Visual word recognition in a second language:

A test of the lexical entrenchment hypothesis with lexical decision times

Marc Brysbaert Evelyne Lagrou Michaël Stevens

Ghent University

Belgium

Running head: Lexical entrenchment hypothesis

Keywords: bilingualism, word recognition, masked priming, lexical decision, lexical entrenchment, diffusion model

Address for correspondence:
Marc Brysbaert
Department of Experimental Psychology
Ghent University
Henri Dunantlaan 2
B-9000 Gent
Belgium
Tel. +32 9 264 94 25
Fax. +32 9 264 64 96
E-mal: marc.brysbaert@ugent.be

Abstract

The word frequency effect is stronger in second language (L2) processing than in first language (L1) processing. According to the lexical entrenchment hypothesis, this difference is not due to a qualitative difference in word processing between L1 and L2, but can be explained by differences in exposure to the target language: People with less exposure to a language show a steeper frequency curve for that language. Exposure differences can be measured with a vocabulary test. The present study tested whether the lexical entrenchment hypothesis provides an adequate explanation for differences in lexical decision times. To this end, we compared the performance of 56 Dutch-English bilinguals to that of 1011 English L1 speakers on 420 English six-letter words. In line with previous research, the differences in the word frequency effect between word processing in L1 and in L2 became vanishingly small once vocabulary size was entered as a predictor. Only in a diffusion model analysis did we find some evidence that the information build-up may be slower in L2 than in L1, independent of vocabulary size. We further report effects of cognates, age-of-acquisition, and neighborhood size that can also be explained in terms of differences in exposure.

Bilingualism is pervasive among people who do not belong to an economically and culturally dominant country (Myers-Scotton, 2006). This has encouraged scholars to investigate commonalities and differences between language processing in the mother tongue (L1) and another known, so-called second language (L2). Reviews of this research can be found in De Groot (2010), Altarriba & Isurin (2014), Heredia & Altarriba (2014), and Tokowicz (2014). We limit ourselves to studies on visual word recognition.

Evidence against selective access

For a long time, researchers started from the hypothesis that words in L1 and L2 were stored in separate lexicons, and tested whether participants had selective access to one or the other lexicon (Kroll & Stewart, 1994). The conclusion from this line of research was that selective access does not exist and that even the existence of distinct lexicons is unlikely (Brysbaert & Dijkstra, 2005; Brysbaert & Duyck, 2010; Jin, 2013; Kroll, Bobb, & Wodniecka, 2006; Tokowicz, 2014).

Much research focused on words shared between the languages, either with the same meaning (called cognates) or with different meanings (interlingual homographs). With respect to cognates, Costa, Caramazza, and Sebastian-Galles (2000) reported that bilinguals name pictures with cognate names faster than matched pictures with non-cognate names. The cognate advantage has been obtained in many other studies involving both language production and comprehension (e.g., Bultena, Dijkstra, & van Hell, 2014; Duyck, Van Assche, Drieghe, & Hartsuiker, 2007). As for interlingual homographs, Dijkstra, Timmermans, and Schriefers (2000) presented Dutch-English bilinguals with lists of English and Dutch words. The participants were to press a button only if an English word appeared. If the presented word belonged to Dutch, they were instructed to wait for the next word (i.e., a go / no-go paradigm). The authors were interested in the comparison between interlingual homographs (such as *room*, which means *cream* in Dutch) and words that only exist in English (e.g., *home*). The idea was that if participants only activated words in their English lexicon, they should not be influenced by whether or not the letter string formed a word with a different meaning in Dutch. Still, Dijkstra et al. (2000) obtained a reliable homograph effect: Participants needed more time to decide that a homograph was an English word than that a non-homograph was an English word, even though the English reading of the homograph was much more frequent than the Dutch reading and even though all test words were readily recognized as valid English words. Interestingly, Dijkstra et al. further showed that performance was affected by the other language not only when the response was required in L2, but also when the response was required for words in L1 (with homographs in L2). Participants took longer to accept a letter string as an existing Dutch word when it was an English homograph (*room*) than when it was not (e.g., *nis* [niche]).

Commonalities in L1 and L2 processing

Research on bilingual language processing has traditionally focused on differences between L1 and L2 processing. For instance, Van Heuven, Dijkstra, and Grainger (1998) examined how the recognition of L2 target words is influenced by similar words in L1 and L2. Dutch-English bilinguals and English native speakers were asked to decide whether strings of letters formed English words or nonwords (English lexical decision task). For the English native speakers, word identification time depended on the number of English orthographic neighbors (i.e., words of the same length that differ by one letter). Participants took longer to decide that a letter string was a word when it had few neighbors (e.g., *deny*, with the neighbors *defy* and *dent*) than when it had many (e.g., *dish*, with the neighbors *fish*, *wish*, *dash*, *dosh*, *disc*, *disk*). In contrast, the Dutch-English bilinguals were more influenced by the number of Dutch neighbors than by the numbers of English neighbors. Furthermore, the Dutch neighborhood effect was different from the English neighborhood effect: Dutch-English bilinguals took longer to accept an English L2 word with many Dutch L1 neighbors (e.g., *poor*, with the Dutch neighbors *boor, door, goor, hoor, koor, moor, noor, voor, pook, pool, poos, poot*) than an English word with few Dutch neighbors (e.g., *bath* with no reasonably well-known Dutch words as neighbor). This was interpreted as evidence for strong inhibitory cross-language interactions in word identification.

To chart the differences between L1 and L2 word recognition more systematically, Lemhöfer, Dijkstra, Schriefer, Baayen, Grainger, and Zwitserlood (2008) set up a large-scale study comparing English word recognition in native speakers, Dutch-English bilinguals, French-English bilinguals, and German-English bilinguals. Participants were given a word identification task (progressive demasking) with 1,025 monosyllabic English words (3-5 letters). Against their own expectations based on van Heuven et al. (1998), the authors found many more commonalities between the groups than differences. They observed a substantial overlap of reaction time patterns across the various groups of participants, indicating that the word recognition data obtained for one group generalized to the other groups. Furthermore, among the set of significant predictors, all but one reflected characteristics of the target language, English. There were virtually no influences of the bilinguals' mother tongue on their responses to English words. As a result, Lemhöfer et al. concluded that to understand English L2 word processing, it is more important to study the properties of the English language itself than possible interactions between English and the participants' mother tongue. The only robust differences Lemhöfer et al. (2008) observed between native speakers and bilinguals were related to the cognate status of the words and the word frequency effect. As for the latter, L2 speakers needed relatively more time to process low-frequency words than

L1 speakers. The larger frequency effect in bilinguals has also been reported by de Groot, Borgwaldt, Bos, & van den Eijnden (2002), Van Wijnendaele & Brysbaert (2002), Duyck, Vanderelst, Desmet, and Hartsuiker (2008), Whitford and Titone (2012), and Cop, Keuleers, Drieghe, and Duyck (2015).

The lexical entrenchment account

Diependaele, Lemhöfer, and Brysbaert (2013) examined whether the larger frequency effect in bilinguals was due to a qualitative distinction between L1 and L2 processing. A qualitative difference meant that an extra variable had to be postulated for L2 processing, that the weight of a variable differed fundamentally between L2 and L1, or that knowledge of more than one language significantly interfered with the processing of each of the languages. In contrast, if the larger frequency effect in L2 could be understood on the basis of the same mechanisms as differences in the frequency effect among L1 speakers, then this would be evidence for a system that processes L1 and L2 words in very much the same way. For instance, in L1 word recognition it has been reported that people with a small vocabulary size have a larger frequency effect than people with a large vocabulary size (Yap, Balota, Sibley, & Ratcliff, 2012). Could the difference in the frequency effect between bilinguals and native speakers also be explained by the fact that people have a smaller vocabulary size in L2 than in L1?

All participants in the Lemhöfer et al. (2008) study completed a vocabulary test and, therefore, Diependaele et al. (2013) could enter this variable as a covariate in their analysis. Once vocabulary size was taken into account, all differences between bilinguals and native speakers disappeared. Bilingual participants showed a larger frequency effect, not because they were processing words in L2, but because on average they had a smaller English vocabulary size. L2 speakers and L1 speakers with matched vocabulary sizes showed similar word frequency effects. Diependaele et al. (2013) named their finding the lexical entrenchment hypothesis: “lexical representations are weaker in low-proficiency individuals and require more energy to be processed; this is particularly true for low-frequency words”.

Kuperman and Van Dyke (2013) offered an explanation why a reduced vocabulary size correlates with an increased word frequency effect. They showed that limited exposure to language hurts the exposure to low-frequency words in particular. Large corpora yield higher frequencies of rare words than small corpora. So, people with limited exposure to a language are likely to have encountered low-frequency words considerably less than people with extensive exposure. High frequency words are encountered in large numbers by both groups and are less affected by additional exposures. The latter is a direct consequence of the fact that learning curves are concave with more impact of additional learning trials in the early stages of learning. To Kuperman and Van Dyke’s (2013) interpretation, one could add that people with a limited exposure to language are also likely to opt for easier materials (i.e., with fewer low-frequency words). For instance, it is well documented that written materials (books, newspapers, magazines) contain a richer choice of words than spoken conversations or television programs (Cunningham & Stanovich, 2001).

Importantly, the lexical entrenchment hypothesis entails that there is no qualitative difference between L1 and L2 word processing, and that any processing differences can be explained by variations in exposure. Exposure is also the driving force behind the word frequency effect and the age of acquisition (AoA) effect (early-acquired words are easier to process than late-acquired words), and arguably exposure is also involved in the cognate effect (as cognates are part of both languages). This suggests that variations in exposure to the words of a language is the main variable determining word processing times for that language, both in L1 and L2.

Following Diependaele et al. (2013) and Kuperman and Van Dyke (2013), we believe that a good vocabulary test is the best measure of language exposure we currently have (see also Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991, for a link between language exposure and vocabulary knowledge in young children). Participants exposed to less language have a smaller vocabulary.

