



# Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment

Emmanuel Keuleers, Michaël Stevens, Paweł Manderer & Marc Brysbaert

**To cite this article:** Emmanuel Keuleers, Michaël Stevens, Paweł Manderer & Marc Brysbaert (2015) Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment, *The Quarterly Journal of Experimental Psychology*, 68:8, 1665-1692, DOI: [10.1080/17470218.2015.1022560](https://doi.org/10.1080/17470218.2015.1022560)

**To link to this article:** <http://dx.doi.org/10.1080/17470218.2015.1022560>



Accepted author version posted online: 25 Feb 2015.  
Published online: 08 Apr 2015.



[Submit your article to this journal](#)



Article views: 204



[View related articles](#)



[View Crossmark data](#)




Citing articles: 2 [View citing articles](#)

# Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment

---

---

Emmanuel Keuleers , Michaël Stevens, Paweł Mandera, and Marc Brysbaert

Department of Experimental Psychology, Ghent University, Gent, Belgium

(Received 4 June 2014; accepted 18 February 2015; first published online 9 April 2015)

We use the results of a large online experiment on word knowledge in Dutch to investigate variables influencing vocabulary size in a large population and to examine the effect of word prevalence—the percentage of a population knowing a word—as a measure of word occurrence. Nearly 300,000 participants were presented with about 70 word stimuli (selected from a list of 53,000 words) in an adapted lexical decision task. We identify age, education, and multilingualism as the most important factors influencing vocabulary size. The results suggest that the accumulation of vocabulary throughout life and in multiple languages mirrors the logarithmic growth of number of types with number of tokens observed in text corpora (Herdan's law). Moreover, the vocabulary that multilinguals acquire in related languages seems to increase their first language (L1) vocabulary size and outweighs the loss caused by decreased exposure to L1. In addition, we show that corpus word frequency and prevalence are complementary measures of word occurrence covering a broad range of language experiences. Prevalence is shown to be the strongest independent predictor of word processing times in the Dutch Lexicon Project, making it an important variable for psycholinguistic research.

*Keywords:* Prevalence; Frequency; Crowdsourcing; Herdan's law; Ageing; Bilingualism.

Experimental research on language processing has traditionally taken place on groups of students attending the institutions where behavioural laboratories are located. As a consequence, not much is known about the variability in language processing in the population at large. However, it can be assumed that language processing is for a large part driven by an individual's accumulated linguistic experiences. This suggests that the better we can model an individual's linguistic experience, the better we can explain language processing. For instance, Van Heuven, Mandera, Keuleers, and Brysbaert (2014) have shown that British English word frequencies are better at explaining lexical

decision data collected from British students than from American students and that, vice versa, the SUBTLEX-US frequencies (Brysbaert & New, 2009) outperform the SUBTLEX-UK frequencies for behavioural data collected on US students. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) showed that HAL (hyperspace analogue to language) frequencies, which are derived from a corpus of internet newsgroups, better predict the lexical decision performance of younger adults than that of older adults, consistent with the possibility that the younger adults in their study were more likely to use the internet. In another study, Kuperman and Van Dyke (2013)

---

Correspondence should be addressed to Emmanuel Keuleers, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Gent, Belgium. E-mail: [emmanuel.keuleers@ugent.be](mailto:emmanuel.keuleers@ugent.be)

This work was supported by an Odysseus grant awarded by the Government of Flanders to Marc Brysbaert.

collected subjective word frequencies from participant groups with different levels of reading experience. They then matched these subjective frequencies to participants in other experiments (based on their reported reading experience level) and found that eye movement and lexical decision latencies were better predicted by the matched subjective frequencies than by corpus word frequency. However well we are able to model the language experience of participants in experiments, we still cannot avoid that in most cases the participants come from a small and relatively homogeneous population of college students. Research must also move from small homogeneous groups of participants to large heterogeneous groups of language users. Some existing applications of *crowdsourcing* in psycholinguistics used the Amazon Mechanical Turk framework for collecting acceptability judgements (Gibson, Piantadosi, & Fedorenko, 2011), age-of-acquisition ratings (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), valence, arousal, and dominance ratings (Warriner, Kuperman, & Brysbaert, 2013), and concreteness ratings (Brysbaert, Warriner, & Kuperman, 2014).

An earlier attempt at recruiting large heterogeneous groups of participants for a lexical decision experiment was made by Dufau et al. (2011), who introduced the ScienceXL platform. Individuals could download an app for iPhone and iPad and perform lexical decision in different languages at leisure. The app is available in many languages, and the results, which are still being collected, will undoubtedly be useful to the scientific community. A particular feature of the app is that the user does not need to be online to do the experiment. Instead, participants can send results by e-mail when the device is connected.

