight some important differences and drawbacks in comparison to L1 studying, which ultimately did not seem to harm the performance on the subsequent recognition tests. Third, we showed that the convergence between lexical decision and eye movement measures, but also between eye-tracking corpora themselves is fairly low, pointing out the crucial role of language context on the reading process. We also recommend to include repeated presentations of target stimuli throughout eye-tracking experiments.

REFERENCES

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., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*, 215–226.

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

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103–1123. https://doi.org/10.1037/0278-7393.26.5.I103

European Union & European Commission for Education and Culture. (2012). Europeans and their languages.Retrieved from http://ec. europa.eu/public_opinion/archives

Gerhand, S., & Barry, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, *27*(4), 592–602. https://doi.org/10.3758/BF03211553

Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica*, *23*, 245–281.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7–8), 846–863. https://doi.org/10.1080/13506280544000075

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

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157. https://doi.org/10.1037/0033-295X.105.1.125

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4–22. https://doi.org/10.1016/j.cogsys.2005.07.002

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

Vander Beken, H., & Brysbaert, M. (2017). Studying texts in a second language: the importance of test type. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728917000189

CHAPTER 9

NEDERLANDSTALIGE SAMENVATTING

Lezen is een taak die we dagelijks uitvoeren: van het moment dat we beginnen met te leren lezen, gaat er waarschijnlijk geen dag meer voorbij in ons leven waarop we niet lezen. Het lijkt een erg gemakkelijke taak, hoewel het leesproces erg complex is en vele deelprocessen bevat (van het selecteren van de juiste letters tot het integreren van woorden in een zin) en kan beïnvloed worden door verschillende bronnen, zoals specifieke woordkarakteristieken, de linguïstische context en de taal waarin we lezen. In dit proefschrift probeerden we verschillende onderzoeksvragen te beantwoorden gerelateerd aan (tweetalige) leesprocessen. De aandacht die we besteden aan tweetalig lezen is uiterst relevant, gezien het grote aantal mensen die tegenwoordig tweetalig zijn (de Europese Unie schat dat momenteel 54% van de mensen in de EU twee- of meertalig zijn; European Union & European Commission for Education and Culture, 2012) en de toename van Engels als een Medium van Instructie (EMI) in het (hoger) onderwijs.

HOOFDSTUK 2 was een onderzoek naar het effect van verwervingsleeftijd (VL) op woordniveau (vroeger geleerde woorden worden sneller verwerkt, vb. Gerhand & Barry, 1999) in het Gentse Eyetracking COrpus (GECO; Cop, Dirix, Drieghe, & Duyck, 2017), een oogbewegingsdatabase van proefpersonen die een volledig boek lazen. Een deel van de GECO proefpersonen waren monolingualen die het boek in het Engels voltooiden, een ander deel waren Nederlands-Engels tweetaligen die een helft in iedere taal lazen. Omdat het doel voor de proefpersonen er enkel uit bestond om zich te concentreren op het lezen van het verhaal, en niets anders, wordt er naar dit onderzoek soms ook gerefereerd als "natuurlijk lezen", voor zijn grote gelijkenis met lezen in het dagelijkse leven. We waren benieuwd of we een sterk effect van VL konden vinden, een woordkarakteristiek, gezien de contexteffecten waarschijnlijk al zwaar zouden wegen op de leestijden (cfr. De resultaten uit HOOFDSTUK 6). We vonden een faciliterend effect van VL doorheen het hele leesproces (kortere fixaties in zowel vroege als late oogbewegingsmaten), bovenop de invloeden van andere lexicale variabelen zoals woordfrequentie en woordlengte, wat in overeenstemming was met voorgaande studies die enkelvoudige zinnen presenteerden (Juhasz & Rayner, 2006). Deze resultaten zijn ook van belang voor de ontwikkeling van modellen van oogbewegingscontrole zoals het E-Z reader model (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Pollatsek, & Rayner, 2006), omdat die momenteel VL nog niet expliciet als voorspeller van fixatiesnelheid opnemen.

In HOOFDSTUK 3 vervolgden we ons onderzoek naar het VL effect, maar nu verlegden we onze aandacht naar tweetalige woordverwerking. Door het leren van L2 woorden doet er zich een interessante situatie voor: aan tweetalige leert een nieuwe woordvorm voor een concept dat hij of zijn reeds heeft geleerd in L1. Het resultaat is dat elk L2 woord een L2 VL heeft voor de woordvorm, maar ook een L1 VL voor het onderliggende semantische concept. We vergeleken de VL effecten in L1 en L2 leesmaten in GECO om vast te stellen of het de volgorde waarin de woordvormen worden geleerd de oorzaak zijn van het VL effect (cfr. de *mapping* hypothese, Ellis & Lambon Ralph, 2000), omdat vroeg geleerde woorden kunnen genieten van de flexibiliteit van het lege netwerk waar ze binnenkomen. Een andere mogelijkheid is dat de specifieke organisatie van het semantische netwerk het VL effect veroorzaakt (cfr. de *semantische* hypothese, Brysbaert, Van Wijnendaele, & De Deyne, 2000): als we nieuwe woorden leren, linken we het onderliggende concept aan de semantische kennis waarover we reeds beschikken, wat maakt dat vroeger geleerde concepten een meer centrale plaats in het semantische netwerk zullen opnemen en daarom sneller toegankelijk zullen zijn. De mapping hypothese voorspelt dat enkel de L1 VL een effect heeft op L1 lezen, L2 VL zou enkel een invloed op L2 lezen, gezien de algemene volgorde van leren het meest belangrijk is. De semantische hypothese voorspelt daarentegen een effect van L1 VL op L2 lezen, gezien de volgorde waarin de semantische concepten in L1 zijn geleerd cruciaal zijn. Dit onderwerp werd al eerder bestudeerd door Izura en Ellis (2002) in een serie van lexicale decisietaken, waarbij ze enkel VL effecten vonden binnen de talen, wat evidentie biedt voor de mapping hypothese. Wij besloten om de eerste big data studie van het L2 VL effect uit te voeren door gebruik te maken van de tweetalige GECO data. Bovendien wilden we de mogelijkheid hebben om de invloed van L1 en L2 VL in meer detail te bestuderen door meerdere oogbewegingsmaten te onderzoeken. De resultaten van deze studie toonden aan dat er een faciliterend binnen-taal effect van L2 VL op L2 lezen was doorheen het hele leesproces (kortere fixaties voor L2 woorden als ze vroeger werden geleerd in L2), maar er was ook een faciliterend effect van L1 VL op de vroege L2 leesmaten. De vroege L1 VL blijkt een boost te geven aan L2 woorden die moeilijker te verwerken zijn, gezien het effect er enkel was voor lange L2 woorden. In het kader van de twee hypotheses suggereert dit dat niet slechts één van de twee, maar mogelijks een combinatie van de besproken hypotheses de oorzaak van het VL effect kan verklaren: de volgorde waarin de woordvormen zijn geleerd is van algemeen belang om te bepalen in welke mate het een invloed kan uitoefenen en een plaats kan innemen in het netwerk, gezien we goed te onderscheiden binnen-taal VL effecten vinden, maar het is niet onredelijk te denken dat de specifieke organisatie van het semantische netwerk toelaat dat vroeg geleerde concepten een meer centrale plaats hebben, waaruit volgt dat ze sneller geactiveerd kunnen worden als ze worden tegengekomen in een tekst en dat ze zelfs de activatie van een corresponderend L2 woord kunnen boosten.

In HOOFDSTUK 4 richtten we ons op een ander interessant onderwerp in het onderzoeksveld van tweetaligheid: lexicale toegang. De hoofdvraag hier is of tweetaligen een gedeeld lexicon hebben met een niet-taalselectieve toegang, of in andere woorden: de taal die tijdens het lezen irrelevant is, is nog steeds actief en kan ook de visuele woordherkenningsprocessen beïnvloeden. De meest conservatieve manier om dit te onderzoeken is door op zoek te gaan naar invloeden van cross-linguale buren op het leesproces. Orthografische buren zijn worden die heel hard op elkaar lijken en verschillen maar in één letter. Woorden kunnen een intra-linguale buur hebben (vb. *balk* en *valk*), maar hier zijn we vooral geïnteresseerd in de cross-linguale buren (vb. *puree* in het Nederlands en *purse* in het Engels). In een gegeneraliseerde lexicale decisietaak (met zowel L1 als L2 stimuli in hetzelfde experiment) vonden van Heuven, Dijkstra, en Grainger (1998) dat Nederlands – Engelse tweetaligen trager reageerden op Engelse targetwoorden met veel Nederlandse buren in vergelijking met woorden met een klein aantal cross-linguale buren, wat evidentie biedt voor de niet-taalselectieve toegang en het gedeelde lexicon. In Experiment 1 van HOOFDSTUK 4 repliceerden we gedeeltelijk deze resultaten met een identieke taak, maar belangrijker: we vonden ook faciliterende invloeden van L1 cross-linguale buren op L2 lezen, zowel in vroege als in late oogbewegingsmaten. Het lijkt zo te zijn dat zelfs een strikt unilinguale verhaalcontext niet kan voorkomen dat de niet-targettaal ook wordt geactiveerd, wat zelfs nog sterkere evidentie biedt voor de niet-taalselectieve toegang.

In HOOFDSTUK 5 was ons doel om de impact van vroeger gerapporteerde inferieure L2 verwerking van tekst lezen (Cop, Drieghe, & Duyck, 2015) op verschillende leesdoelen in te schatten: informatie lezen vs studeren in L2. Zoals eerder vermeld zijn onderzoeken zoals deze uiterst relevant gezien de toename van EMI in hoger onderwijs, waarbij er bijvoorbeeld het gevolg is dat studenten vaker moeten studeren uit internationale, Engelstalige handboeken. We vroegen aan proefpersonen om vier teksten ofwel informatief te lezen, of anders om ze te studeren. De helft van de teksten werd in L1 aangeboden, de andere helft in L2, tijdens het verwerken van de teksten werden de oogbewegingen opgenomen. Dit liet ons toe om zowel het effect van leerdoel binnen elke taal als die van de cruciale L1 vs L2 conditie op het oogbewegingspatroon vast te stellen. De resultaten toonden vrij grote verschillen tussen L1 en L2 studeren: in L1 scanden proefpersonen die moesten studeren sneller een eerste keer de tekst dan proefpersonen die informatief lazen, maar dit verschil werd niet teruggevonden in L2, mogelijks door een plafondeffect van L2 lezen (ze konden simpelweg niet sneller lezen); er werden ongeveer 15% meer en 20% langere fixaties gemaakt in L2 studeren in vergelijking met L2 studeren; bij een toenemende zins-lengte en –complexiteit hebben proefpersonen in L2 meer extra verwerkingstijd nodig dan in L1, wat zeker relevant is voor complex materiaal zoals academische handboeken. Om de context realistisch te maken, maar ook om te kunnen inschatten wat de academische performantie in L1 en L2 is, vervolledigden onze proefpersonen ook nog testen met waar/vals herkenningsvragen na het studeren. Vander Beken en Brysbaert (2017) rapporteerden reeds een L2 kost bij een herinneringstest, maar gelijke scores voor L1 en L2 herkenningstesten. Onze resultaten zijn gelijkaardig, wat suggereert dat de moeizamer verlopende initiële verwerking de oppervlakkige opslag-, behoud- en ophalingsprocessen die benodigd zijn voor een gecuede herkenning niet schaadt.