Lexical decision and a diffusion model analysis

A limitation of the Lemhöfer et al. (2008) and the Diependaele et al. (2013) studies is that they were based on word identification in the progressive demasking paradigm. In this paradigm a word is presented between masks for increasing durations until the participant is able to identify the word. Although this task is known to correlate with other word processing times (e.g., Carreiras, Perea, & Grainger, 1997; Ferrand, Brysbaert, Keuleers, New, Bonin, Meot, Augustinova, & Pallier, 2011; Ploetz & Yates, in press), it is not the most common task in word recognition research. Many more studies are based on the lexical decision task, which shows a very clear word frequency effect (Balota et al., 2007; Ferrand, New, Brysbaert, Keuleers, Bonin, Meot, Augstinova, & Pallier, 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012). So, a test of the effect of the lexical entrenchment hypothesis on lexical decision times is needed.

A challenge for a between-groups design is to test enough participants to make sure that the participants form a representative group and that intermediate effect sizes can be detected. Lemhöfer et al. (2008) compared four groups of 21 participants (university undergraduates) each. This is good, but still provides a rather limited picture. In particular, one would like to have a larger group of L1 speakers, so that the performance of L2 speakers can be compared to the full range of L1 performances. Such a study was recently published by Adelman et al.

(2014), who tested 1011 native English speakers from 14 different universities on 420 six-letter words. By running an additional sample of Dutch-English bilingual participants, we can get a detailed picture of the position of L2 speakers relative to L1 speakers.

The large number of observations per participants and the large number of participants also allowed us to do more in-depth analyses than a simple comparison of mean reaction times (RTs). A model increasingly used to understand performance in binary forced choice RT tasks is Ratcliff's (1978) diffusion model (Dutilh, Vandekerckhove, Forstmann, Keuleers, Brysbaert, & Wagenmakers, 2012; Gomez & Perea, 2014; Ratcliff, Gomez, & McKoon, 2004). The advantage of using such a model is that it takes into account the full distribution of RTs both for correct and incorrect responses, words and nonwords, and that it captures differences between conditions with a small set of parameters, which can be linked to processing aspects. The model will be explained in more detail in the Results section, when we report the outcome of the analysis.

Method

Participants. Participants were 56 psychology undergraduates from Ghent University, Belgium. They had normal or corrected-to-normal vision and knew that the experiment involved English word recognition. All participants were native Dutch speakers and saw themselves as reasonably proficient in English. Because Adelman et al. (2014) used 28 counterbalanced lists of stimuli (see below), two participants were tested per list. To be included in the data analysis, participants had to obtain accuracy scores above 75% in the lexical decision task. A similar criterion was used in Adelman et al., as that study's focus was on the orthographic priming effect of 28 different types of stimuli expressed in milliseconds. Because 16 students did not reach the 75% criterion, they were replaced (using the same stimulus list). Ghent University students also have reasonable knowledge of French (taught in the last two years of primary school and in all years of secondary education) and sometimes of a fourth language (German, Spanish, Turkish, Hebrew, …), but this knowledge is not expected to affect the results in a way that invalidates the conclusions.

Stimuli. The 420 words and 420 nonwords from Adelman et al. (2014) were used. They were all 6 letters long. As in the Adelman et al. study, targets were preceded by a briefly presented, masked non-word prime that had various letters in common with the target word. There were 28 types of primes varying from primes that had all letters in common with the target word (i.e., identity priming) to primes that had no letters in common (unrelated primes), as shown in Table 2 below. The primes were included to test various theories of orthographic processing (the original aim of the Adelman et al. study) and were not visible to the participants. Adelman et al. used a Latin-square design to obtain data from all prime-target combinations in a group of participants who saw the target list only once. Consequently, 28 different stimulus lists were composed with 15 target words in each priming condition. As orthographic priming is expected to take place at the very first, prelexical stages of word processing, we did not expect differences in orthographic priming between our L2 participants and the L1 participants tested by Adelman et al., also because Dutch and English have very similar orthographies. Targets were presented in uppercase letters, primes in lowercase letters.

Design. The design followed the Adelman et al. (2014) study as closely as possible.[1] Participants started with the lexical decision experiment. They then proceeded with a word spelling test (not reported here) and a vocabulary test. The latter was based on Shipley (1940)

[1] The authors thank Colin Davis and Sam McCormick who kindly helped them with this.

and consisted of 40 words of increasing difficulty with four alternatives to choose from. Participants had to select the correct alternative.

Results

The full dataset, containing all information of the lexical decision task at the trial level, is available on the website of the Open Science Framework (https://osf.io/wsdxm/). This is also the case for the mixed-effects models we report, so that the analyses we report can be replicated. Our discussion involves various parts, starting with the vocabulary test. As the lexical entrenchment hypothesis makes predictions about RTs we focus on this variable (see the diffusion model below for an analysis incorporating accuracy data). Following common practice, RTs were calculated on correct trials only. Outliers were detected and removed per participant using the adjusted boxplot criterion by Hubert & Vandervieren (2008), which takes into account the positive skewness of RT distributions. Because it became clear that the vocabulary sizes of our participants were at the low end of the L1 range, we included all L1 participants available in the Adelman et al. (2014) database, so that we had a full overlap of the range of vocabulary sizes in both groups. This gave a total of 1,011 participants rather than the 924 analyzed by Adelman et al. (2014). Table 1 shows the number of participants per university.

Vocabulary test. Our participants scored on average 59.3% (SD = 9.1%) on the Shipley vocabulary test. Table 1 illustrates how this compares to the universities tested in Adelman et al. (2014). As can be seen, the average score of the L2 participants was below that of the L1 participants, although it came close to the least scoring universities. As could be expected, the vocabulary scores correlated with the accuracy data on the lexical decision task (r = .91, N = 15). Surprisingly, they did not correlate with the response times (r = .13, N = 15).

- -

Insert Table 1 about here

- -

Masked priming. Before we analyze the lexical decision data, it is important to check whether the orthographic priming effects are similar in L1 and L2, as expected. Table 2 shows the priming effects for the 28 different types of primes. As can be seen, the effects are pretty similar (correlation between the L1 and L2 effects = 0.84, N = 27, p < .0001). A mixed-effects analysis[2] on the lexical decision times confirmed that there were main effects of language (L1 vs. L2, $\chi^2_{(1)}$ = 17.21, p < .001), vocabulary size ($\chi^2_{(1)}$ = 19.83, p < .001), and type of prime ($\chi^2_{(27)}$ = 1503.6, p < .001). Participants responded faster when English was their first language, when they had a large vocabulary size, and when the orthographic overlap between prime and target increased (Table 2). Importantly, there were no interactions between prime type and language ($\chi^2_{(27)}$ = 23.34, p = .66) or between prime time and vocabulary size ($\chi^2_{(27)}$ = 37.92, p = .08)

- -

Insert Table 2 about here

[2] Linear mixed-effects models were estimated using the lme4 package in R. We followed a bottom-up model building strategy. In the first step the model included the fixed effects we wanted to test and random intercepts for items and participants. If a fixed (main) effect was significant, we added the corresponding random slopes and used a likelihood ratio test to assess whether this improved the model. Random effects were only added for measures that were repeated, as there was no variability otherwise. Word frequency, for instance, only has a random slope per participant (each participant sees items of different frequencies) but not per item (each item only has one frequency). Similarly, a random slope of vocab size was only added per item: an item is seen by participants with different vocab sizes, but a participant has only one vocab size. Applied to the analysis of the priming data, likelihood ratio tests showed that the model needed random slopes of language and vocab size per item (respectively $\chi^2_{(2)}$ = 448, p < .001 and $\chi^2_{(2)}$ = 918, p < .001) and a random slope of prime condition per item ($\chi^2_{(28)}$ = 80.66, p < .001) but not per participant ($\chi^2_{(28)}$ = 25.49, p = 0.60). To keep the computation feasible we estimated only the variances and not the covariances of the random effect of prime type.

- -

Lexical decision performance. As can be seen in Table 1, average performance of the L2 participants was in line with that of the L1 participants, although the RT was at the high end of the universities tested and the accuracy rate was at the low end. To further investigate the similarities/differences between the groups, we correlated the RTs of the groups across the 420 target words. The correlations are shown in the upper right half of Table 3. This table also includes an estimate of the reliabilities of the estimates per university placed on the diagonal (based on the Intraclass Correlation Coefficient). The reliabilities differ because the number of students tested per university varied from 28 to 217 (Table 1). Correlations can be corrected for the lack of reliability with the equation: corrected correlation = (correlation / sqrt(reliability $_{test1}$ * reliability $_{test2}$). The corrected correlations are given in the lower left half of Table 3. They clearly show the high correlation between L2 and L1 processing times (around r = .8), but the still higher correlations between the L1 data collected at the various universities (around r = .9). As was found by Lemhöfer et al. (2008), the commonalities of L1 and L2 processing outweigh the differences, but there is room for a few discrepancies, which will be outlined in the remainder of the text.

- -

Insert Table 3 about here

- -

The frequency effect and the lexical entrenchment hypothesis. The lexical entrenchment hypothesis makes two predictions: (1) participants with a small vocabulary size will show a stronger word frequency effect than participants with a large vocabulary size, and (2) once vocabulary size is taken into account, no more difference in frequency effect is expected between L1 speakers and L2 speakers.

To test the frequency effect, we made use of the SUBTLEX-UK word frequency estimates[3], expressed as Zipf-values (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The Zipf values are a standardized measure of word frequency, equal to log10(frequency per billion words), and have the following interpretation: A Zipf value of 2 equals 1 occurrence per 10 million words, Zipf 3 = 1 occurrence per million words, Zipf 4 = 10 occurrences per million words, and Zipf 5 = 100 occurrences per million words. As a rule of thumb, Zipf-values of 3 and lower can be considered as low-frequency words (equal to or lower than 1 occurrence per million words) and values of 4 and higher as high frequency words (equal to or higher than 10 occurrences per million words).

The usual finding related to the frequency effect is that the frequency effect is strong in the middle part of the continuum but levels off at the low and the high end (Keuleers et al., 2010, 2012). The leveling-off at the high end is most likely due to a floor effect in RTs. The levelling-off at the low end seems to be related to the fact that many low frequency words are not well known.[4] The consequence is that the RTs are based on smaller numbers of observations, which in addition come from the few people who know the word (and arguably have processed it more often). Keuleers, Stevens, Mandera, and Brysbaert (2015) showed that the percentage of people who know a word (a variable called ‘word prevalence’) is more informative for low-frequency words than frequency itself.

[3] Given that most data were collected in universities using British English.

[4] For empirical evidence, see the frequency effect as a function of vocabulary size in Figure 3.