Although offline platforms like ScienceXL have advantages, a permanent connection to the internet is quickly becoming the norm. Since web browsers are also ubiquitous on connected devices, the choice to perform experiments online becomes much more appealing. An immense advantage of this approach is the direct communication between the device on which the experiment is performed and the infrastructure on which the

data are collected, increasing the probability that the data that are generated by the users are actually collected. In addition, online experiments do not require participants to install special software and can be designed to run on a large range of devices. Beyond this, online experiments make it easier to add educational and social components to a study. As we discuss in more detail later, participants in our study were able to look up the meaning of stimuli in an online dictionary and could easily share their score via social media, which we believe increased participation and participant satisfaction.

The current study combines elements of language proficiency testing with elements of megastudies, but instead of using a limited set of validated items to test vocabulary we used random samples from a very large list of words. This has the advantage that participants can do the test as often as they like and, more importantly, that we can collect data on tens of thousands of words. While a lexical decision task defines word knowledge as the ability to distinguish a word from a nonword, other tests that focus on knowing the correct meaning of words (e.g., Nation & Beglar, 2007) may give a different estimate of vocabulary size. In this context, it is important to note that research suggests that part of the same underlying construct is being measured: Stubbe (2012) found a correlation of .82 between testing vocabulary size with a yes/no tests and a multiple choice test. Lemhöfer and Broersma (2012) found a correlation of .72 between the yes/no test and translation scores. Yap, Balota, Sibley, and Ratcliff (2012) found a lower correlation of .62 between individual participants' lexical decision accuracy in the English Lexicon Project and their performance on the Shipley Vocabulary of Living scale (Zachary & Shipley, 1986).

In this paper, we use the results of a large-scale vocabulary test to address two issues. First, what are the effects of age, gender, degree of multilingualism, second language (L2), L2 proficiency, education, handedness, and location on vocabulary size in a very large population? Second how can word prevalence be used as a measure of word occurrence?

## Variables contributing to vocabulary size

In our choice of variables, we were motivated by two questions: (a) How does the variable contribute to our understanding of differences in vocabulary size, which was our immediate research goal, and (b) how can the variable contribute to research in a more general way? We were also restricted in the number of data we could collect, as we were entirely dependent on the willingness of a large number of individuals to contribute to our experiment on a voluntary basis, and this willingness is likely to decrease with increasing survey length.

### *Age*

Research investigating the relation between cognitive skills and age generally concludes that vocabulary size increases with age (McCabe, Roediger, McDaniel, Balota, & Hambrick, 2010; Park et al., 2002) or shows a smaller decrease than tasks emphasizing speed and short-term or working memory (Singh-Manoux et al., 2012). It is worth noting that the typical vocabulary tests used in research on the relation between age and cognitive skills consist of relatively few items and use the same words for all participants. For instance, the vocabulary part of the Shipley scale (Zachary & Shipley, 1986) uses 40 words, and the Mill Hill vocabulary test (Raven, 1965) uses 33 words. While the current study uses around 70 words per participant, its biggest advantage is that the stimuli were drawn from a master list containing over 50,000 words. On the aggregate level, this virtually eliminates list bias and allows for a more detailed evaluation of the relation between age and vocabulary size.

### *Education*

Education is commonly used as a control variable in research on vocabulary size. It can be assumed that formal education exposes people to specific vocabulary associated with new knowledge domains and that dictionaries codify many words that are associated with formal education. Hence, an increase in vocabulary with increasing education is expected, and it would be surprising if we find that education does not have an effect on vocabulary size.

### *Multilingualism and foreign language proficiency*

Almost every person living in Belgium and the Netherlands is exposed to foreign languages to some extent. Exposure to English is ubiquitous through TV, popular music, and advertising. In addition, the proximity of areas where French or German is the dominant language and the foreign language background of a sizeable part of the population contribute to widespread multilingualism, spanning the entire proficiency spectrum. As is detailed later, we asked participants three questions regarding their knowledge of foreign languages: “How many foreign languages do you know?”, “What is your best foreign language?” (henceforth L2), and “How well do you know this language?” (henceforth L2 proficiency). The current study therefore makes it possible to evaluate two important questions on a large scale. The first concerns the effect of L2 knowledge and proficiency on L1 vocabulary size. The second question covers the effect of the number of foreign languages known on L1 vocabulary size.