Ten slotte wilden we in HOOFDSTUK 6 de bevinding bevestigen dat er slechts een kleine hoeveelheid gedeelde variantie is tussen twee populaire paradigma's in het onderzoeksveld van visuele woordherkenning: de lexicale decisietaak en eyetracking (Kuperman, Drieghe, Keuleers, & Brysbaert, 2013). De lexicale decisietaak bestaat uit de presentatie van een reeks van letters op een scherm, waarbij de proefpersoon zo snel en accuraat mogelijk moet beslissen of het om een bestaand woord of nonwoord gaat door op de juiste knop te drukken. Bij eyetracking worden zinnen of langere stukken tekst gepresenteerd op een scherm. Proefpersonen moeten simpelweg de tekst lezen, in tussentijd worden hun oogbewegingen opgenomen. Meerdere oogbewegingsmaten kunnen worden afgeleid van de ruwe data, zoals de duur van de eerste fixatie op een woord, de som van alle fixatiaduren, het aantal fixaties, etc. Deze maten worden geacht verschillende processen in visuele woordherkenning te representeren, zoals vroege woordherkenning of late verificatie of integratie (Boston, Hale, Kliegl, Patil, & Vasishth, 2008). We berekenden correlaties tussen reactietijden (RTs) van lexicale decisie megastudies en leestijden van een oogbewegingscorpus, en vonden inderdaad dat de overeenkomst tussen deze maten dramatisch laag was. Dit suggereert dat deze taken op zijn minst deels gelinkt zijn aan verschillende onderliggende processen, en dat oogbewegingsmaten waarschijnlijk minder beïnvloed worden door woordkarakteristieken, maar meer door contexteffecten in vergelijking met lexicale decisie RTs. We vonden gelijkaardige resultaten in datasets van de eerste (L1) en tweede (L2) taal. We konden een zwakke consistentie als verklaring voor de lage correlaties uitsluiten, gezien we de betrouwbaarheid van elke maat berekenden en deze tamelijk hoog waren. We vergeleken daarnaast ook nog de data van twee oogbewegingscorpora om de impact van de context op leesmaten verder te bepalen. Verrassend genoeg blijkt dat de oogbewegingsleestijden van identieke woorden, maar die in zinnen staan in een verschillende overkoepelende context, totaal niet overeenkomen, wat nogmaals wijst op de kritieke rol van de context in het bepalen van fixatieduren van oogbewegingen. Ten slotte bleek dat oogbewegingsmaten van een enkele presentatie van een woord geen goede predictor blijkt te zijn voor de leestijden van volgende presentaties van dat woord, en dat herhaalde presentaties de correlatie met lexicale decisie doet stijgen, wat er op wijst dat het ook belangrijk is om meermaals de targetstimuli te presenteren, in verschillende taalcontexten in oogbewegingsexperimenten, om de kwaliteit van de data te verbeteren.

De studies die werden gepresenteerd in dit proefschrift hebben op verschillende manieren bijgedragen aan het onderzoeksveld van visuele woordherkenning in de psycholinguïstiek. Ten eerste toonden we dat een corpus zoals GECO, dat bestaat uit oogbewegingsdata van participanten die een volledig boek lazen, zeker geschikt is om exploratieve analyses te runnen, maar ook om gevestigde resultaten te bevestigen en ze zelfs uit te breiden naar natuurlijk lezen en/of L2 verwerking. Bovendien konden we in onze onderzoeken met GECO data enkele factoren identificeren die duidelijk bijdragen aan fixatietijden, wat geïmplementeerd kan worden in modellen van oogbewegingscontrole. We hebben ook bijgedragen aan de theoretische discussie over leesprocessen in het algemeen (vb. de oorsprong van het VL effect), maar ook tweetalige woordherkenning specifiek (vb. het gedeelde lexicon en niet-taalselectieve toegang). Ten tweede hebben wij als eerste de oogbewegingspatrone van L2 studeren onderzocht. We konden enkel belangrijke verschillen en mogelijke nadelen blootleggen in vergelijking met L1 studeren, maar dit leed uiteindelijk niet tot een verminderde performantie in de daarop volgende herkenningstesten. Ten derde hebben we aangetoond dat de overeenkomst tussen lexicale decisie en oogbewegingsmaten, maar ook tussen oogbewegingscorpora onderling, erg laag is. Dit toont aan dat de taalcontext een cruciale rol speelt in het leesproces. We raden ook aan om steeds meerdere presentaties van de targetstimuli doorheen oogbewegingsexperimenten in te voeren.

REFERENTIES

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., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*, 215–226.

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). 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

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103–1123. https://doi.org/10.1037/0278-7393.26.5.I103

European Union & European Commission for Education and Culture. (2012). Europeans and their languages.Retrieved from http://ec. europa.eu/public_opinion/archives

Gerhand, S., & Barry, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, *27*(4), 592–602. https://doi.org/10.3758/BF03211553

Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica*, *23*, 245–281.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7–8), 846–863. https://doi.org/10.1080/13506280544000075

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

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157. https://doi.org/10.1037/0033-295X.105.1.125

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4–22. https://doi.org/10.1016/j.cogsys.2005.07.002

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

Vander Beken, H., & Brysbaert, M. (2017). Studying texts in a second language: the importance of test type. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728917000189

APPENDIX 2A

First Fixation Duration Analysis

First fixations more than 2.5 SD away from the subject means were excluded (2.34%). The final model is presented in Table S1.

The main effect of AoA was significant. First fixations were shorter for earlier learned nouns. There was a main effect of word frequency: shorter fixations for higher frequency words. The main effect of word length was marginally significant; this predictor interacted with language proficiency. Post hoc contrasts showed that fixations became longer with an increasing word length, but this effect diminished for participants scoring 90.65 or higher on the LexTALE (χ = 3.87, df = 1 p < .05).

Table A1. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for first fixation duration.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.3151	0.0145	159.769	<.001	***
Age of Acquisition	0.0018	0.0004	4.092	<.001	***
Word Frequency	-0.0077	0.0016	-4.718	<.001	***
Word Length	0.0012	0.0006	1.882	.070	.
Neighborhood Density	-0.0001	0.0002	-0.493	.622	
Language Proficiency	-0.0009	0.0015	-0.572	.578	
Rank of Occurrence	<0.0001	<0.0001	0.480	.613	
Word Length * Language Proficiency	-0.0002	0.0001	-3.046	.009	**

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0003	0.017
Subject		
(Intercept)	0.003	0.054
Age of Acquisition	<0.0001	0.001
Word Frequency	<0.0001	0.004
Word Length	<0.0001	0.002

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

APPENDIX 3A

L2 AoA rating instructions

"Welcome to this experiment. You will have to rate words on the age at which you've learned them. By this we mean the age at which you completely understood the word when someone used it, even if you didn't use it yourself. Please only use round numbers in the list (you'll get an error message if you don't). Some examples:

If you think you learned the word "love" at age 8, fill in 8.

If you think you learned the word "neuroscientist" at age 18, fill in 18.

You have to fill in the number in the column "age learned". If you don't know the word, put an 'x' in the column "word unknown". When your input is correct (either a number or an x), the column next to your input will turn from red to green. Please make sure to fill in every row. Try to use the whole age range from when you started to learn English words up to now. Try to make an estimation as good as possible, but don't think too long about a word."

APPENDIX 3B

As an additional experiment, we ran a lexical decision task with some of the nouns included in GECO (Cop, Dirix, Drieghe, & Duyck, 2017). This allowed us to (a) compare the results of this task with those of the eye movement measures of Experiment 2 and (b) compare our lexical decision results with those of Izura and Ellis (2002).

Method

Participants and Materials. We split up the data collection of the L1 and L2 reaction times (RTs): we took the L1 lexical decision data from the Dutch Lexicon Project 2 (DLP2; Brysbaert, Stevens, Mandera & Keuleers, 2016). We found 921 Dutch GECO words in the DLP data. For L2 reading (and more specifically, L2 reading by our sort of bilinguals), such a database is not available. So, for the L2 data collection, we randomly selected 800 English words of the 966 stimuli included in the current study (see Table S1 for the characteristics of the Dutch and English words). These were divided in four lists and paired with non-words using the "Wuggy" tool (Keuleers & Brysbaert, 2010). These were embedded in four L2 lexical decision tasks, for which we recruited 20 participants each. A total of 80 Dutch-English bilingual first year psychology students of Ghent University took part in this study for course credit (64 females, Mage = 18.4, SDage = 1.1). Their L1 (Dutch) and L2 (English) proficiency was rated with the Dutch and English version of the LexTALE (Lemhöfer & Broersma, 2012). For Dutch, their average LexTALE score was 81.53 (SD = 6.25, range = [62.5 – 90]). For English, their average LexTALE score was 71.45 (SD = 9.01, range = [50 – 90]). We refer to Brysbaert et al. (2016) for information on the participants of the L1 data collection.

Table B1. *Descriptive Statistics for the GECO nouns analyzed in the current experiment, averaged over stimuli per language (standard deviations between parentheses).*

	Word Frequency[a]	Word Length	L1 AoA[b]	L2 AoA[c]	CLD[d]
Dutch	1.13 (0.83)	7.08 (2.45)	7.84 (2.10)	13.11 (1.56)	0.32 (0.26)
English	1.64 (0.63)	5.46 (1.47)	6.85 (1.84)	12.47 (1.43)	0.38 (0.31)

[a]Log10 Subtlex per million frequencies: SUBTLEX-NL for Dutch words (Keuleers, Brysbaert & New, 2010), SUBTLEX-UK for English words (van Heuven, Mandera, Keuleers & Brysbaert, 2013); [b]For "Dutch", this means the L1 AoA of the Dutch words, for "English" the L1 AoA of the Dutch translation of the English words (Brysbaert, Stevens, De Deyne, Voorspoels & Storms, 2014); [c]For "Dutch", this means the L2 AoA of the English translation of the Dutch words, for English the L2 AoA of the English words (from the Experiment 1 ratings, see Supplementary Material).[d]The average amount of orthographic overlap, expressed by the Corrected Levenshtein Distance.