The shape of the frequency effect outlined above is also present in the current dataset (Figure 1), although the leveling off at the low end starts at much higher word frequencies than seen in other megastudies (possibly because the participants of the word megastudies had larger vocabulary sizes). We tried out various ways to best capture the nonlinear nature of the frequency effect, but the most easily understandable (without loss of accuracy) is the one suggested by Harrell (2001) and depicted in Figure 1. In this approach the frequency effect is estimated via linear regression in three ranges: Low end, middle, high end. In line with Harrell's (2001) recommendation, the inflection knots were placed at the frequency percentiles 20 and 80 (i.e., the lower end included the 20% words with the lowest frequencies and the higher end included the 20% words with the highest frequencies). For the present stimulus set, these knots coincided with the Zipf values 3.047 and 4.302.

Based on a mixed-effects model with frequency as a fixed effect, a random intercept per item and participant and random slopes of the frequency effect per participant, frequency is highly significant in the middle part (β=-60.88, z=-15.14, $\chi^2_{(1)}$ = 229.315, p < 0.001) but not in the low part (β=-11.29, z=-1.01, $\chi^2_{(1)}$ = 1.022 , p=0.31) or the high part (β=-10.12, z=-1.53, $\chi^2_{(1)}$ = 2.346, p=0.13). As will become clear below, the middle range is the part where the individual differences were situated.

- -

Insert Figure 1 about here

- -

To check whether the L2 speakers had a stronger word frequency effect than the L1 speakers, as previously reported, we added language group and the interaction between language group and frequency to the above model (together with a random effect of language per item). In this analysis, the interaction between language group and frequency was significant for the middle part, but not for the lower and the higher end (see Table 4). In addition, there was a strong main effect of language group, because the L2 speakers were on average 88 ms slower (740 ms) than the L1 speakers (652 ms). Figure 2 shows the frequency effects for the L1 and L2 group.

- -

Insert Table 4 and Figure 2 about here

- -

The specific prediction of the lexical entrenchment hypothesis is that the difference in the word frequency effect between L1 and L2 speakers disappears once vocabulary size is taken into account. To test this prediction, we added vocabulary size, its random slope per item and its interaction with frequency to the model. This analysis (Table 5) showed a strong main effect of vocabulary size: The participants with the lowest vocabulary sizes (estimated as 2SD below the mean) were 64 ms slower than the participants with the highest vocabulary sizes (estimated as 2SD above the mean), with RTs of 685 ms and 621 ms respectively. More importantly, there was a strong interaction between vocabulary size and word frequency in the middle range of the frequency, but not at the lower end or the higher end, as shown in Figure 3. The word frequency effect was larger for participants with a small vocabulary than for participants with a large vocabulary. Furthermore, after adding vocabulary size, the interaction between frequency and language was not significant any more, either for the middle, lower, or higher part of the frequency range. The main effect of language remained significant.

- -

Insert Table 5 and Figure 3 about here

- -

A diffusion model analysis. In the previous analyses we saw clear evidence for a modulation of the frequency effect by vocabulary size, combined with overall slower reaction times for the Dutch-English bilinguals (even though the RTs of our bilinguals were not much longer than those of the students from the University of Arizona and Colby College; Table 1). Another way to investigate the origins of these effects is to make use of a model of the underlying processes. A model increasingly used to understand performance in binary forced choice RT tasks is Ratcliff's (1978) diffusion model (Dutilh et al., 2012; Gomez & Perea, 2014; Ratcliff et al., 2004). The advantages of the model are that it takes into account the full distribution of RTs both for correct and incorrect responses, words and nonwords, and that it captures differences between conditions with a small set of parameters. Figure 4 shows the model as it applies to a lexical decision situation. The model assumes that the information for a word or a nonword response accumulates over time, beginning from a start position until a threshold value is exceeded. The starting value, the speed with which information increases, and the position of the threshold values are parameters of the model.

- -

Insert Figure 4 about here

- -

The standard version of the diffusion model makes use of seven parameters:

1. Mean drift rate (v): This is the speed with which information accumulates. It depends on task difficulty and participant ability. Word frequency typically affects this parameter, with higher drift rates for high-frequency words than for low-frequency words (Dutilh et al., 2012; Gomez & Perea, 2014; Ratcliff et al., 2004). We expect vocabulary size to have a strong effect on this parameter. The lexical entrenchment hypothesis predicts that there will be no additional effect of L2 vs L1 once vocabulary size is taken into account. There are separate drift rates for word and nonwords.
2. Across–trial variability in drift rate (η). This parameter reflects the fact that drift rate may fluctuate from one trial to the next. As people with a large vocabulary size are more practiced, it seems sensible to expect that η decreases with vocabulary size.
3. Boundary separation (a). This variable indicates how far the boundaries are separated from each other. It quantifies response caution and modulates the speed–accuracy tradeoff. Given that bilinguals took longer to respond but made more errors, it is not clear what to expect for this parameter.
4. Mean starting point (z): This variable reflects the bias participants have towards word or nonwords responses. It might be hypothesized, for instance, that participants with a small vocabulary size show a stronger bias towards nonwords responses, as they know fewer words.
5. Across–trial variability in starting point (s_z). This parameter reflects the fact that the starting point may fluctuate from one trial to the next. Given that participants with a large vocabulary have more practice with words, a likely expectation is that variability will decrease with vocabulary size.
6. The non–decision component of processing (T_{er}). This parameter represents the time needed to encode the stimulus and execute the response, irrespective of information accumulation and decision. Finding a difference between L2 and L1 speakers on this

parameter would suggest that the main effect of language group has little to do with word processing. On the other hand, both Dutilh et al. (2012) and Gomez and Perea (2014) found a clear effect of word frequency on T_{er}. So, the interpretation of this variable is less clear for word processing than originally assumed.

7. Across–trial variability in the non–decision component of processing (s_T). As for the previous variability parameters, the explanation would be most straightforward if the variability decreased as a function of vocabulary size.

By fitting the model to the data of each participant, we can enter the resulting parameter estimates in multiple regression analyses with language group (L1, L2) and vocabulary size as predictors. To estimate the parameters of the diffusion model, we made use of the fast-dm algorithm written by Voss & Voss (2007).

- -

Insert Table 6 about here

- -

Table 6 shows the estimates of the various parameters, together with the z-values for the effects of language group and vocabulary size. Language group has a significant effect on the drift rate for words and on the non-decision time. Vocabulary size had a significant effect on nearly all parameters.

Starting with the most interesting parameter, we see that the drift rate *v* differs as a function of vocabulary size, as expected: Participants with a large vocabulary size have a higher drift rate than participants with a low vocabulary size. At the same time, L2 speakers have a lower drift rate than L1 speakers for words. Figure 5 shows both effects. The variability in drift rate (η) was smaller for participants with a high vocabulary size, in line with the assumption that processing went more smoothly for them.

- -

Insert Figure 5 about here

- -

There were no clear effects on boundary separation (parameter a) when we corrected for multiple comparisons. If a more lenient criterion is used, L2 speakers had their boundaries slightly lower than L1 speakers, meaning that they based their decisions on less information. This explains their higher error rates. Interestingly, the boundaries were not influenced by vocabulary size. Figure 6 shows how the a-parameter changes as a function of language group and vocabulary size.

- -

Insert Figure 6 about here

- -

All participants had a bias towards words (i.e., the starting point was closer to the word boundary than to the nonword boundary, as shown in Figure 7). Against expectation, participants with a large vocabulary had a less strong word bias than participants with a small vocabulary.

- -

Insert Figure 7 about here

- -

There was a 70 ms difference in T_{er} between L2 and L1 speakers, indicating that the main effect of language group on RT was largely due to factors outside the word recognition and decision processes. At the same time, there was no difference between people with a small and a large vocabulary. These findings agree with the observation that a considerable variability was observed in the mean RTs between the English-speaking universities as well, without corresponding differences in vocabulary size (Table 1).

- -

Insert Figure 8 about here

- -

Finally, the variabilites of T_{er} and z had opposite effects as a function of vocabulary size. Whereas the variability in T_{er} decreased for participants with a large vocabulary, as expected, the variability in z (the starting point) increased. It is not clear how to interpret the latter finding. Maybe good participants are more flexible in their starting point and make it shift more as a function of the stimulus sequence just processed (e.g., a streak of words or nonwords; Dufau, Grainger, & Ziegler, 2012)?

Cognates, age-of-acquisition, and neighbors. Given the richness of the dataset, it is worthwhile to further test three variables that have been claimed to affect L2 word recognition differently than L1 word recognition. This allows us not only to further chart the differences between L1 and L2 processing, but also to test the quality of the dataset. If none of these effects could be found, we would have to conclude that the dataset is less interesting than we had hoped for. The three variables claimed to have different effects in L1 and L2 are cognates, age-of-acquisition (AoA), and neighbors in L1 and L2. Importantly for bilingualism researchers, AoA refers to the age at which English words are acquired in English L1 speakers, not the age at which an L2 is learned. These variables were added simultaneously to the model of Table 5 (see Table 7).

- -

Insert Table 7 about here

- -

As indicated in the Introduction, cognate words are expected to be easier for bilinguals than non-cognate words. Based on the Dutch-English cognate list compiled by Schepens, Dijkstra, and Grootjen (2012), 126 of the 420 target words were Dutch-English cognates. As predicted, bilinguals were 26 ms faster on cognates than on noncognates (z=-4.81, p < 0.001). This was significantly larger than the difference seen in L1 speakers (z=-3.56, p<0.001; Figure 9), even though the L1 speakers also responded 11 ms faster to the cognates than the noncognates (z=-3.20, p<0.001), indicating that researchers must be very careful when they investigate the cognate effect, as the effect could be due to other variables if it is not contrasted against an L1 group. Also reassuring is that the cognate effect did not depend on vocabulary size, as the cognate effect is thought to be present in all bilinguals.