A common assumption in the bilingualism literature is that bilinguals do not speak as much as monolinguals in each language, but that they divide their word usage among the different languages they know (Gollan, Montoya, Cera, & Sandoval, 2008). Further assuming that usage and exposure are similar in this respect, a first language (L1) speaker who also has exposure to other languages will have less exposure to L1 than a monolingual person. A naive interpretation of this position would lead to the prediction that, all else being equal, Dutch L1 vocabulary size should decrease with the number of foreign languages a participant knows.

However, we can think of two ways in which exposure to multiple languages may mediate the effect that reduced L1 exposure has on vocabulary size, even when the assumption of equal total exposure is maintained. The first is that languages often do not have completely independent vocabularies. In the case of our study, the acquisition of vocabulary in a foreign but related language may partly contribute to the acquisition of vocabulary in Dutch through cognates—that is, words with the same meaning and a very similar form.

Schepens, Dijkstra, and Grootjen (2012, Table 1) estimate that about 20% of English and French words and over 40% of German words have a Dutch cognate. The vocabulary size for a particular language will therefore be the sum of the vocabulary that is exclusive to that language and the vocabulary that is shared with other known languages. This shared vocabulary could compensate the decrease in L1 vocabulary by decreased exposure.

The second way in which foreign language exposure can mitigate decreased L1 exposure can be derived from the typical relation between exposure and vocabulary growth (Herdan, 1960). Herdan's law tells us that the probability of encountering a new word type decreases with the number of encountered word tokens. In other words, the more of a language one has been exposed to, the slower the rate of increase in vocabulary in that language. Since proficient L1 speakers have already been exposed to a great amount of L1, the rate at which they will acquire new L1 vocabulary will be quite small. Hence, the disadvantage of being exposed to other languages instead of L1 will also be quite small. However, since the probability of encountering new types is much higher when being exposed to a foreign language, the initial growth rate of foreign vocabulary will be higher. As a first consequence, the total multilingual vocabulary for a person who is exposed to multiple languages could be larger than the total vocabulary of an L1 speaker with same amount of total exposure. Adding to this, the more that foreign language exposure leads to shared vocabulary with L1, the less that decreased L1 exposure will impact vocabulary size.

The effects of foreign language knowledge and foreign language proficiency on the vocabulary size of the participants in our study can give us a clearer insight in how L1 vocabulary size is determined by the interplay between the number of tokens that one has been exposed to in L1, the structural relationship between types and tokens (Herdan's law), and the shared vocabulary between the languages. Depending on the balance of these factors, we can expect a number of results. If L1 exposure is the dominant factor, we can expect a

decrease of vocabulary size with the number of foreign languages spoken. If Herdan's law and/or shared vocabulary play an important role, then we can expect an increase with the number of foreign languages spoken. If shared vocabulary plays an independent role, languages with more cognates should lead to a larger vocabulary size. Independent of shared vocabulary, simultaneous decreased exposure in L1 and a slowing vocabulary growth rate in L2 due to Herdan's law imply that vocabulary size should not increase forever with foreign language proficiency. To anticipate our results: A *U*-shaped curve for the effects of second language proficiency on L1 vocabulary size is a clear indicator of the interplay between Herdan's law and L1 exposure.

#### *Gender*

The factor gender was primarily included in our study to differentiate between male and female vocabulary. However, it is of auxiliary interest regarding the folk assumption that women are more talkative than men, characterized by Mehl, Vazire, Ramirez-Esparza, Slatcher, and Pennebaker (2007, p. 82) as "deeply engrained in Western folklore and often considered a scientific fact". Mehl et al. did not find evidence for this assumption. Likewise, we do not know of any scientific basis for assuming a gender difference in vocabulary size.

#### *Handedness*

We included handedness in our study to control for a dominant hand bias on reaction times. We do not expect any differences of this variable on vocabulary size.

#### *Location*

Location was included in our study primarily to document specific vocabulary differences in Belgium and the Netherlands. To anticipate our results, while we did not initially expect an effect of location on vocabulary size, this turned out to be an interesting factor.

## Prevalence

While the analysis of participant scores gives us information about vocabulary size in different subsets of participants, psycholinguistic research can also benefit from having precise information about the actual knowledge of each word. For instance, when setting up factorial experiments where reaction time is the measure of interest, it is generally not useful to include words that are unknown by participants, as incorrect responses are not taken into account.