Procedure. Participants were seated approximately 45-60cm from the screen. All instructions were presented on the screen. They were instructed to judge whether a presented letter string was either an English word or a non-word by pressing the appropriate button on a response box (response mapping was counterbalanced across participants). They were asked to decide as quickly and accurately as possible. Prior to the actual experiment, participants had to complete a practice block with 10 trials, which was repeated if their accuracy was below 80% on the first try. After successful completion of the practice block, two experimental blocks were presented, with a presentation of the stimuli in a pseudo-random order (no more than four consecutive words or nonwords were presented). Between the two experimental blocks, participants could take a short break.

A trial started with the presentation of a fixation cross (800ms), followed by a blank screen of 300ms. The stimulus was presented until the participant responded or for a maximum duration of 2500ms. The inter-trial interval was kept constant at 700ms.

After finishing the experiment, participants had to complete the English and Dutch version of the LexTALE. We refer to Brysbaert et al. (2016) for the procedure of the DLP2 lexical decision task.

Results

The transformation of (in)dependent variables and the analysis for the reaction times (RTs) was performed in the same way as the reading measures in Experiment 2. Only correct trials were analyzed (1.9% data loss) and RTs that differed more than 2.5 standard deviations from the subject means were considered as outliers and excluded (0.85%). The final model is presented in Table S2. The maximum VIF was 3.94.

For the L1 lexical decision task, there was a facilitatory influence of L1 AoA (faster RTs for earlier learned words), but no effect of L2 AoA. For the L2 lexical decision task, the pattern was opposite: L2 AoA had a facilitatory effect on the L2 RTs, but not L1 AoA.

Effects of L1 AoA. The main effect of L1 AoA was significant: RTs were shorter for earlier learned words ($\beta = 0.0020$, se = 0.0007, $t = 3.082$, $p < .01$). There were also significant two-way interactions with language ($\beta = -0.0059$, se = 0.0013, $t = -4.461$, $p < .001$), word frequency ($\beta = -0.0017$, se = 0.0006, $t = -2.849$, $p < .01$), and L1 proficiency ($\beta = -0.0002$, se < 0.0001, $t = -3.480$, $p < .001$). The L1 AoA effect was significant for words with a log frequency up to 1.828 ($\chi^2 = 3.89$, df = 1, $p < 0.05$) and for L1 proficiency scores smaller than 97.80 ($\chi^2 = 3.89$, df = 1, $p < 0.05$). The three-way interaction between L1 AoA, language and word frequency was also significant ($\beta = 0.0038$, se = 0.0012, $t = 3.156$, $p < .01$).

Effects of L2 AoA. The main effect of L2 AoA was not significant. There was however a significant two-way interaction with language ($\beta = 0.0176$, se = 0.0014, $t = 12.171$, $p < .001$) and a three-way interaction with language and word frequency ($\beta = -0.0044$, se = 0.0015, $t = -3.034$, $p < .01$).

L1 Lexical Decision. ***Effects of L1 AoA.*** In the analysis for the L1 lexical decision task, there was a main effect of L1 AoA ($\beta = 0.0023$, se = 0.0005, $t = 4.113$, $p < .001$): RTs were faster for earlier learned words. The interactions with word frequency ($\beta = -0.0016$, se = 0.0006, $t = -2.748$, $p < .01$) and L1 proficiency ($\beta = -0.0002$, se < 0.0001, $t = -3.134$, $p < .01$) were significant. Post-hoc contrasts revealed that the L1 AoA effect was significant for words with a word frequency op to 1.843 ($\chi^2 = 3.88$, df = 1, $p < 0.05$) and L1 proficiency scores up to 98.45 ($\chi^2 = 3.86$, df = 1, $p < 0.05$).

Effects of L2 AoA. There were no significant L2 AoA effects in the L1 lexical decision task.

L2 Lexical Decision. ***Effects of L2 AoA.*** The main effect of L2 AoA was significant ($\beta = 0.0144$, se = 0.0011, $t = 13.380$, $p < .001$), with faster RTs for earlier learned words, as well as its interaction with word frequency (β = -0.0049, se = 0.0013, t = -3.726, p < .001). Contrasts showed that this two-way interaction was significant for words with a word frequency up to 3.213 (χ^2 = 3.89, df = 1, p < 0.05).

Effects of L1 AoA. There were no significant effects of L1 AoA on the L2 lexical decision task.

Table B2. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final linear mixed effect model for lexical decision reaction times.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.7082	0.0075	363.039	<.001	***
L1 Age of Acquisition	0.0024	0.0006	3.918	<.001	***
L2 Age of Acquisition	-0.0001	0.0007	-0.116	.907	
Language	0.0687	0.0145	4.726	<.001	***
Word Frequency	-0.0161	0.0015	-10.744	<.001	***
Word Length	0.0001	0.0004	0.237	.813	
L1 Proficiency	-0.0010	0.0009	-1.066	.288	
L2 Proficiency	-0.0001	0.0005	0.115	.909	
Orthographic Overlap	-0.0012	0.0002	-0.526	.599	
L1 AoA * Language	-0.0053	0.0011	-4.792	<.001	***
L1 AoA * Word Frequency	-0.0016	0.0006	-2.739	.006	**
L1 AoA * L1 Proficiency	-0.0002	<0.0001	-3.484	<.001	***
L2 AoA * Language	0.0151	0.0012	12.613	<.001	***
L2 AoA * Word Frequency	-0.0001	0.0001	-0.999	.318	
Language * Word Frequency	-0.0056	0.0028	-2.026	.043	*
Language * Word Length	-0.0025	0.0008	-3.029	.002	**
Word Frequency * Word Length	-0.0016	0.0005	-3.539	<.001	***
L1 AoA * Language * Word Frequency	0.0037	0.0012	3.094	.002	**
L2 AoA * Language * Word Frequency	-0.0044	0.0015	-3.036	.002	**
Language * Word Frequency * Word Length	0.0025	0.0012	2.149	.032	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0003	0.0170
Subject		
(Intercept)	0.0037	0.0605
L1 AoA	<0.0001	0.0018
L2 AoA	<0.0001	0.0020
Word Frequency	0.0001	0.0080

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Discussion

We obtained no effects of L1 AoA on L2 reading, or vice versa. The absence of such crosslingual effects in isolated word reading is consistent with Izura & Ellis (2002) and (again) illustrates possible task differences between lexical decision and eye movement data. Also, this observation is inconsistent with the semantic hypothesis, which would predict an L1 AoA effect on L2 word recognition through the shared semantic representations of the words.

We did however find within-language AoA effects for reading in both L1 and L2, just like in the eye tracking data. These AoA effects point towards a general order of acquisition effect which is predicted by the mapping hypothesis. Note however that these lexical decision results do not necessarily exclude the semantic hypothesis, in favor of this mapping hypothesis. Kuperman, Drieghe, Keuleers, and Brysbaert (2013) correlated lexical decision RTs and various timed eye movement measures. These correlations were in general quite low, but the highest ones were found between RTs and the total reading time. We could therefore presume that these two measures reflect the same underlying processes, to some extent. In accordance with the current lexical decision data, we also didn't find a cross-lingual AoA effect in the late measures in our eye tracking analysis. This illustrates the value of the multiple measures that eye tracking allows, representing different stages in the time course of word recognition. In the GECO analyses, we did obtain an effect of L1 AoA on L2 reading in the early measures, which supports the semantic hypothesis, whereas such an effect is not observed in the lexical decision task, nor in the late eye tracking measures.

References

Brysbaert M., Stevens M., De Deyne S., Voorspoels W. and Storms G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words, *Acta Psychologica, 150*, 80–84. doi: http://dx.doi.org/10.1016/j.actpsy.2014.04.010.

Brysbaert M., Stevens M., Mandera P. and 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. doi: http://dx.doi.org/10.1037/xhp0000159.

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. doi: http://dx.doi.org/10.3758/s13428-016-0734-0.

Izura C. and Ellis A.W. (2002). Age of acquisition effects in word recognition and production in first and second languages, *Psicológica, 23*, 245–281.

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

Keuleers E., Brysbaert M. and New B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles, *Behavior Research Methods, 42*(3), 643–650. doi: http://dx.doi.org/10.3758/BRM.42.3.643.

Kuperman V., Drieghe D., Keuleers E. and 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. doi: http://doi.org//10.1080/17470218.2012.658820.

Lemhöfer K. and Broersma M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English, *Behavior Research Methods*, *44*(2), 325–343. doi: http://dx.doi.org/10.3758/s13428-011-0146-0

van Heuven W.J.B., Mandera P., Keuleers E. and Brysbaert M. (2013). SUBTLEX-UK: A new and improved word frequency database for British English, *Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. doi: http://dx.doi.org/10.1080/17470218.2013.850521.

APPENDIX 4A

Formula for the used measure of Orthographic Overlap: the Corrected Levenshtein Distance (taken from Schepens, Dijkstra, & Grootjen, 2012).

$$Orthographic\ Overlap = 1 - \frac{Distance}{Length}$$

Distance = min (number of insertions, deletions and substitutions needed to edit target word into translation word)

Length = max (length of target word, length of translation word)

APPENDIX 4B

Proficiency scores.

Table B1. *Average percentage scores (standard deviations between brackets and range between square brackets) on the LexTALE. Average rating on the self-report questionnaire (standard deviations between brackets).*

		Dutch	English	t-value L1-L2
LexTALE-score (%)		87.58 (7.03) [70.00-96.25]	73.04 (9.08) [57.50-88.75]	6.519***
Self Report				
	Listening	4.9 (0.4)	4 (0.58)	5.141***
	Speaking	4.87 (0.34)	3.5 (0.612	7.628***
	Reading	4.9 (0.3)	3.93 (0.63)	5.604***
	Writing	4.8 (0.48)	3.43 (0.72)	6.899***
	Average	4.87 (0.29)	3.72 (0.47)	7.523***

* p<0.05, **p<0.01, ***p<0.001

Participants had on average a higher proficiency for Dutch then English, both on the LexTALE, $t(29) = 7.518$, $p < .001$, and the average self-proficiency ratings, $t(29) = 10.891$, $p < .001$.