- -

Insert Figure 9 about here

- -

Izura and Ellis (2002) reported that the AoA effect in L2 depends on the order of acquisition of the L2 words and not on the order of acquisition of the L1 words. Given that most of our bilingual participants started to learn English at the age of 12-14 years, the words they first acquired were different from the words an English toddler is learning. So, if Izura and Ellis (2002) are right, we ought to find a stronger AoA effect, based on English L1 AoA estimates, for L1 speakers than for L2 speakers. The AoA measures were taken from Kuperman, Stadthagen-Gonzalez, and Brysbaert (2013). As Figure 10 and Table 7 show, there was indeed a significant interaction between AoA and language group in the predicted direction. We found an AoA effect for L1 speakers (β=3.61, z=5.34, $p < 0.001$), but not quite for L2 speakers (β=1.58, z=1.41, p=0.156), although there was a trend in the right direction. AoA did not interact with vocabulary size, as was expected given that the AoA effect is assumed to be present for all L1 speakers.

- -

Insert Figure 10 about here

- -

As described in the introduction, van Heuven et al. (1998) reported that intra-language neighbors had a facilitation effect on English lexical decision times, but that inter-language neighbors had an inhibition effect for bilinguals. We could test this pattern of results in our data as well.[5] Because the length of the stimuli was longer in the present dataset (6 letters) than in Van Heuven et al. (3-5 letter words), the number of neighbors is considerably less. However, this is likely to be an advantage, because the effect of word neighbors is particularly

[5] The authors thank Nicolas Dirix for pointing them to this possibility.

robust between 0 and 1 neighbor (Davis, 2010). As it happens, 221 out of the 420 words did not have an English neighbor, and only 74/420 words had at least one Dutch neighbor.[6]

As can be seen in Figure 11, the effect of English neighborhood size was facilitatory, both for the L1 and the L2 speakers. The effect was best captured with the log(neighborhood size + 1) transformation as predictor. This transformation takes into account that the effect of word neighborhood size is particularly strong for differences between small sizes. The effect of English neighborhood was larger for participants (both L1 and L2) with a small vocabulary size.

- -

Insert Figure 11 about here

- -

The Dutch neighborhood size had no effect, also not for the L2 speakers separately. There was a hint of an interaction with vocabulary size, as the effect tended to be facilitatory for participants with a small vocabulary but inhibitory for participants with a large vocabulary size. However, this interaction was present to the same extent for L1 and L2 speakers and, hence, is unlikely to be specific to knowledge of the Dutch language.

Discussion

Bilinguals show a stronger frequency effect in L2 than in L1 (Cop et al., 2015; de Groot et al. 2002; Duyck et al., 2008; Lemhöfer et al., 2008; Van Wijnendaele & Brysbaert, 2002; Whitford & Titone, 2012). According to the lexical entrenchment hypothesis (Diependaele et

[6] Neighbors were calculated on the basis of Celex (Baayen, Piepenbrock, & Gulikers, 1995) and had to have a frequency of at least 2 per million in that database. The same criteria were used in Van Heuven et al. (1998). The authors thank Walter van Heuven for providing them with the neighborhood sizes.

al., 2013), this difference can be explained on the basis of a more limited exposure to L2 than to L1, and requires no further explanation. A good proxy of language exposure is vocabulary size (see also Kuperman & Van Dyke, 2013). Once a person's vocabulary size is taken into account, there are no further differences between L2 and L1 processing.

The present study tests the lexical entrenchment hypothesis with lexical decision data. We made use of a database in which lexical decision times for 420 six-letter English words had been collected from 1011 native speakers at 14 different universities. To this database, we added the records of 56 Dutch-English bilinguals with overlapping vocabulary sizes. In line with previous findings, there was a clear interaction between language group and word frequency: The frequency effect was stronger for the L2 speakers than for the L1 speakers (Table 4 and Figure 2). More importantly, when vocabulary size was introduced as a covariate, the interaction largely disappeared (Table 5), as reported by Diependaele et al. (2013). Bilinguals show a stronger word frequency effect in L2, not because a second language is harder to process, but because participants have had less exposure to this language than the average native speaker. Once the degree of exposure (estimated via vocabulary size) is taken into account, the frequency effects in L1 and L2 become equivalent.

Further evidence that L2 word processing is better explained in terms of exposure to L2 than in terms of interactions with L1 can be seen in the effects of cognates, AoA, and word neighbors. Each of these effects can be explained in terms of exposure. Because cognates exist in both languages and have the same meaning, bilingual participants have been exposed to them more often and, hence, show a cognate advantage (Figure 9). Interestingly, the English L1 speakers also showed a (smaller) cognate effect. This has been reported before (Mulder, Dijkstra, Schreuder, & Baayen, 2014) and related to the fact that cognates tend to be the same in many languages. As a result, they are the words that English speaking students may pick up most easily when they are abroad or have some shallow knowledge of another language.

The age-of-acquisition effect is attributed to the order of acquisition and to the fact that a learning network loses plasticity the more stimuli of a particular kind it already knows (Monaghan & Ellis, 2010). Interestingly, the AoA effect in L2 is related to the order of word acquisition in L2 and not to the order of acquisition in L1 (Izura & Ellis, 2002). As a result, English AoA estimates should be better predictors of L1 processing times than of L2 processing times, as we indeed observed (Figure 10). The fact that the AoA effect is not completely absent for L2 speakers is in line with the hypothesis that the AoA effect is not entirely situated in the connections between the representations but also has an effect on the organization of the semantic system, with the meaning of early-acquired words being more accessible than the meaning of late-acquired words (Brysbaert & Ellis, in press; Brysbaert, Van Wijnendaele, & De Deyne, 2000). Importantly for the present discussion, the most straightforward interpretation of the difference in AoA effect between L1 and L2 word processing refers to differences in (the order of) exposure to the English words.

Finally, we observed that reaction times to English words were influenced by the number of English orthographic neighbors, but not by the number of Dutch orthographic neighbors. The former is in line with van Heuven et al. (1998). The effect is present to a similar extent in the English Lexicon Project (as checked on the basis of Balota et al., 2007) and, therefore, is not something peculiar to the present experiment (e.g., due to the fact that the target words were preceded by orthographic primes). The absence of an effect due to Dutch neighbors contrasts with van Heuven et al. (1998), who found an inhibitory effect of Dutch neighbors for Dutch-

English bilinguals. As indicated in the introduction, the pattern of results reported by van Heuven et al. (1998) did not agree with the later findings of Lemhöfer et al. (2008) or Diependaele et al. (2013). Our findings are further evidence that this aspect of the van Heuven et al. (1998) data may be less solid than assumed thus far. On the other hand, it should be taken into account that our study was not well suited to measure the effects of cross-language, Dutch neighbors. Less than 20% of the words had Dutch neighbors and no attempts were made to make the Dutch neighborhood size orthogonal to the English neighborhood size. So, the null-effect has to be treated very cautiously.

The facilitation effect of within-language English neighbors was stronger for participants with a small vocabulary size than for participants with a large vocabulary size (Figure 11). This is in line with the hypothesis that the neighborhood size effect on lexical decision times is the result of a balance between (a) facilitation due to the fact that a word looks more wordlike when it has neighbors, and (b) inhibition because it is more difficult to distinguish two visually similar words (Andrews, 1997; Grainger & Jacobs, 1996). Because a lexical decision can often be made on the basis of an overall familiarity feeling rather than the identification of the exact word presented, word neighborhood facilitation effects are often observed in lexical decision experiments (Andrews, 1997). This is particularly true for participants with lower English proficiency levels (Andrews & Hersch, 2010). Important for the present discussion is that the effect of orthographic neighbors depends on the English vocabulary size of the participants and not on whether English was their L2 or L1 (Figure 11).

So far, the analyses are all in line with the lexical entrenchment hypothesis: Differences between L1 and L2 processing can be explained in terms of differences in exposure to the target language, which can be measured with a good vocabulary test, and do not need the inclusion of further mechanisms. A slightly more complicated picture emerges, however, when we analyze the data with the diffusion model (Ratcliff, 1978). Then we see that the similar RTs in L1 and L2, once vocabulary size is filtered out, are not achieved in exactly the same way. In particular, there is some evidence that lexical information builds up more slowly in L2 than in L1, and that this is compensated by a stronger word bias and more risky decision boundaries in L2 speakers (Figures 5-7). This would suggest that L2 word processing is genuinely harder than L1 word processing (e.g., because of extra competition from the L1 words). A complicating factor for this explanation is that the slower information build-up is not observed for non-words, making it hard to decide whether there is a genuine difference between L1 and L2 processing in terms of the diffusion model parameters, or whether the differences observed are due to some overfitting of the model or because the vocabulary test we used failed to pick up all differences between L1 and L2 speakers. Given that the effects of language on the parameters of the diffusion model are rather modest and not entirely convergent, for the time being we prefer to treat them as an observation, to be kept in mind when analyzing new data but not strong enough to refute the lexical entrenchment hypothesis. A further interesting research question may be to investigate whether similar effects would be found in L1 processing between bilinguals and monolinguals, to find out whether knowledge of another language has an impact on the processing of the native language. Such research would require a considerable investment, however, as the participant samples must be large enough to have good power to disentangle the effect of language status from the effect due to differences in vocabulary size.

All in all, our findings largely agree with the conclusions of Lemhöfer et al. (2008) and Diependaele et al. (2013) that in order to understand L2 word processing, it is much more important to study the characteristics of the L2 words, rather than possible ways in which L1 and L2 words interfere with each other. All the differences between L1 and L2 word processing we obtained could be understood on the basis of discrepancies in the exposure to the English language, which can be estimated by means of an objective vocabulary test.[7]

Although it may be tempting to interpret the absence of an interaction between Dutch and English words as evidence for separate lexicons (in which the English L2 words are insulated from the Dutch L1 words), we do not think such a conclusion is warranted. As indicated in the Introduction, there is quite a lot of evidence that the bilingual lexicon is unitary (Brysbaert & Dijkstra, 2005; Brysbaert & Duyck, 2010; Jin, 2013; Kroll, Bobb, & Wodniecka, 2006; Tokowicz, 2014). In addition, interpreting a lack of interaction between Dutch and English words as evidence for distinct lexicons only makes sense in the presence of clear interactions between the English words themselves. Such interactions should have taken the form of an inhibition effect between English orthographic neighbors. The fact that we found a facilitation effect can only be explained by assuming that the lexical decision times were partly based on the overall “English” activity in the mental lexicon (Andrews, 1997; Grainger & Jacobs, 1996). Such overall activity can as well be present in a bilingual Dutch-English lexicon as in a full English lexicon. Apparently, RTs from a lexical decision task are not well suited to expose the competition process between orthographically similar entries in the mental lexicon, contrary to what the data of van Heuven et al. (1998) originally suggested.[8] Ferrand et al. (2011) reported a similar lack of orthographic competition effect on response times in the progressive demasking task. The most likely reason for the insensitivity of both tasks to

[7] Therefore, we strongly recommend all language researchers to use such tests (whether studying L1 or L2), so that the findings from various studies can be related to each other. Two tests in English are Shipley (1940) and LexTALE (Lemhöfer & Broersma, 2012). The LexTALE test was also administered to the participants of our test and correlated r = .74 with the Shipley scores (N = 56, p < .01).