Table B2. *Average percentage scores (standard deviations between brackets and range between square brackets) on the LexTALE, Spelling test and Lexical Decision task for the bilingual and monolingual group in Experiment 2.*

	Monolinguals	Bilinguals L1	Bilinguals L2	t-value L1-L2	t-value L1-mono
LexTALE-score (%)	91.07(8.92) [71.25-100]	92.43 (6.34) [73.75-100]	75.63(12.87) [51.25-98.75]	7.59 ***	0.49
Spelling score (%)	80.78 (7.26) [73.81-90.48]	83.16(7.80) [67.00-93.00]	69.92 (8.74) [52.00-83.00]	8.15 ***	0.99
Lexical Decision score (%)	77.89 (12.01) [54.61-95.23]	80.47 (5.45) [68.87-88.76]	56.75 (11.01) [38.46-75.86]	9.87 ***	0.67

* p<0.05, **p<0.01, ***p<0.001

The Dutch (L1) proficiency of the bilinguals was matched with the English proficiency of the monolinguals indicating that both groups were equally proficient in their first language. Neither the LexTALE (t=0.488, df=22.254, p=0.630), the spelling test (t=0.989, df=29.282, p=0.331), nor the lexical decision tasks (t=0.667, df=17.092, p=0.514) yielded significant differences for these two groups performing in L1. The bilingual L2 LexTALE scores were significantly lower than their L1 scores (t=7.587, df=18, p<0.001). The bilingual L2 Spelling scores were lower than the L1 scores (t=8.154, df=18, p<0.001). The performance of the bilinguals on the classic lexical decision task was significantly better in L1 (t=9.873, df=18, p<0.001) than in L2.

APPENDIX 4C

Stimulus materials used in Experiment 1.

Dutch Words

Large Dutch N, Large English N. Bons, borg, bril, dolk, hiel, klam, knie, oord, plek, rund, sein, spar, takt, tolk, vork, wolk, worp, woud, wrak, zalf

Large Dutch N, Small English N. Berg, beul, bouw, deun, dief, eter, fuik, kelk, kies, knal, kous, rede, snik, teug, touw, twee, unie, vals, verf, vies

Small Dutch N, Large English N. Brug, bult, draf, drie, fris, galg, hemd, heup, lach, meid, melk, munt, nota, pret, prik, smid, stug, vete, welp, wilg

Small Dutch N, Small English N. Akte, ambt, blad, erwt, ezel, gesp, gids, gips, inkt, joch, muts, ober, pech, pion, rots, snor, stro, toga, trui, veld

English Words

Large Dutch N, Large English N. Aunt, blue, farm, hawk, knit, left, loan, loud, maid, monk, moon, path, quit, shoe, suit, tool, verb, weak, wrap, zero

Large Dutch N, Small English N. Army, atom, bias, bird, diet, edge, germ, huge, butt, jerk, keen, knee, liar, lion, myth, noon, nude, obey, poem, poor

Small Dutch N, Large English N. Bath, bomb, busy, clue, coin, desk, dial, dirt, dish, firm, grey, hurt, iron, joke, lamb, limb, loss, milk, prey, rude

Small Dutch N, Small English N. Deny, duty, earl, envy, evil, folk, frog, guts, idol, kiss, okay, oral, oval, soup, true, twin, ugly, used, vein, view

Nonwords

Large Dutch N, Large English N. Aril, aunk, blag, boul, boup, braf, bret, dris, duef, elap, fram, frip, furk, gonk, heud, jeef, knat, knub, koup, loem, meem, merd, mots, oram, peit, pern, piot, pral, pred, rama, sluf, sluk, snus, sols, stui, tess, trum, tult, vene, zork

Large Dutch N, Small English N. Alof, besp, bito, bouf, daus, drot, epoe, etel, feik, goep, grul, heut, irok, jees, jeul, jund, jurf, kalp, kelf, kerd, keun, loga, morp, muig, mups, nazz, noge, nont, noto, obel, oune, pris, puif, reug, reun, slen, smir, viem, woup, zuls

Small Dutch N, Large English N. Aute, bele, bulf, ceot, chah, cham, clet, dolo, drid, dulp, feul, foug, fran, genk, girs, jant, jero, jert, liry, lurd, lurp, lusp, naul, nirk, nudo, orim, pani, prad, prog, puet, raut, reud, rion, ruze, seto, snam, tirk, tran, vich, vorn

Small Dutch N, Small English N. Aler, anas, arns, aurd, baun, cafa, chof, deim, dilm, drio, durs, enip, fenk, feup, frig, frus, giep, heif, hilp, jalp, jofe, kach, kiot, knaf, luet, maup, moug, nige, omil, paby, ridi, siom, taur, torp, tuni, twol, unar, vota, zous, zuke

APPENDIX 4D

Results of the linear mixed effects analysis of the generalized lexical decision data (Experiment 1)

Table D1. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for reaction times for Dutch words.*

Dutch Words					
	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.806	0.0156	179.90	<.001	***
Dutch N density	-0.00006	0.0014	-0.04	.969	
English N density	0.00001	0.0012	0.01	.991	

Dutch N Frequency	-0.0019	0.0119	-0.16	.876	
English N Frequency	-0.0003	0.0125	-0.02	.981	
Word frequency	-0.0391	0.0069	-5.66	<.001	***
Average bigram frequency	0.0163	0.0176	0.93	.357	
Orthographic Overlap	-0.0073	0.0167	-0.44	.664	
	Variance	SD			
Random Effects					
Word					
(Intercept)	0.001	0.032			
Subject					
(Intercept)	0.003	0.051			
p<0.1 . p<0.05 * p<0.01 ** p<0.001***					

Table D2. *Estimates, standard errors, z-values and p-values for the fixed and random effects of the final general linear mixed effect model for error rate for Dutch words.*

Dutch Words					
	Estimate	SE	z-value	p-value	
Fixed Effects					
Intercept	-3.30	0.47	-7	< 0.001	***
Dutch N density	0.05	0.05	1.19	0.233	
English N density	-0.01	0.04	-0.19	0.849	
Dutch N Frequency	-0.22	0.39	-0.58	0.56	
English N Frequency	0.06	0.39	-0.58	0.88	
Word frequency	-1.22	0.24	-5.14	< 0.001	***
Average bigram frequency	-0.5	0.55	-0.93	0.355	
Orthographic Overlap	-0.15	0.54	-0.28	0.781	
English N density * Word frequency	0.13	0.07	1.93	0.053	.
	Variance	SD			
Random effects					
Word					
(Intercept)	0.443	0.666			
Subject					
(Intercept)	0.389	0.624			
p<0.1 . p<0.05 * p<0.01 ** p<0.001***					

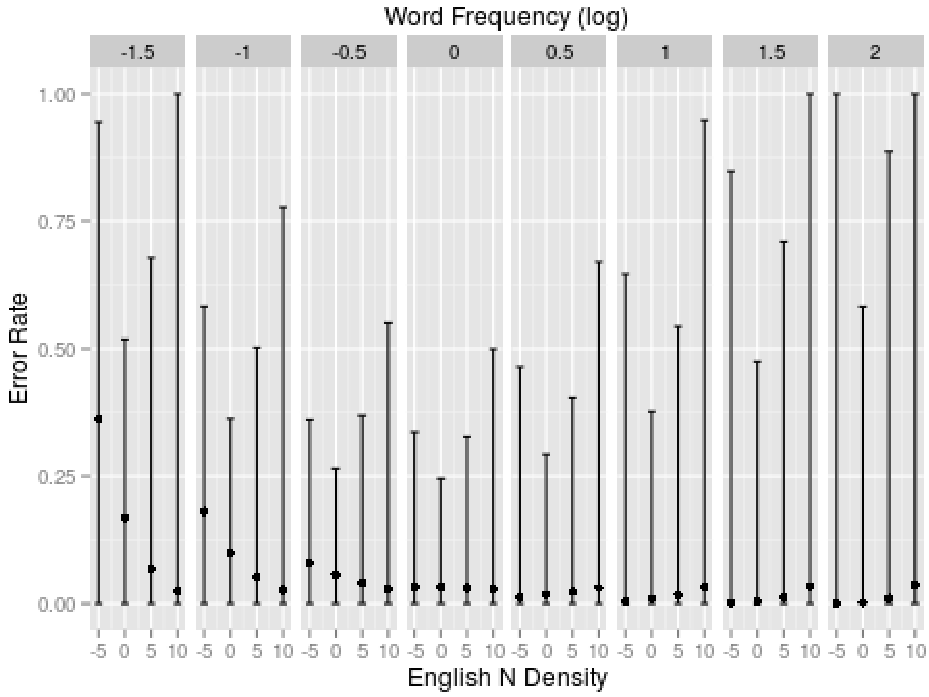

Figure D1. *Error rate (on the y-axis) for Dutch words dependent on English N density (on the x-axis) dependent on word frequency of the word (panels) for a generalized lexical decision task.*

Table D3. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for reaction times for English words.*

English Words					
	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.807	0.0128	218.66	<.001	***
Dutch N density	0.0013	0.0011	1.15	.254	
English N density	0.00007	0.0011	0.07	.946	
Dutch N Frequency	0.0087	0.0079	1.10	.277	
English N Frequency	0.0081	0.087	0.94	.353	
Word frequency	-0.0076	0.0129	-0.59	.556	
Average bigram frequency	0.0250	0.0258	0.97	.337	
Orthographic Overlap	-0.0117	0.0124	-0.95	.349	
Dutch N density * Average bigram frequency	-0.0134	0.0066	-2.04	.046	*
English N density * Word Frequency	0.0044	0.0025	1.75	.086	.

English N Frequency * Word Frequency	-0.0402	0.016	-2.51	.015	*
English N Frequency * Average bigram frequency	-0.0564	0.0315	-1.79	.078	.

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0005	0.021
Subject		
(Intercept)	0.0027	0.052
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table D4. *Estimates, standard errors, z-values and p-values for the fixed and random effects of the final general linear mixed effect model for error rates for English words.*

English Words					
	Estimate	SE	z-value	p-value	
Fixed Effects					
Intercept	-3.50	0.40	-8.71	< .001	***
Dutch N density	0.10	0.04	2.32	.021	*
English N density	-0.07	0.05	-1.35	.177	
Dutch N Frequency	-0.06	0.33	-0.18	.857	
English N Frequency	0.228	0.40	0.57	.566	
Word frequency	-0.48	0.43	-1.10	.270	
Average bigram frequency	-1.36	0.67	-2.01	.044	*
Orthographic Overlap	0.05	0.50	0.11	.914	
English N density * Average bigram frequency	-0.29	0.17	-1.65	.099	.
English N Frequency * word frequency	-1.01	0.55	-1.84	.065	.

	Variance	SD
Random Effects		
Word		
(Intercept)	0.542	0.737
Subject		
(Intercept	0.487	0.698
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

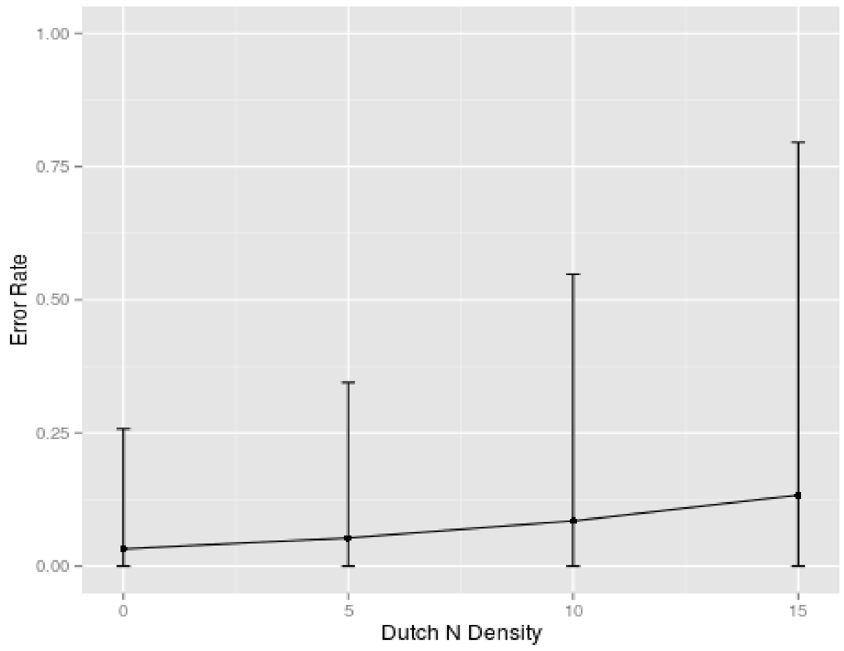

Figure D2. *Error rates (on the y-axis) for English words dependent on Dutch N density (on the x-axis) in a generalized lexical decision task.*

APPENDIX 4E

Results of the linear mixed effects analysis of the natural reading data (Experiment 2).

Table E1. *Estimates, standard errors, z-values and p-values for the fixed and random effects of the final general linear mixed effect model for Skipping Rates for bilingual L1 reading.*

Bilingual L1					
	Estimate	SE	z-value	p-value	
Fixed Effects					
(Intercept)	-0.903	0.111	-8.130	<.001	***
Dutch N Density	0.003	0.007	0.446	.655	
English N Density	0.0005	0.004	0.111	.9111	
Dutch N Frequency	-0.011	0.037	-0.289	.773	
English N Frequency	0.050	0.043	1.141	.254	
Word Frequency	0.099	0.021	4.736	<.001	***
Word Length	-0.227	0.013	-17.078	<.001	***
Average Bigram Frequency	-0.031	0.067	-0.456	.648	
Orthographic Overlap	-0.028	0.051	-0.548	.583	
Rank of Occurrence	0.0005	0.0005	0.911	.362	
Dutch N Density * Word Frequency	-0.011	0.003	-3.266	.001	**

Dutch N Density * Word Length	-0.007	0.002	-2.918	.004	**
English N Frequency * Word Frequency	0.078	0.043	1.790	.074	.

	Variance	SD
Random Effects		
Word		
(Intercept)	0.039	0.198
Subject		
(Intercept)	0.205	0.453
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E2. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for Single Fixation Durations for bilingual L1 reading.*

Bilingual L1					
	Estimate	SE	t-value	p-value	
Fixed Effects					
(Intercept)	2.306	0.010	222.955	<.001	***
Dutch N Density	-0.0002	0.0003	-0.657	.511	
English N Density	0.0001	0.0003	0.437	.662	
Dutch N Frequency	-0.001	0.003	-0.396	.693	
English N Frequency	0.005	0.004	1.424	.155	
Word Frequency	-0.010	0.001	-7.262	<.001	***
Word Length	0.004	0.001	6.354	<.001	***
Average Bigram Frequency	0.001	0.005	0.243	.808	
Orthographic Overlap	-0.002	0.004	-0.550	.583	
Rank of Occurrence	<-0.0001	<0.0001	-1.056	.291	
Dutch N Density * Word Frequency	0.001	0.0002	3.595	<.001	***

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0003	0.019
Subject		
(Intercept)	0.002	0.043
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E3. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for Gaze Durations for bilingual L1 reading.*

Bilingual L1					
	Estimate	SE	t-value	p-value	
Fixed Effects					
(Intercept)	2.334	0.013	182.702	<.001	***
Dutch N Density	-0.001	0.0004	-1.460	.145	
English N Density	0.0001	0.0004	0.157	.875	
Dutch N Frequency	0.001	0.003	0.285	.776	

English N Frequency	0.004	0.004	0.901	.368	
Word Frequency	-0.016	0.002	-8.547	<.001	***
Word Length	0.008	0.001	11.919	<.001	***
Average Bigram Frequency	-0.001	0.006	-0.129	.897	
Orthographic Overlap	-0.005	0.004	-1.027	.305	
Rank of Occurrence	<-0.0001	<0.0001	-0.662	.508	
Dutch N Density * Word Frequency	0.001	0.0003	3.662	<.001	***
Dutch N Frequency * Word Frequency	0.006	0.003	2.017	.044	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0006	0.025
Subject		
(Intercept)	0.0029	0.054
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E4. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for Total Reading Times for bilingual L1 reading.*

Bilingual L1					
	Estimate	SE	t-value	p-value	
Fixed Effects					
(Intercept)	2.381	0.014	175.051	<.001	***
Dutch N Density	-0.0005	0.0005	-0.950	.342	
English N Density	0.0001	0.001	0.203	.839	
Dutch N Frequency	0.004	0.004	0.888	.375	
English N Frequency	-0.002	0.005	-0.415	.678	
Word Frequency	-0.022	0.002	-9.553	<.001	***
Word Length	0.010	0.001	12.979	<.001	***
Average Bigram Frequency	0.003	0.010	0.350	.727	
Orthographic Overlap	0.0004	0.005	0.067	.947	
Rank of Occurrence	<-0.0001	<0.0001	-1.369	.171	
Dutch N Density * Word Frequency	0.001	0.0004	3.281	.001	**
Dutch N Frequency * Word Frequency	0.009	0.004	2.394	.017	*
Dutch N Frequency * Average Bigram Frequency	0.027	0.015	1.859	.063	.
English N Frequency * Average Bigram Frequency	-0.030	0.017	-1.754	.080	.

	Variance	SD
Random Effects		
Word		
(Intercept)	0.001	0.032
Subject		
(Intercept)	0.003	0.057
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E5. *Estimates, standard errors, z-values and p-values for the fixed and random effects of the final general linear mixed effect model for Regressions for bilingual L1 reading.*

Bilingual L1					
	Estimate	SE	z-value	p-value	
Fixed Effects					
(Intercept)	-2.143	0.098	-21.859	<.001	***
Dutch N Density	0.017	0.008	2.155	.031	*
English N Density	0.008	0.008	1.000	.317	
Dutch N Frequency	-0.023	0.068	-0.333	.739	
English N Frequency	-0.169	0.087	-1.951	.051	.
Word Frequency	-0.060	0.031	-1.919	.055	.
Word Length	-0.054	0.013	-3.992	<.001	***
Average Bigram Frequency	0.163	0.126	-1.299	.194	
Orthographic Overlap	0.057	0.095	0.601	.548	
Rank of Occurrence	0.0004	0.0008	0.469	.639	

	Variance	SD
Random Effects		
Word		
(Intercept)	.262	.512
Subject		
(Intercept)	.146	.382
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E6. *Estimates, standard errors, z-values and p-values for the fixed and random effects of the final general linear mixed effect model for Skipping Rates for bilingual L2 reading.*

Bilingual L2					
	Estimate	SE	z-value	p-value	
Fixed Effects					
(Intercept)	-1.074	0.126	-8.527	<.001	***
Dutch N Density	0.006	0.005	1.409	0.159	
English N Density	0.009	0.003	2.730	0.006	**
Dutch N Frequency	0.008	0.034	0.228	0.820	
English N Frequency	0.038	0.030	1.269	0.205	
Word Frequency	0.139	0.178	7.813	<.001	***
Word Length	-0.190	0.010	-18.677	<.001	***
Average Bigram Frequency	0.039	0.100	0.387	0.698	
Orthographic Overlap	0.120	0.045	2.676	0.007	**
Rank of Occurrence	0.002	0.001	3.164	0.002	**
English N Density * Average Bigram Frequency	0.018	0.009	1.986	0.047	*
Dutch N Frequency * Average Bigram Frequency	0.256	0.127	2.022	0.043	*
English N Frequency * Average Bigram Frequency	-0.300	0.134	-2.239	0.025	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.029	0.171
Subject		
(Intercept)	0.291	0.540
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E7. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for Single Fixation Durations for bilingual L2 reading.*

Bilingual L2					
	Estimate	SE	t-value	p-value	
Fixed Effects					
(Intercept)	2.336	0.011	217.144	<.001	***
Dutch N Density	-0.002	0.001	-2.508	.013	*
English N Density	<0.0001	0.0003	0.140	.888	
Dutch N Frequency	0.002	0.003	0.689	.491	
English N Frequency	0.002	0.003	0.842	.400	
Word Frequency	-0.016	0.001	-11.286	<.001	***
Word Length	0.002	0.001	1.520	.129	
Average Bigram Frequency	0.012	0.005	2.440	.015	*
Orthographic Overlap	-0.004	0.004	-1.166	.244	
Rank of Occurrence	-0.0001	0.0001	-1.819	.069	.
Dutch N Density * Word Length	-0.001	0.0004	-2.736	.006	**

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0003	0.018
Subject		
(Intercept)	0.002	0.045
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E8. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for Gaze Durations for bilingual L2 reading.*

Bilingual L2					
	Estimate	SE	t-value	p-value	
Fixed Effects					
(Intercept)	2.375	0.014	169.101	<.001	***
Dutch N Density	-0.002	0.001	-1.871	.062	.
English N Density	0.0002	0.0004	0.406	.685	
Dutch N Frequency	0.002	0.004	0.503	.615	
English N Frequency	0.001	0.003	0.352	.725	
Word Frequency	-0.018	0.002	-10.867	<.001	***
Word Length	0.008	0.001	5.487	<.001	***