[8] The ideal paradigm to reveal inhibition effects makes use of masked priming with high frequency orthographic neighbors preceding low frequency target words (Davis & Lupker, 2006; De Moor, Verguts, & Brysbaert, 2005; Segui & Grainger, 1990). Measures other than RT may also be indicated. For instance, Massol, Grainger, Dufau, & Holcomb (2010) and Laszlo & Federmeier (2011) reported stronger effects on ERP signals.

orthographic competition is that the size of the effect is considerably smaller than the exposure-based effects reported here and in Diependaele et al. (2013). This, in our view, is the reason why the lexical entrenchment hypothesis is such a good account for the RTs obtained in progressive demasking and lexical decision.

References

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., ... & Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods, 46 (4)*, 1052-1067.

Altarriba, J., & Isurin, L. (Eds.) (2014). *Memory, Language, and Bilingualism: Theoretical and Applied Approaches.* Cambridge University Press.

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

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General, 139(2),* 299-318.

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39(3),* 445-459.

Brysbaert, M. & Dijkstra, T. (2006). Changing views on word recognition in bilinguals. In J. Morais & G. d'Ydewalle (Eds.), *Bilingualism and second language acquisition*. Brussels: KVAB.

Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the revised hierarchical model of bilingual language processing after 15 years of service? *Bilingualism: Language and Cognition, 13,* 359-371.

Brysbaert, M., & Ellis, A.W. (In press). Aphasia and age-of-acquisition: Are early-learned words more resilient? *Aphasiology*.

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. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *The Quarterly Journal of Experimental Psychology, 67(6),* 1214-1241.

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

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency Effects in Monolingual and Bilingual Natural Reading. *Psychonomic Bulletin and Review, 20*, 963-972.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(5),* 1283-1296.

Cunningham, A. E., & Stanovich, K. E. (2001). What reading does for the mind. *Journal of Direct Instruction, 1*, 137–149.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117(3)*, 713-758.

Davis, C. J., & Lupker, S. J. (2006). Masked inhibitory priming in English: Evidence for lexical inhibition. *Journal of Experimental Psychology: Human Perception & Performance, 32*, 668–687.

de Groot, A.M.B. (2010). *Language and Cognition in Bilinguals and Multilinguals: An Introduction.* Hove: Psychology Press.

de Groot, A. M. B., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language, 47,* 91–124.

De Moor, W., Verguts, T., & Brysbaert, M. (2005). Testing the “multiple” in the multiple read-out model of visual word recognition*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1502-1508.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first and second language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology, 66,* 843-863.

Dijkstra, T., Timmermans, M., & Schriefers, H. (2000). On being blinded by your other language: Effects of task demands on interlingual homograph recognition. *Journal of Memory and Language, 42,* 445-464.

Dufau, S., Grainger, J., & Ziegler, J. C. (2012). How to say “no” to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(4),* 1117.

Dutilh, G., Vandekerckhove, J., Forstmann, B.U., Keuleers, E., Brysbaert, M., & Wagenmakers, E.J. (2012). Testing theories of post–error slowing. *Attention, Perception, & Psychophysics, 74,* 454-465.

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.

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.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from Chronolex. *Frontiers in Psychology, 2:306*. doi: 10.3389/fpsyg.2011.00306.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42,* 488-496.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from Chronolex. Frontiers in Psychology, 2:306. doi: 10.3389/fpsyg.2011.00306.

Gomez, P., & Perea, M. (2014). Decomposing encoding and decisional components in visual-word recognition: A diffusion model analysis. *Quarterly Journal of Experimental Psychology, 67*, 2455-2466.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological Review, 103(3),* 518-565.

Harrell, F. E., Jr. (2001). *Regression modeling strategies.* Berlin, Germany: Springer.

Heredia, R., & Altarriba, J. (Eds.) (2014). *Foundations of bilingual memory*. New York: Springer.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis, 52,* 5186-5201.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27(2),* 236-248.

Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicológica, 23(2)*, 245-282.

Jin, Z. (2013). Nonselective access of English phonology in bi-scriptal Chinese–Korean visual word recognition. *Cognitive Processing, 14(4),* 435-441.

Johnson, P.C.D. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution, 5*, 944-946.

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 1:174*. doi: 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*, 287-304.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (in press). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*.

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition, 9(02),* 119-135.

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.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44,* 978-990.

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance, 39(3),* 802-823.

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology, 48(2),* 176-186.

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

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwisterlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory & Cognition, 34,* 12-31.

Massol, S., Grainger, J., Dufau, S., & Holcomb, P. (2010). Masked priming from orthographic neighbors: An ERP investigation. *Journal of Experimental Psychology: Human Perception and Performance, 36(1),* 162-174.

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.

Myers-Scotton, C. (2006). *Multiple voices: An introduction to bilingualism*. Malden, MA: Blackwell Publishing.

Ploetz, D. M., & Yates, M. (In press). Age of acquisition and imageability: A cross‐task comparison. *Journal of Research in Reading*.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111(1),* 159.

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.

Segui, J., & Grainger, J. (1990). Priming word recognition with orthographic neighbors: Effects of relative prime-target frequency. *Journal of Experimental Psychology: Human Perception and Performance, 16(1),* 65-76.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology, 9,* 371–377.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86(2),* 420-428.

Tokowicz, N. (2014). *Lexical processing and second language acquisition*. Routledge.

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.

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,* 1176-1190.

Van Wijnendaele, I., & Brysbaert, M. (2002). Visual word recognition in bilinguals: Phonological priming from the second to the first language. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 616-627.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39,* 767–775.

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.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53-79.

Table 1: Comparison between the universities tested (in increasing order of vocabulary size). Notice that at all universities, except for Ghent University, English was the native language of the participants. N $_{\text{students}}$ = the number of participants tested at each university present in the database. Vocab = the score on the Shipley vocabulary test. Accuracy $_{\text{LDT}}$ = the accuracy in the lexical decision task. RT $_{\text{unrelated pseudoword prime}}$ = the average reaction time to the target words preceded by unrelated pseudoword primes (see Table 2 to learn how the RTs differed as a function of the type of orthographic prime).

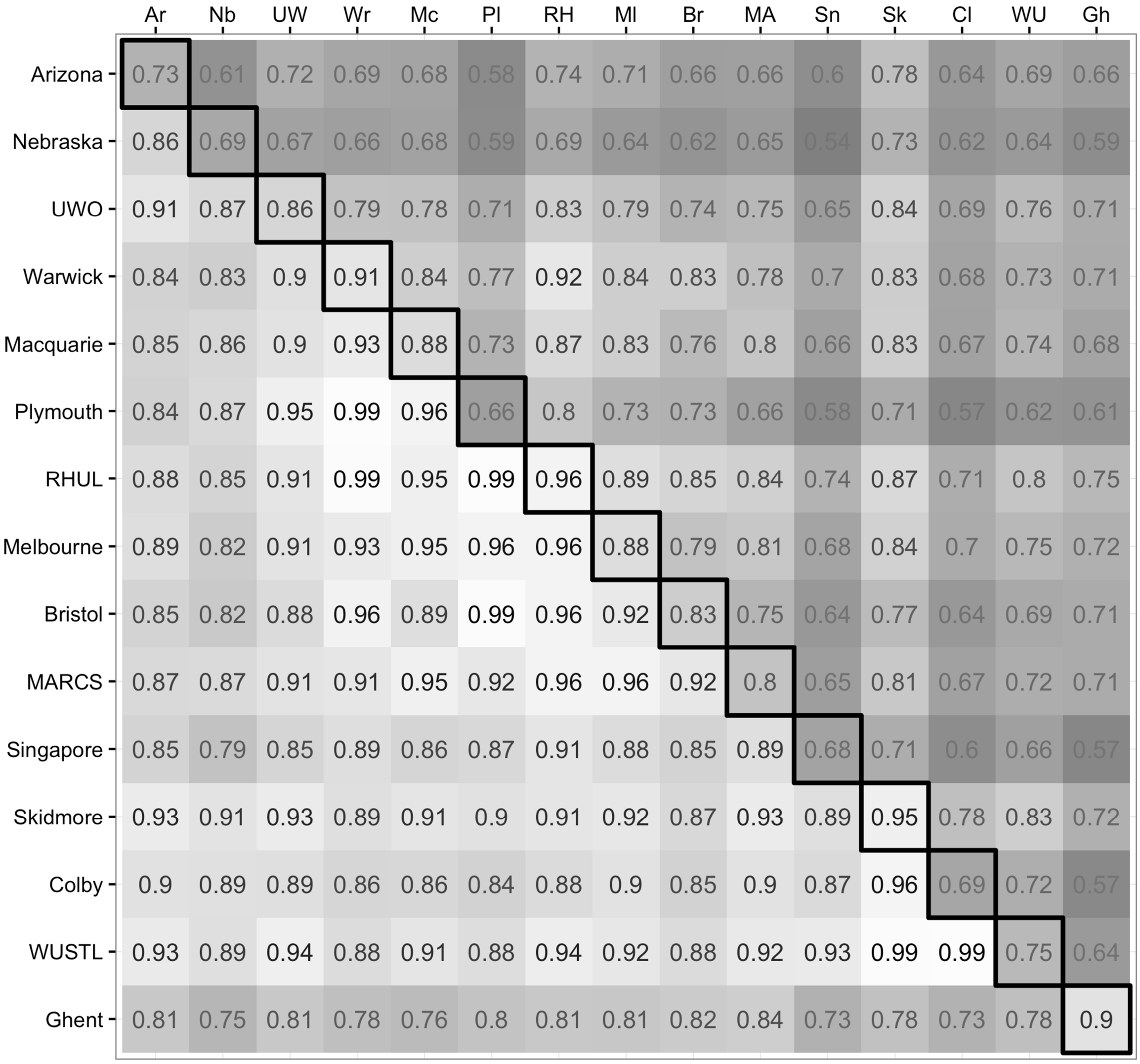
Site	N $_{\text{students}}$	Vocab	Vocab sd	Vocab min	Vocab max	Accuracy LDT	RT unrelated pseudo word prime
Ghent	56	59.3	9.1	40.0	82.5	78.8	742
Arizona	28	64.6	9.8	42.5	85.0	87.7	727
Nebraska	29	66.8	12.5	35.0	87.5	90.2	338[9]
UWO	60	68.4	11.2	32.5	92.5	88.9	668
Warwick	119	71.1	8.3	52.5	95.0	91.2	686
Macquarie	65	72.7	12.8	27.5	90.0	89.8	624
Plymouth	28	72.9	10.6	52.5	92.5	92.2	703
RHUL	217	72.9	9.9	40.0	97.5	91.2	624
Melbourne	66	73.1	9.9	47.5	92.5	90.7	698
Bristol	59	73.6	9.4	45.0	100.0	90.0	690
MARCS	31	75.2	11.7	42.5	97.5	90.1	644
Singapore	28	76.1	7.9	52.5	90.0	92.2	687
Skidmore	197	76.1	9.3	40.0	95.0	93.6	709
Colby	28	80.2	7.5	65.0	92.5	94.5	726
WUSTL	56	81.6	9.0	55.0	95.0	94.1	667