Average Bigram Frequency	0.017	0.006	2.826	.005	**
Orthographic Overlap	-0.003	0.004	-0.742	.458	
Rank of Occurrence	-0.0001	0.0001	-1.019	.308	
Dutch N Density * Word Length	-0.001	0.0005	-2.174	.030	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0009	0.030
Subject		
(Intercept)	0.004	0.059
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E9. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the final general linear mixed effect model for Total Reading Times for bilingual L2 reading.*

Bilingual L2					
	Estimate	SE	t-value	p-value	
Fixed Effects					
(Intercept)	2.376	0.029	83.308	<.001	***
Dutch N Density	-0.003	0.002	-2.066	.039	*
English N Density	0.0002	0.0005	0.514	.607	
Dutch N Frequency	0.001	0.005	0.187	.852	
English N Frequency	0.001	0.004	0.391	.696	
Word Frequency	-0.028	0.002	-14.460	<.001	***
Word Length	0.011	0.002	6.041	<.001	***
Average Bigram Frequency	0.020	0.007	2.722	.007	**
Orthographic Overlap	-0.006	0.005	-1.160	.246	
Rank of Occurrence	-0.0002	0.0001	-1.865	.062	.
Dutch N Density * Word Length	-0.001	0.001	-1.984	.048	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.001	0.037
Subject		
(Intercept)	0.004	0.064
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table E10. *Estimates, standard errors, z-values and p-values for the fixed and random effects of the final general linear mixed effect model for Regressions for bilingual L2 reading.*

Bilingual L2					
	Estimate	SE	z-value	p-value	
Fixed Effects					
(Intercept)	-2.093	0.113	-18.530	<.001	***
Dutch N Density	0.010	0.009	1.164	.244	
English N Density	0.009	0.006	1.390	.165	

Dutch N Frequency	0.026	0.063	0.415	.678	
English N Frequency	-0.004	0.052	-0.078	.937	
Word Frequency	-0.096	0.028	-3.447	<.001	***
Word Length	-0.066	0.015	-4.361	<.001	***
Average Bigram Frequency	0.087	0.104	0.843	.399	
Orthographic Overlap	-0.101	0.077	-1.312	.189	
Rank of Occurrence	-0.003	0.001	-2.002	.045	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.203	0.451
Subject		
(Intercept)	0.211	0.459
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

APPENDIX 4F

Analyses of nonwords in experiment 1

LMER analyses

For the nonwords, only N densities were available as word characteristics of the stimuli. Therefore, only Dutch and English N density were included as fixed factors in the analyses.

Results for the analysis of RTs are presented in Table F1. There was an inhibitory effect of Dutch N density.

Table F1. *Estimates, standard errors, t-values and p values for the fixed and random effects of the final general linear mixed effect model for reaction times for nonwords.*

	Estimate	SE	t-value	*p*	
Fixed Effects					
Intercept	2.857	0.0106	269.31	<.001	***
Dutch N density	0.0043	0.0008	5.15	<.001	***
English N density	0.0011	0.0007	1.47	.223	

	Variance	SD
Random Effects		
Stimulus		
(Intercept)	0.0004	0.021
Subject		
(Intercept)	0.0032	0.057

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Results for the analysis of error rates are presented in Table F2. There was again an inhibitory effect of Dutch N density.

Table F2. *Estimates, standard errors, z-values and p values for the fixed and random effects of the final general linear mixed effect model for accuracy for nonwords.*

	Estimate	SE	z-value	*p*	
Fixed Effects					
Intercept	-3.77	0.25	-15	<.001	***
Dutch N density	0.10	0.04	2.33	0.020	*
English N density	0.02	0.04	0.50	.615	

	Variance	SD
Random Effects		
Stimulus		
(Intercept)	0.874	0.935
Subject		
(Intercept)	1.179	1.086

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

For nonwords there seemed to be no effect of L2, but a robust inhibitory effect of L1 neighborhood density.

APPENDIX 4G

Results for the English monolinguals in Experiment 2

Early measures.

We fitted a logistic linear mixed model for skipping probability. The outcome of the final model for skipping rates is presented in Table G3. For the single fixation analyses, only the nouns that received one fixation were selected (56.9%). Single fixation durations that differed more than 2.5 standard deviations from the subject means were excluded (1.98%). The outcome of the final model for single fixation durations is presented in Table G4.

Table G3. *Estimates, standard errors, z-values and p values for the fixed and random effects of the final general linear mixed effect model for Skipping Rates for monolingual reading.*

Monolingual					
	Estimate	SE	z-value	*p*	
Fixed Effects					
(Intercept)	-0.683	0.957	-7.136	<.001	***
Dutch N Density	0.002	0.004	0.465	.642	
English N Density	0.014	0.003	4.579	<.001	***

Dutch N Frequency	-0.023	0.029	-0.791	.429	
English N Frequency	0.021	0.025	0.846	.398	
Word Frequency	0.106	0.015	7.161	<.001	***
Word Length	-0.179	0.009	-20.200	<.001	***
Average Bigram Frequency	-0.147	0.050	-2.952	.003	**
Rank of Occurrence	0.002	0.001	2.931	.003	**
English N Density * Word Frequency	-0.004	0.002	-1.937	.053	.
English N Frequency * Word Length	-0.027	0.0123	-2.100	.036	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.030	0.174
Subject		
(Intercept)	0.122	0.349
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table G4. *Estimates, standard errors, t-values and p values for the fixed and random effects of the final general linear mixed effect model for Single Fixation Durations for monolingual reading.*

Monolingual					
	Estimate	SE	t-value	*p*	
Fixed Effects					
(Intercept)	2.315	0.015	157.955	<.001	***
Dutch N Density	0.0002	0.0004	0.719	.473	
English N Density	-0.001	0.0004	-2.780	.006	**
Dutch N Frequency	-0.002	0.003	-0.840	.401	
English N Frequency	0.003	0.002	1.276	.202	
Word Frequency	-0.011	0.001	-9.715	<.001	***
Word Length	0.001	0.001	0.475	.635	
Average Bigram Frequency	0.012	0.004	2.980	.003	**
Rank of Occurrence	<0.0001	<0.0001	0.029	.977	
English N Density * Word Length	-0.0004	0.0002	-2.463	.014	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0004	0.019
Subject		
(Intercept)	0.003	0.054
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

We found a main effect of within-language neighborhood density for skipping rates: words were more likely to be skipped with an increasing N density. English N frequency interacted with word length: the effect of N frequency was facilitatory for very short words (4 letters or less, χ^2=3.99, df=1, p < .05).

Fixations were shorter fixations for words with more within-language neighbors. This variable interacted significantly with word length. Post hoc contrasts showed that the facilitation of neighborhood density was present only for words of 5 letters or more (χ^2=6.56, df=1, $p < .05$).

Cross-language N frequency had no effect on the skipping rates or single fixation times.

Intermediate measures.

Gaze durations that differed more than 2.5 standard deviations from the subject means were excluded (2.44%). The outcome of the final model for gaze durations is presented in Table G5

Table G5. *Estimates, standard errors, t-values and p values for the fixed and random effects of the final general linear mixed effect model for Gaze Durations for monolingual reading.*

Monolingual					
	Estimate	SE	t-value	*p*	
Fixed Effects					
(Intercept)	2.338	0.017	141.698	<.001	**
Dutch N Density	0.0005	0.0005	1.082	.279	
English N Density	-0.001	0.001	-2.528	.012	*
Dutch N Frequency	-0.001	0.003	-0.517	.606	
English N Frequency	0.002	0.002	1.070	.285	
Word Frequency	-0.014	0.001	-11.037	<.001	***
Word Length	0.005	0.001	4.003	<.001	***
Average Bigram Frequency	0.015	0.005	3.175	.002	**
Rank of Occurrence	<0.0001	0.0001	0.437	.662	
English N Density * Word Length	-0.0005	0.0002	-2.404	.016	*

	Variance	SD
Random Effects		
Word		
(Intercept)	0.0005	0.023
Subject		
(Intercept)	0.004	0.061
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

There was again a main effect of English neighborhood density for gaze durations; these were shorter with increasing N density. We also found a significant interaction of neighborhood density with word length. Post hoc contrasts showed that there was a facilitatory effect for nouns of 5 characters or more (χ^2=4.81, df=1, $p < .05$). There was no effect of English N frequency, Dutch N density or Dutch N frequency.

Late measures.

Total reading times that differed more than 2.5 standard deviations from the subject means were excluded (2.83%). The outcome of the final model for total reading times is presented in

Table G6. For regression rate a logistic linear mixed model was fitted. The outcome of the final model for regression rate is presented in Table G7

Table G6. *Estimates, standard errors, t-values and p values for the fixed and random effects of the final general linear mixed effect model for Total Reading Times for monolingual L1 reading.*

Bilingual L1					
	Estimate	SE	t-value	*p*	
Fixed Effects					
(Intercept)	2.388	0.019	128.712	<.001	***
Dutch N Density	0.001	0.001	1.603	.109	
English N Density	-0.001	0.0004	-2.170	.030	*
Dutch N Frequency	-0.004	0.004	-1.077	.282	
English N Frequency	0.001	0.003	0.481	.631	
Word Frequency	-0.019	0.002	-10.871	<.001	***
Word Length	0.009	0.001	10.351	<.001	***
Average Bigram Frequency	0.018	0.006	2.908	.004	**
Rank of Occurrence	-0.0002	0.0001	-2.635	.008	**
English N Density * Word Frequency	0.001	0.0002	3.620	<.001	***

	Variance	SD
Random Effects		
Word		
(Intercept)	0.001	0.034
Subject		
(Intercept)	0.005	0.069
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

Table G7. *Estimates, standard errors, z-values and p values for the fixed and random effects of the final general linear mixed effect model for Regressions for monolingual L1 reading.*

Bilingual L1					
	Estimate	SE	z-value	*p*	
Fixed Effects					
(Intercept)	-2.222	0.162	-13.693	<.001	***
Dutch N Density	<-0.0001	0.009	-0.003	.998	
English N Density	0.012	0.006	1.989	.047	*
Dutch N Frequency	-0.038	0.062	-0.620	.535	
English N Frequency	0.090	0.050	1.797	.072	.
Word Frequency	-0.122	0.027	-4.587	<.001	***
Word Length	-0.079	0.014	-5.494	<.001	***
Average Bigram Frequency	0.229	0.103	2.216	.027	*
Rank of Occurrence	-0.002	0.001	-1.495	.135	

	Variance	SD
Random Effects		
Word		

(Intercept)	0.253	0.503
Subject		
(Intercept)	0.346	0.588
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

For total reading times, we found a significant main effect of English neighborhood density: reading times became shorter with increasing neighborhood density. Furthermore, there was a significant interaction with word frequency. Post hoc contrasts showed that there was a facilitatory effect of neighborhood density for words with a log frequency up to 4.00 (χ^2=3.87, df=1, $p < .05$).

For regressions there was a significant main effect of neighborhood density and marginal one of neighborhood frequency: nouns were more likely to be regressed to with an increasing within-language N density or with a higher frequent within language neighbor.

APPENDIX 5A

Results of study 6.

Table A1. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the linear mixed effect model for the first pass time.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	1.188	0.120	9.904	<.001	***
Language	-0.004	0.146	-0.026	.979	
Reading goal	-0.331	0.130	-2.541	.012	*
Information centrality	-0.053	0.141	-0.374	.710	
Average word frequency	-0.239	0.183	-1.304	.197	
Average word length	-0.024	0.072	-0.327	.745	
Unit length	0.068	0.020	3.396	.001	**
L1 Proficiency	0.007	0.015	0.456	.650	
L2 Proficiency	-0.002	0.008	-0.247	.806	
Language * Reading goal	0.231	0.143	1.617	.108	
Language * Information centrality	0.145	0.201	0.724	.471	
Language * Average word frequency	0.342	0.240	1.426	.158	
Language * Average word length	0.000	0.102	0.004	.997	
Language * Unit length	-0.030	0.027	-1.144	.257	
Language * L1 proficiency	-0.012	0.014	-0.811	.420	
language * L2 proficiency	0.001	0.008	0.074	.941	
Reading goal * Information centrality	0.096	0.127	0.756	.450	
Reading goal * Average word frequency	0.172	0.166	1.037	.301	
Reading goal * Average word length	0.115	0.066	1.753	.081	.
Reading goal * Unit length	-0.052	0.018	-2.839	.005	**
Reading goal * L1 proficiency	-0.021	0.021	-0.966	.337	
Reading goal * L2 proficiency	0.003	0.012	0.280	.780	
Language * Reading goal * Information centrality	-0.137	0.182	-0.754	.452	
Language * Reading goal * Average word frequency	-0.342	0.219	-1.561	.120	
Language * Reading goal * Average word length	-0.097	0.093	-1.038	.300	
Language * Reading goal * unit length	0.055	0.024	2.286	.023	*
Language * Reading goal * L1 proficiency	0.021	0.021	0.982	.329	
Language * Reading goal * L2 proficiency	-0.007	0.011	-0.604	.548	

	Variance	SD
Random Effects		
Unit		
Intercept	0.111	0.333

Reading goal	0.013	0.115			
Subject					
Intercept	0.168	0.410			
Language	0.089	0.299			
Residual	1.293	1.137			
p<0.1 . p<0.05 * p<0.01 ** p<0.001***					

Table A2. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the linear mixed effect model for the total reading time.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.802	0.420	6.668	<.001	***
Language	0.464	0.341	1.359	.176	
Reading goal	6.765	0.736	9.190	<.001	***
Information centrality	-0.179	0.299	-0.598	.550	
Average word frequency	-0.306	0.390	-0.784	.433	
Average word length	0.338	0.155	2.174	.030	*
Unit length	0.230	0.043	5.355	<.001	***
L1 Proficiency	-0.003	0.072	-0.038	.970	
L2 Proficiency	-0.002	0.041	-0.043	.966	
Language * Reading goal	1.638	0.778	2.105	.038	*
Language * Information centrality	0.154	0.429	0.360	.719	
Language * Average word frequency	0.047	0.516	0.091	.927	
Language * Average word length	0.047	0.219	0.213	.831	
Language * Unit length	0.075	0.057	1.315	.189	
Language * L1 proficiency	-0.002	0.050	-0.045	.964	
language * L2 proficiency	0.003	0.028	0.119	.905	
Reading goal * Information centrality	-0.109	0.757	-0.144	.886	
Reading goal * Average word frequency	0.402	0.986	0.408	.685	
Reading goal * Average word length	1.336	0.388	3.441	.001	***
Reading goal * Unit length	0.439	0.108	4.062	<.001	***
Reading goal * L1 proficiency	0.106	0.105	1.014	.314	
Reading goal * L2 proficiency	-0.037	0.057	-0.640	.524	
Language * Reading goal * Information centrality	-0.534	1.079	-0.495	.622	
Language * Reading goal * Average word frequency	-1.979	1.293	-1.531	.130	
Language * Reading goal * Average word length	-0.748	0.548	-1.364	.177	
Language * Reading goal * unit length	0.395	0.143	2.757	.007	**
Language * Reading goal * L1 proficiency	0.119	0.072	1.652	.103	
Language * Reading goal * L2 proficiency	-0.029	0.040	-0.729	.468	
	Variance	SD			

Random Effects					
Unit					
Intercept	0.029	0.169			
Reading goal	3.718	1.928			
Subject					
Intercept	4.830	2.198			
language	1.120	1.059			
Residual	14.638	3.826			
p<0.1 . p<0.05 * p<0.01 ** p<0.001***					

Table A3. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the linear mixed effect model for the fixation count.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	11.367	1.642	6.923	<.001	***
Language	1.273	1.343	0.948	.345	
Reading goal	26.990	2.819	9.574	<.001	***
Information centrality	-0.792	1.105	-0.716	.474	
Average word frequency	-0.644	1.447	-0.445	.657	
Average word length	1.496	0.577	2.594	.010	**
Unit length	0.946	0.158	5.970	<.001	***
L1 Proficiency	0.014	0.284	0.050	.961	
L2 Proficiency	-0.038	0.162	-0.237	.813	
Language * Reading goal	5.154	2.943	1.751	.083	.
Language * Information centrality	0.647	1.582	0.409	.683	
Language * Average word frequency	0.040	1.900	0.021	.983	
Language * Average word length	0.203	0.808	0.251	.802	
Language * Unit length	0.260	0.210	1.236	.217	
Language * L1 proficiency	0.012	0.203	0.059	.953	
language * L2 proficiency	0.017	0.115	0.152	.880	
Reading goal * Information centrality	0.423	2.790	0.152	.880	
Reading goal * Average word frequency	0.994	3.637	0.273	.785	
Reading goal * Average word length	5.312	1.432	3.710	<.001	***
Reading goal * Unit length	1.898	0.398	4.771	<.001	***
Reading goal * L1 proficiency	0.420	0.413	1.017	.313	
Reading goal * L2 proficiency	-0.133	0.226	-0.585	.560	
Language * Reading goal * Information centrality	-2.357	3.976	-0.593	.555	
Language * Reading goal * Average word frequency	-7.233	4.762	-1.519	.133	
Language * Reading goal * Average word length	-3.195	2.018	-1.584	.118	
Language * Reading goal * unit length	1.470	0.527	2.788	.007	**

Language * Reading goal * L1 proficiency	0.197	0.296	0.664	.509	
Language * Reading goal * L2 proficiency	0.005	0.162	0.030	.976	

	Variance	SD
Random Effects		
Unit		
Intercept	0.324	0.569
Reading goal	50.022	7.073
Subject		
Intercept	76.261	8.733
language	23.173	4.814
Residual	205.587	14.338

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Table A4. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the linear mixed effect model for the regression count.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	0.835	0.349	2.395	.018	*
Language	0.127	0.447	0.284	.777	
Reading goal	3.270	0.290	11.277	<.001	***
Information centrality	0.074	0.439	0.169	.866	
Average word frequency	0.173	0.572	0.303	.762	
Average word length	0.199	0.224	0.887	.377	
Unit length	0.055	0.062	0.876	.383	
L1 Proficiency	-0.005	0.034	-0.147	.884	
L2 Proficiency	0.003	0.020	0.157	.875	
Language * Reading goal	-0.012	0.308	-0.041	.968	
Language * Information centrality	0.062	0.625	0.100	.921	
Language * Average word frequency	-0.061	0.747	-0.081	.935	
Language * Average word length	-0.050	0.317	-0.157	.876	
Language * Unit length	0.025	0.083	0.306	.761	
Language * L1 proficiency	0.011	0.033	0.342	.733	
language * L2 proficiency	0.003	0.019	0.153	.879	
Reading goal * Information centrality	0.228	0.255	0.892	.372	
Reading goal * Average word frequency	1.084	0.334	3.248	.001	**
Reading goal * Average word length	0.811	0.132	6.122	<.001	***
Reading goal * Unit length	0.234	0.037	6.354	<.001	***
Reading goal * L1 proficiency	0.081	0.050	1.635	.107	
Reading goal * L2 proficiency	0.000	0.027	-0.018	.986	
Language * Reading goal * Information centrality	0.140	0.366	0.383	.701	

Language * Reading goal * Average word frequency	-1.346	0.441	-3.053	.002	**
Language * Reading goal * Average word length	-0.469	0.187	-2.501	.012	*
Language * Reading goal * unit length	0.046	0.049	0.944	.345	
Language * Reading goal * L1 proficiency	-0.003	0.048	-0.053	.958	
Language * Reading goal * L2 proficiency	0.012	0.026	0.452	.653	

	Variance	SD
Random Effects		
Unit		
Intercept	1.454	1.206
Subject		
Intercept	0.967	0.984
Language	0.596	0.772
Residual	5.725	2.393

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Table A5. *Estimates, standard errors, t-values and p-values for the fixed and random effects of the linear mixed effect model for the saccadic amplitude.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	2.050	0.089	23.117	<.001	***
Language	-0.267	0.065	-4.077	<.001	***
Reading goal	0.307	0.120	2.554	.012	*
Information centrality	0.021	0.066	0.318	.751	
Average word frequency	0.158	0.086	1.843	.068	.
Average word length	0.053	0.032	1.632	.106	
Unit length	0.007	0.010	0.761	.448	
L1 Proficiency	0.013	0.015	0.905	.368	
L2 Proficiency	0.020	0.008	2.320	.023	*
Language * Reading goal	-0.006	0.084	-0.073	.942	
Language * Information centrality	-0.095	0.093	-1.017	.312	
Language * Average word frequency	-0.098	0.121	-0.808	.421	
Language * Average word length	0.012	0.050	0.233	.816	
Language * Unit length	-0.006	0.013	-0.448	.655	
Language * L1 proficiency	-0.002	0.008	-0.263	.793	
language * L2 proficiency	0.000	0.004	-0.035	.972	
Reading goal * Information centrality	0.041	0.083	0.493	.624	
Reading goal * Average word frequency	0.220	0.108	2.033	.045	*
Reading goal * Average word length	0.092	0.041	2.259	.027	*
Reading goal * Unit length	-0.002	0.012	-0.167	.868	
Reading goal * L1 proficiency	-0.007	0.021	-0.328	.744	
Reading goal * L2 proficiency	-0.019	0.011	-1.631	.107	