[9] This value is the one obtained from the dataset. In all likelihood, it is caused by a different starting point of the timer, as the RTs correlate as well with the other data as can be expected on the basis of the reliability of the data. Importantly, all analyses we report can handle a constant subtraction (e.g., due to inclusion of an intercept difference between participants or to the inclusion of Ter in the diffusion model). So, the conclusions we draw are not influenced by this measurement error.

Table 2: Orthographic priming effects for 28 different types of primes, expressed in milliseconds relative to the unrelated pseudoword condition. The L1 data correspond to the values reported by Adelman et al. (2014) but based on 1011 participants; the L2 data are the average values of the 56 Dutch-English bilinguals.

Prime	Example Target = DESIGN	L1 priming	L2 priming
Identity	design	31.2	43.9
Initial transposition	edsign	21.5	22.5
Medial transposition	desgin	22.2	16.3
Final transposition	desing	22.9	33.9
2-apart transposition	degisn	13.2	22.9
3-apart transposition	dgsien	4.4	12.5
Medial deletion	dsign	20.8	25.0
Final deletion	desig	24.2	33.8
Central double deletion	degn	17.6	15.8
All-transposed	edisng	11.3	18.3
Transposed halves	igndes	5.3	13.8
Half	des	18.2	25.1
Reversed halves	sedngi	6.0	5.9
Interleaved halves	idgens	1.5	11.6
Reversed (except initial)	dngise	-2.2	3.7
Initial substitution	pesign	20.3	34.5
Medial substitution	desihn	14.3	18.2
Final substitution	desigj	20.0	22.9
Neighbor once removed	dslign	13.3	14.5
Central double substitution	dewvgn	9.0	5.4
Central insertion	desrign	20.3	24.5
Central double insertion	desaxign	12.4	28.4
As above, repeated letter	deshhign	17.6	20.4
Central quadruple subst.	dzbtkn	-3.6	11.5
Prefix	mdesign	18.3	24.3
Suffix	designl	24.1	32.7
Unrelated pseudoword	voctal	-0.0	-0.0
Unrelated arbitrary	cbhaux	-5.8	4.2

Table 3: Correlations between the reaction times of the various universities (based on the 420 word targets). Values on the diagonal represent the reliability of the RT estimates for each university (measured by means of the intraclass correlation coefficient; Shrout & Fleiss, 1979). Values above the diagonal show the raw correlations; values below the diagonal show the correlations corrected for the reliability of the variables. The lighter the cell, the higher the correlation.



	Ar	Nb	UW	Wr	Mc	Pl	RH	Ml	Br	MA	Sn	Sk	Cl	WU	Gh
Arizona	0.73	0.61	0.72	0.69	0.68	0.58	0.74	0.71	0.66	0.66	0.6	0.78	0.64	0.69	0.66
Nebraska	0.86	0.69	0.67	0.66	0.68	0.59	0.69	0.64	0.62	0.65	0.54	0.73	0.62	0.64	0.59
UWO	0.91	0.87	0.86	0.79	0.78	0.71	0.83	0.79	0.74	0.75	0.65	0.84	0.69	0.76	0.71
Warwick	0.84	0.83	0.9	0.91	0.84	0.77	0.92	0.84	0.83	0.78	0.7	0.83	0.68	0.73	0.71
Macquarie	0.85	0.86	0.9	0.93	0.88	0.73	0.87	0.83	0.76	0.8	0.66	0.83	0.67	0.74	0.68
Plymouth	0.84	0.87	0.95	0.99	0.96	0.66	0.8	0.73	0.73	0.66	0.58	0.71	0.57	0.62	0.61
RHUL	0.88	0.85	0.91	0.99	0.95	0.99	0.96	0.89	0.85	0.84	0.74	0.87	0.71	0.8	0.75
Melbourne	0.89	0.82	0.91	0.93	0.95	0.96	0.96	0.88	0.79	0.81	0.68	0.84	0.7	0.75	0.72
Bristol	0.85	0.82	0.88	0.96	0.89	0.99	0.96	0.92	0.83	0.75	0.64	0.77	0.64	0.69	0.71
MARCS	0.87	0.87	0.91	0.91	0.95	0.92	0.96	0.96	0.92	0.8	0.65	0.81	0.67	0.72	0.71
Singapore	0.85	0.79	0.85	0.89	0.86	0.87	0.91	0.88	0.85	0.89	0.68	0.71	0.6	0.66	0.57
Skidmore	0.93	0.91	0.93	0.89	0.91	0.9	0.91	0.92	0.87	0.93	0.89	0.95	0.78	0.83	0.72
Colby	0.9	0.89	0.89	0.86	0.86	0.84	0.88	0.9	0.85	0.9	0.87	0.96	0.69	0.72	0.57
WUSTL	0.93	0.89	0.94	0.88	0.91	0.88	0.94	0.92	0.88	0.92	0.93	0.99	0.99	0.75	0.64
Ghent	0.81	0.75	0.81	0.78	0.76	0.8	0.81	0.81	0.82	0.84	0.73	0.78	0.73	0.78	0.9

Table 4: Fixed effects in the mixed-effects model comprising frequency and language. The residual standard deviation of the model was 143.1 ms.

	Chisq	df	p
language	32.029	1	0.000
low end frequency	1.174	1	0.279
medium frequency	222.906	1	0.000
high end frequency	2.024	1	0.155
low end frequency : language	0.796	1	0.372
medium frequency : language	13.223	1	0.000
high end frequency : language	2.673	1	0.102

Table 5: Fixed effects in the mixed-effects model comprising frequency, language, and vocabulary size. The residual standard deviation of the model was 142.9 ms.

	Chisq	df	p
language	17.515	1	0.000
vocabulary	19.730	1	0.000
low end frequency	1.115	1	0.291
medium frequency	225.406	1	0.000
high end frequency	1.996	1	0.158
low end frequency : language	0.206	1	0.650
medium frequency : language	1.386	1	0.239
high end frequency : language	1.744	1	0.187
low end frequency : vocabulary	2.379	1	0.123
medium frequency : vocabulary	96.622	1	0.000
high end frequency : vocabulary	1.653	1	0.199

Table 6: Values of the estimates of the diffusion parameters for the L1 and L2 speakers, and the corresponding z-values for the effects of language group and vocabulary size. The significance tests took into account the fact that multiple post-hoc comparisons were made using Dunn-Sidak correction. Given the fact that we were looking at 7 separate analyses, the critical absolute z-values corresponding to p-values of 0.05, 0.01 and 0.001 were 2.69, 3.19 and 3.81. The estimates of T_{er} and s_T are in milliseconds.

	v_{words}	$v_{nonwrds}$	η	a	z	s_z	T_{er}	s_T
L1	2.59	-3.27	1.12	1.34	0.62	0.16	470	160
L2	1.86	-3.35	1.22	1.24	0.64	0.12	540	180
Language group	6.53**	0.75	1.78	-2.58	2.53	-2.45	6.68**	2.14
Vocab size	19.2**	-12.22**	-3.85**	1.33	-5.56**	4.30**	-0.23	-6.61**

** p < .001

Table 7: Fixed effects in the mixed-effects model comprising frequency, language, vocabulary size, cognates, AoA, and neighbors in L1 and L2. The marginal R^2 was 4.49%, the conditional R^2 was 43.11%. Adding the item predictors to the model significantly increased the fit relative to the previous model ($\chi^2_{(37)}$=101701, $p < 0.001$). The residual standard deviation of the model was 142.8 ms.