Language * Reading goal * Information centrality	-0.102	0.118	-0.867	.388	
Language * Reading goal * Average word frequency	-0.138	0.153	-0.907	.367	
Language * Reading goal * Average word length	-0.036	0.063	-0.577	.565	
Language * Reading goal * unit length	-0.003	0.016	-0.185	.854	
Language * Reading goal * L1 proficiency	0.002	0.009	0.196	.845	
Language * Reading goal * L2 proficiency	0.013	0.005	2.448	.016	*

	Variance	SD
Random Effects		
Unit		
Intercept	0.014	0.118
Reading goal	0.030	0.173
Subject		
Intercept	0.208	0.456
language	0.015	0.122
Residual	3.937	1.984

p<0.1 . p<0.05 * p<0.01 ** p<0.001***

Table A6. *Estimates, standard errors, z-values and p-values for the fixed and random effects of the generalized linear mixed effect model for the accuracy scores.*

	Estimate	SE	t-value	p-value	
Fixed Effects					
Intercept	1.029	0.559	1.843	.065	.
Language	0.328	0.326	1.003	.316	
Reading goal	0.935	0.187	5.005	<.001	***
Information centrality	0.063	0.329	0.192	.848	
Prior knowledge	-0.074	0.047	-1.577	.115	
Text interest	-0.010	0.033	-0.313	.754	
Content difficulty	-0.106	0.038	-2.809	.005	**
Reading motivation	0.087	0.044	2.010	.044	*
Perceived reading importance	-0.026	0.072	-0.364	.716	
Reading self-efficacy	0.006	0.056	0.100	.920	
L1 proficiency	0.011	0.010	1.049	.294	
L2 proficiency	-0.001	0.006	-0.216	.829	
Language * Reading Condition	-0.192	0.254	-0.756	.450	
Language * Information centrality	-0.205	0.466	-0.441	.659	
Reading goal * Information centrality	-0.002	0.259	-0.009	.993	
Language * Reading goal * Information centrality	-0.034	0.367	-0.092	.927	

	Variance	SD
Random Effects		

Unit		
Intercept	0.761	0.872
Subject		
Intercept	0.035	0.186
p<0.1 . p<0.05 * p<0.01 ** p<0.001***		

DATA STORAGE FACT SHEETS

DATA STORAGE FACT SHEET CHAPTER 2

Author: Nicolas Dirix

Date: 23-02-2017

1. Contact details

==

1a. Main researcher

- name: Nicolas Dirix
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: nicolas.dirix@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Wouter Duyck
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: wouter.ducyk@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

==

* Reference of the publication in which the datasets are reported:

Dirix, N. (2018). Reading in a second language: An eyetracking study (Doctoral dissertation). Ghent University, Ghent, Belgium.

* Which datasets in that publication does this sheet apply to?:

All datasets reported in Chapter 2 of the doctoral dissertation

3. Information about the files that have been stored

==

3a. Raw data

----

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO

If NO, please justify:

* On which platform are the raw data stored?

- [X] researcher PC
- [X] research group file server
- [X] researcher external hard drive

* Who has direct access to the raw data (i.e., without intervention of another person)?

- [X] main researcher
- [X] responsible ZAP
- [X] all members of the research group
- [ ] all members of UGent
- [ ] other (specify): ...

3b. Other files

----

* Which other files have been stored?

- [ ] file(s) describing the transition from raw data to reported results. Specify: ...
- [X] file(s) containing processed data. Specify: Excel data files
- [ ] file(s) containing analyses. Specify:
- [ ] files(s) containing information about informed consent
- [ ] a file specifying legal and ethical provisions
- [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify:
- [ ] other files. Specify: ...

* On which platform are these other files stored?

- [X] individual PC
- [X] research group file server
- [X] researcher external hard drive

* Who has direct access to these other files (i.e., without intervention of another person)?

- [X] main researcher
- [X] responsible ZAP
- [X] all members of the research group
- [ ] all members of UGent
- [ ] other (specify): ...

4. Reproduction
==

* Have the results been reproduced independently?: [ ] YES / [X] NO
* If yes, by whom (add if multiple):

DATA STORAGE FACT SHEET CHAPTER 3

Author: Nicolas Dirix

Date: 23-02-2017

1. Contact details
==

1a. Main researcher

- name: Nicolas Dirix
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: nicolas.dirix@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Wouter Duyck
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: wouter.ducyk@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==

* Reference of the publication in which the datasets are reported:

Dirix, N. (2018). Reading in a second language: An eyetracking study (Doctoral dissertation). Ghent University, Ghent, Belgium.

* Which datasets in that publication does this sheet apply to?:

    All datasets reported in Chapter 3 of the doctoral dissertation

3. Information about the files that have been stored

===

3a. Raw data

----

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO

If NO, please justify:

* On which platform are the raw data stored?

 - [X] researcher PC
 - [X] research group file server
 - [X] researcher external hard drive

* Who has direct access to the raw data (i.e., without intervention of another person)?

 - [X] main researcher
 - [X] responsible ZAP
 - [X] all members of the research group
 - [ ] all members of UGent
 - [ ] other (specify): ...

3b. Other files

----

* Which other files have been stored?

 - [ ] file(s) describing the transition from raw data to reported results. Specify: ...
 - [X] file(s) containing processed data. Specify: Excel data files
 - [ ] file(s) containing analyses. Specify:
 - [ ] files(s) containing information about informed consent
 - [ ] a file specifying legal and ethical provisions
 - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify:
 - [ ] other files. Specify: ...

* On which platform are these other files stored?

 - [X] individual PC
 - [X] research group file server
 - [X] researcher external hard drive

* Who has direct access to these other files (i.e., without intervention of another person)?

 - [X] main researcher

- [X] responsible ZAP
- [X] all members of the research group
- [ ] all members of UGent
- [ ] other (specify): ...

4. Reproduction

===

* Have the results been reproduced independently?: [ ] YES / [X] NO
* If yes, by whom (add if multiple):

DATA STORAGE FACT SHEET CHAPTER 4

Author: Nicolas Dirix

Date: 23-02-2017

1. Contact details

===

1a. Main researcher

- name: Nicolas Dirix
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: nicolas.dirix@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Wouter Duyck
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: wouter.ducyk@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

===

* Reference of the publication in which the datasets are reported:

Dirix, N. (2018). Reading in a second language: An eyetracking study (Doctoral dissertation). Ghent University, Ghent, Belgium.

* Which datasets in that publication does this sheet apply to?:
  All datasets reported in Chapter 4 of the doctoral dissertation

3. Information about the files that have been stored
===
3a. Raw data
----
* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:
* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [X] researcher external hard drive
* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
----
* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported results. Specify: ...
  - [X] file(s) containing processed data. Specify: Excel data files
  - [ ] file(s) containing analyses. Specify:
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify:
  - [ ] other files. Specify: ...
* On which platform are these other files stored?

- [X] individual PC
- [X] research group file server
- [X] researcher external hard drive

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction

===

* Have the results been reproduced independently?: [ ] YES / [X] NO
* If yes, by whom (add if multiple):

DATA STORAGE FACT SHEET CHAPTER 5

Author: Nicolas Dirix

Date: 23-02-2017

1. Contact details

===

1a. Main researcher

- name: Nicolas Dirix
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: nicolas.dirix@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Wouter Duyck
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: wouter.ducyk@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

===

* Reference of the publication in which the datasets are reported:

Dirix, N. (2018). Reading in a second language: An eyetracking study (Doctoral dissertation). Ghent University, Ghent, Belgium.

* Which datasets in that publication does this sheet apply to?:

All datasets reported in Chapter 5X of the doctoral dissertation

3. Information about the files that have been stored

===

3a. Raw data

----

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO

If NO, please justify:

* On which platform are the raw data stored?

- [X] researcher PC
- [X] research group file server
- [X] researcher external hard drive

* Who has direct access to the raw data (i.e., without intervention of another person)?

- [X] main researcher
- [X] responsible ZAP
- [X] all members of the research group
- [ ] all members of UGent
- [ ] other (specify): ...

3b. Other files

----

* Which other files have been stored?

- [ ] file(s) describing the transition from raw data to reported results. Specify: ...
- [X] file(s) containing processed data. Specify: Excel data files
- [ ] file(s) containing analyses. Specify:
- [ ] files(s) containing information about informed consent
- [ ] a file specifying legal and ethical provisions

- [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify:
- [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server
  - [X] researcher external hard drive
* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction

===

* Have the results been reproduced independently?: [ ] YES / [X] NO
* If yes, by whom (add if multiple):

DATA STORAGE FACT SHEET CHAPTER 6

Author: Nicolas Dirix

Date: 23-02-2017

1. Contact details

===

1a. Main researcher

- name: Nicolas Dirix
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: nicolas.dirix@ugent.be

1b. Responsible Staff Member (ZAP)

- name: Wouter Duyck
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: wouter.ducyk@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
===
* Reference of the publication in which the datasets are reported:
Dirix, N. (2018). Reading in a second language: An eyetracking study (Doctoral dissertation). Ghent University, Ghent, Belgium.

* Which datasets in that publication does this sheet apply to?:
All datasets reported in Chapter 6 of the doctoral dissertation

3. Information about the files that have been stored
===
3a. Raw data
----
* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:
* On which platform are the raw data stored?
 - [X] researcher PC
 - [X] research group file server
 - [X] researcher external hard drive
* Who has direct access to the raw data (i.e., without intervention of another person)?
 - [X] main researcher
 - [X] responsible ZAP
 - [X] all members of the research group
 - [ ] all members of UGent
 - [ ] other (specify): ...

3b. Other files
----
* Which other files have been stored?
 - [ ] file(s) describing the transition from raw data to reported results. Specify: ...

- [X] file(s) containing processed data. Specify: Excel data files
- [ ] file(s) containing analyses. Specify:
- [ ] files(s) containing information about informed consent
- [ ] a file specifying legal and ethical provisions
- [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify:
- [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server
  - [X] researcher external hard drive
* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
===

* Have the results been reproduced independently?: [ ] YES / [X] NO
* If yes, by whom (add if multiple):