	Chisq	df	p
language	20.463	1	0.000
vocabulary	20.677	1	0.000
low end frequency	1.293	1	0.255
medium frequency	129.017	1	0.000
high end frequency	0.122	1	0.727
aoa	28.542	1	0.000
cognates	10.256	1	0.001
English neighbors	3.089	1	0.079
Dutch neighbors	0.099	1	0.753
low end frequency : language	1.019	1	0.313
medium frequency : language	3.352	1	0.067
high end frequency : language	1.575	1	0.209
aoa : language	4.943	1	0.026
cognates : language	12.639	1	0.000
English neighbors : language	0.074	1	0.785
Dutch neighbors : language	0.064	1	0.801
low end frequency : vocabulary	2.572	1	0.109
medium frequency : vocabulary	70.763	1	0.000
high end frequency : vocabulary	0.437	1	0.508
aoa : vocabulary	1.235	1	0.266
cognates : vocabulary	3.815	1	0.051
English neighbors : vocabulary	10.989	1	0.001
Dutch neighbors: vocabulary	5.536	1	0.019

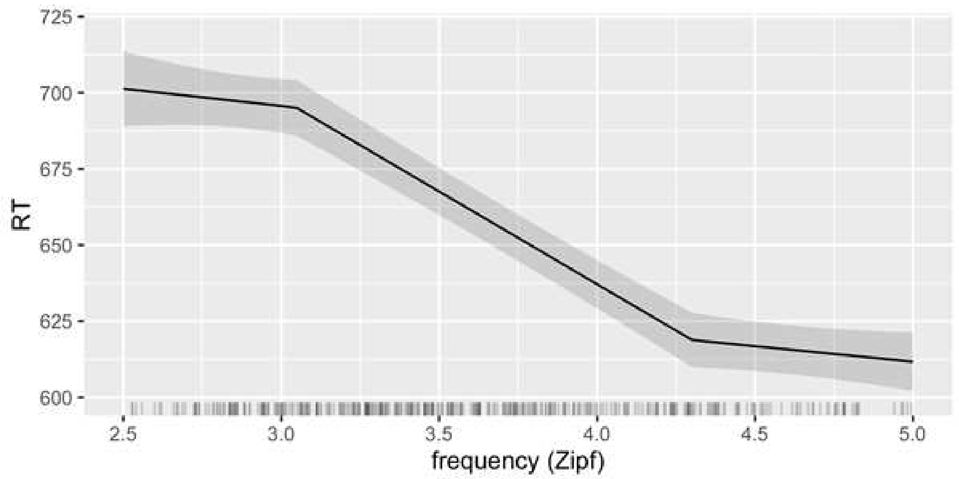


Figure 1: The mean frequency effect for all participants, based on the model with frequency as the only fixed effect. This shows that the frequency effect was particularly strong for the middle part of the frequency range (see the text for the factors causing this pattern and for the break points used to distinguish between low frequency, medium frequency, and high frequency words). The short vertical lines on the abscissa show the distribution of the stimulus words. The marginal R^2 (fixed effects only) of the model was 2.81%, the conditional R^2 (fixed and random effects) was 42.01%. See Johnson (2014) for a discussion of R^2 for mixed-effects models. The grey area indicates the 95% confidence interval.

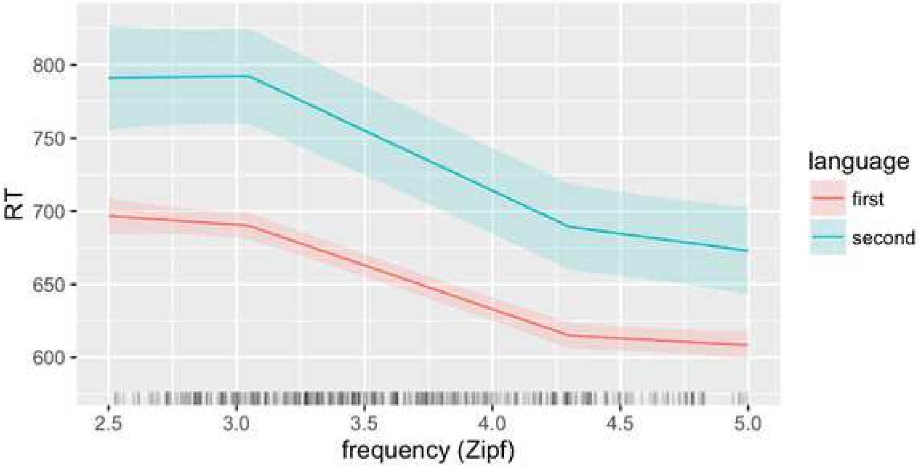


Figure 2: Frequency effect split up by language group, based on the model with frequency and language as fixed effects. The marginal R^2 was 3.66%, the conditional R^2 was 42.11%. Adding the effect(s) of language to the model significantly increased the fit relative to the frequency-only model ($\chi^2_{(6)}$=414, p < 0.001). See the digital version for a colored graph.

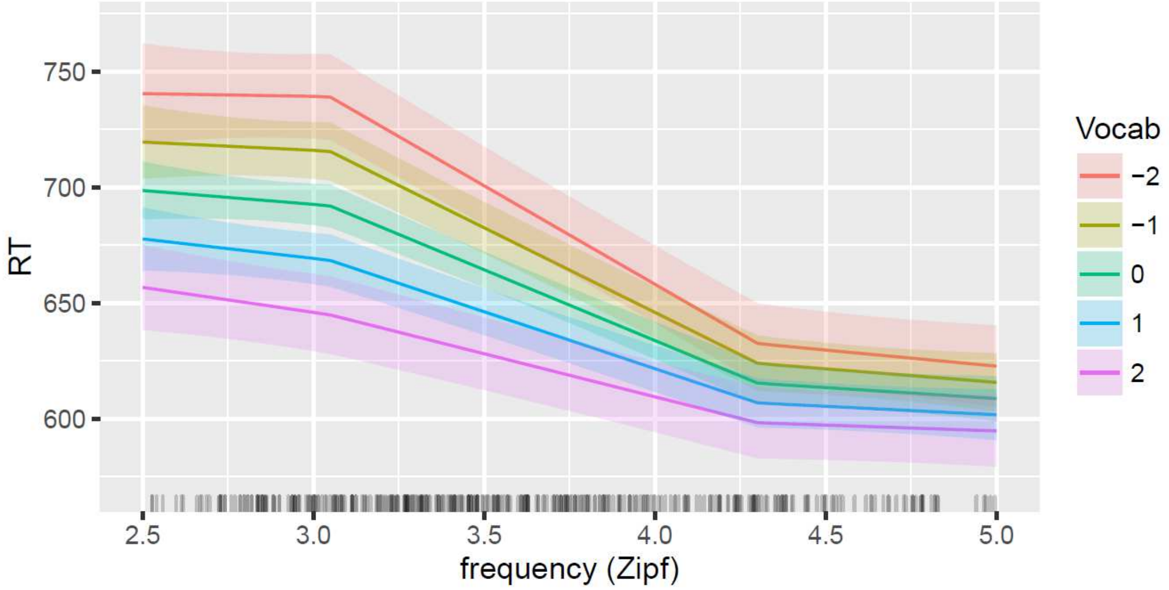


Figure 3: Interaction between vocabulary size (Cvocab) and word frequency, based on the model with frequency, language and vocabulary size as fixed effects. The lowest line represents the RTs of the participants with the highest vocabulary size (2 standard deviations above the mean); the highest line represents the RTs of the participants with the lowest vocabulary size (2 standard deviations below the mean). The marginal R^2 was 4.29%, the conditional R^2 was 42.21%. Adding the effect(s) of vocabulary size to the model significantly increased the fit relative to the frequency plus language model ($\chi^2_{(7)}$=500, $p < 0.001$). See the digital version for a colored graph.

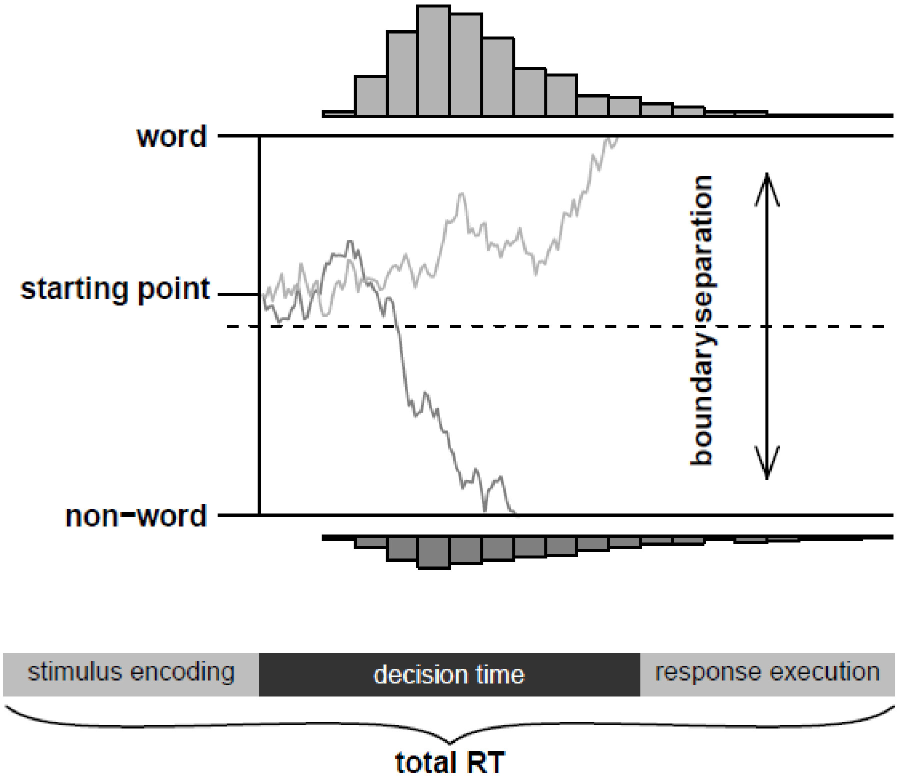


Figure 4: A diffusion model analysis of the lexical decision task. When a stimulus is presented, noisy evidence accumulates either towards the word (top) or the nonword decision boundary (bottom). In the figure the accumulation of two different stimuli is shown, one which results in a word decision and one that results in a nonword decision. The reaction time distributions (represented by the bar charts at the top and the bottom of the figure) and the errors are used to estimate the best fitting parameters of the model.
(Source: Dutilh et al., 2012)

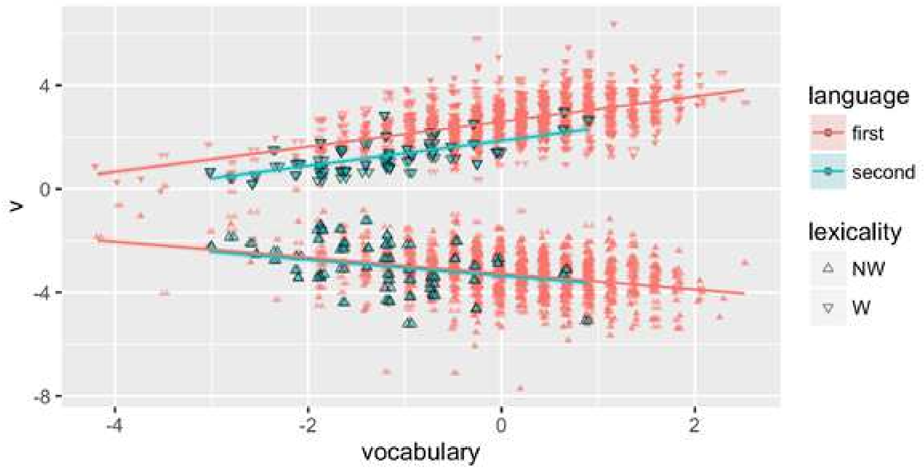


Figure 5: Drift rates (v) as a function of vocabulary size (centered with 0 equal to the median value), language group, and word (top half) vs. nonwords (bottom half). This figure shows that the drift rate is steeper for participants with a large vocabulary size than for participants with a small vocabulary size. In addition, it shows that for words, but not for nonwords, there is an additional difference between L1 and L2 speakers. In order to show all the data, the points are slightly jittered around the obtained vocabulary values. See the digital version for a colored graph.

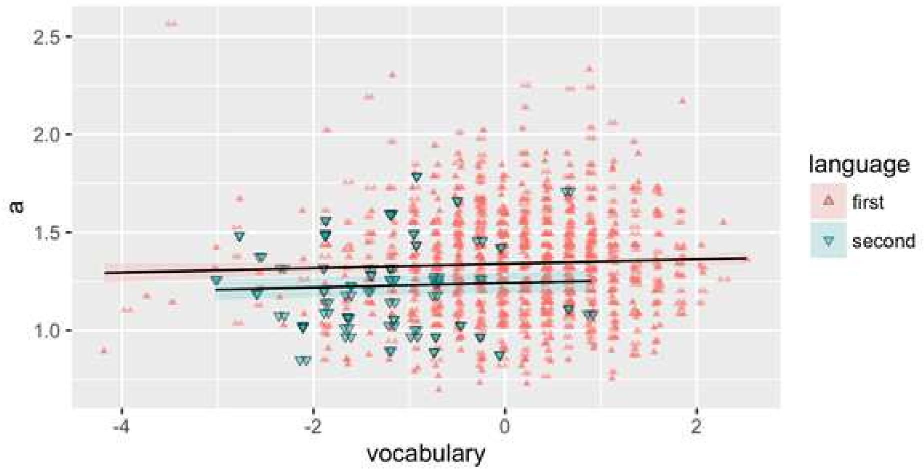


Figure 6: Boundary (a) as a function of vocabulary size (centered with 0 equal to the median value) and language group. This figure shows that the boundaries were slightly further apart for the L1 speakers than for the L2 speakers. There was no effect of vocabulary size. See the digital version for a colored graph.

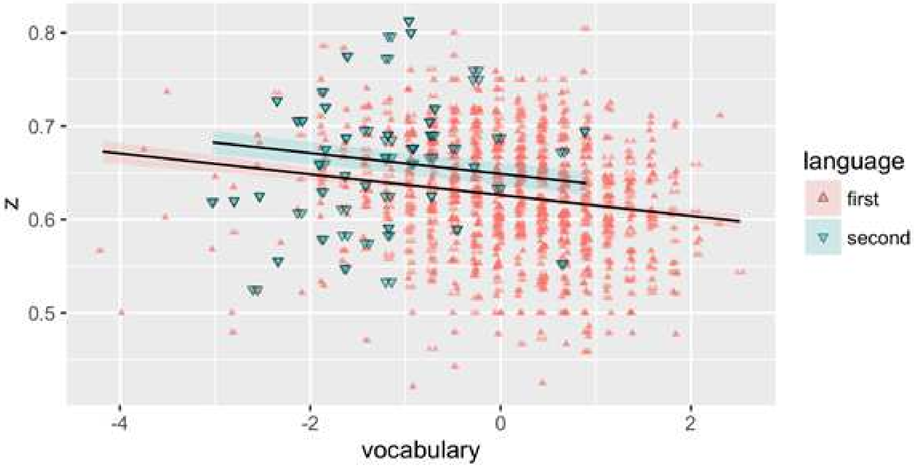


Figure 7: Bias (z) as a function of vocabulary size and language group. All participants showed a bias towards words (positive z-values). The bias decreased as vocabulary size increased, and tended to be stronger for L2 speakers. See the digital version for a colored graph.

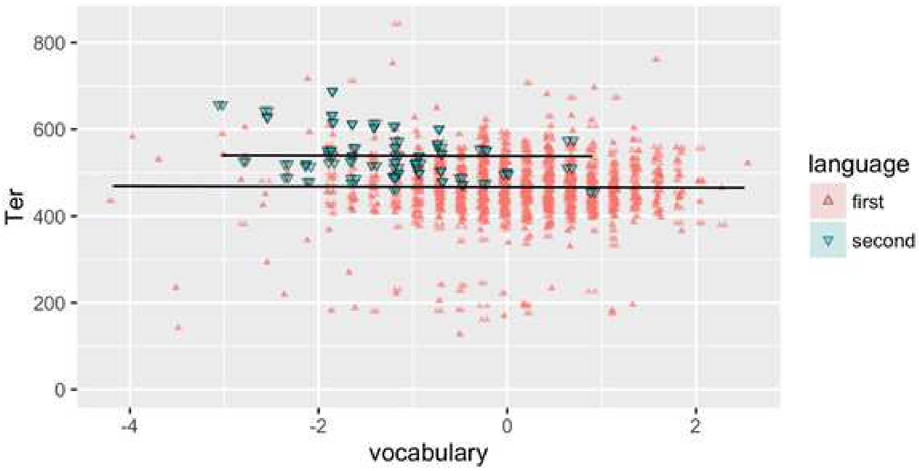


Figure 8: Non-decision time (T_{er}) as a function of vocabulary size and language group. L2 participants had Ter values 70 ms longer than L1 speakers. There was no effect of vocabulary size. See the digital version for a colored graph.

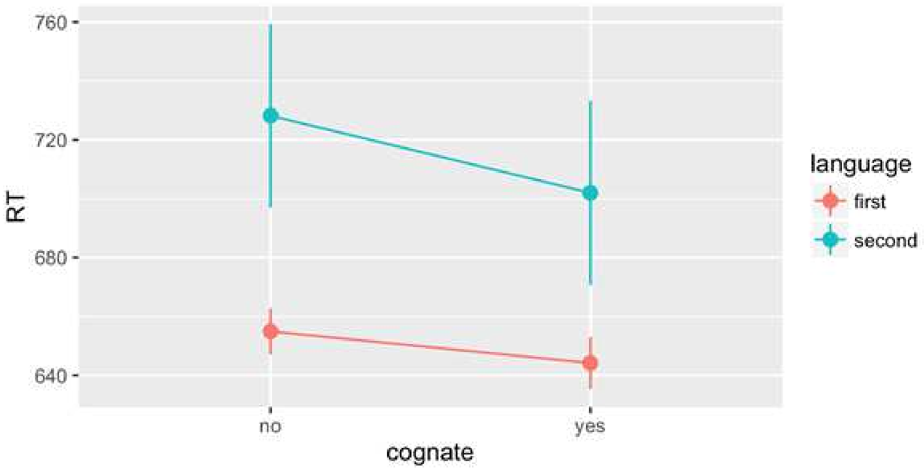


Figure 9: The cognate effect for bilinguals and monolinguals. The cognate advantage is present in both groups but significantly stronger for the L2 group. See the digital version for a colored graph.

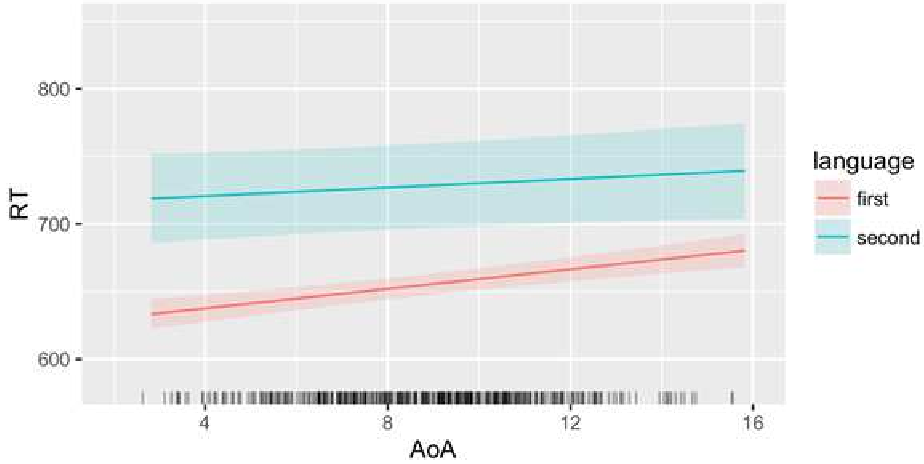


Figure 10: The AoA effect for bilinguals and monolinguals. AoA refers to the age (in years) at which words are thought to be acquired in English, based on the ratings collected by Kuperman et al. (2012). The effect is present for the L1 group. See the digital version for a colored graph.

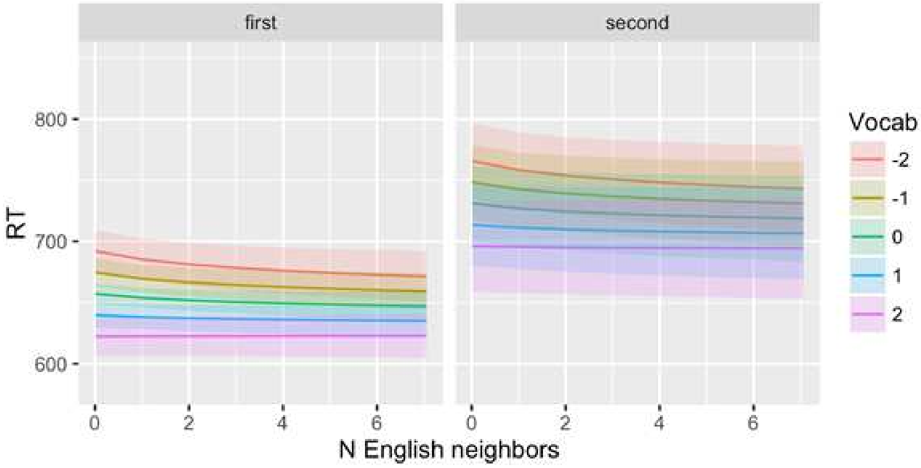


Figure 11: Effect of English neighborhood N on RTs. N stands for the number of English words that are orthographic neighbors of the target words. The effect was facilitatory, in particular for participants with a small vocabulary. There was no difference between L1 and L2 speakers. See the digital version for a colored graph.