We introduce the notion of *word prevalence* to mean the proportion of a population knowing a particular word. Of course, the larger and the more diverse the sample is, the better an estimate becomes. Therefore, the current collection of yes/no responses on a large set of words and on a large sample of participants may give us a reliable indication of the degree to which each word is known in the population. While accuracy data from megastudies also tell us something about the prevalence of a large number of words, the number of observations for each word is usually quite small (e.g., about 29 in the English Lexicon Project and about 40 in the Dutch Lexicon Project and the British Lexicon Project). In addition, the participants usually come from a homogeneous population of university students.

Word prevalence may be an important theoretical measure of word occurrence. At first, this may seem odd, since we are used to approximating word occurrence with word frequency counts from corpora. However, a drawback of frequency counts is that, regardless of corpus size, lower counts are unreliable. As an example, consider asking a random sample of 100 people whether they know each of the word types that occur just once in a large corpus. Although frequency for all these types is equal, the number of judges knowing each word will vary widely. As the judges are also producers of language, words known to many of them may be considered to

occur more often in language than words that are known by fewer of them. Following this reasoning, the estimate of the number of language users who know a word may be a better indication of occurrence than corpus frequency counts for low-frequency words. On the other hand, consider presenting the same random sample of people with words from the language's core vocabulary. Since these words will be known to all of the judges, prevalence will be singularly high and uninformative. In this case, corpus counts should be a much better estimate of occurrence.

To our knowledge, the notion that prevalence can be thought of as a measure of occurrence has not been discussed in the literature. Interestingly, prevalence should work where frequency counts are low and uninformative, and frequency counts should be predictive where prevalence is uninformative.

To test the idea that word frequency and word prevalence are complementary measures of occurrence, it is necessary to evaluate the measure on a task where the effect of word frequency is clear and well understood and where there are data for a wide range of stimuli. For Dutch, the lexical decision data of the Dutch Lexicon Project (Keuleers, Diependaele, & Brysbaert, 2010) satisfy both criteria.

## EXPERIMENTAL STUDY

### Method

#### *Participants*

We analysed data collected from 16th March to 15th September 2013. Up to that point, 572,146 tests had been finished by an estimated 368,798 individual participants.<sup>1</sup> Figure 1 shows that the peak of participation happened right after the test started, indicating that participation in the experiment was both viral and ephemeral. We gathered about 55% of the data in the first four days of

<sup>1</sup>Whenever profile data were saved or changed on a device, we saved a new profile identifier to that device. The number of unique profile identifiers associated to one or more finished sessions gives us a rough estimate of the number of participants, keeping in mind that multiple participants could have used the same identifier and that the same individual can have different identifiers because they use multiple devices.



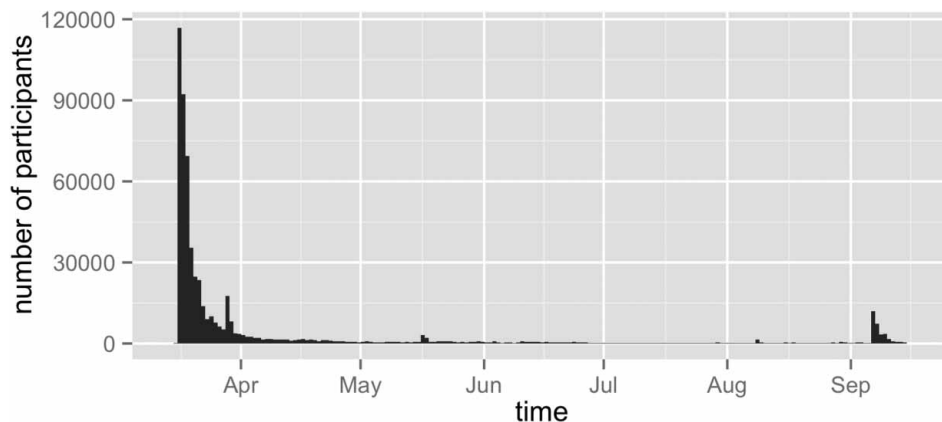


Figure 1. Number of participants over time from 16th March to 15th September 2013.

testing. In the three months before September 15th, we only gathered 7.7% of the data.

### Materials

We started with a collection of several hundred thousand words from multiple dictionary sources and corpora. From these sources, we first removed those words that were identifiable as proper nouns. We also removed most regularly inflected forms of nouns, verbs, and adjectives. While these forms are undoubtedly of great interest to psycholinguistic research in morphology, including them would have multiplied the number of items while yielding relatively little insight in vocabulary. We also removed many nonlexicalized transparent compounds and derived forms, using our best judgement to decide whether including a word would yield additional information compared to only the base form or constituent morphemes. At the start of the experiment, our list contained 52,847 words.

Nonwords were constructed with Wuggy (Keuleers & Brysbaert, 2010). This meant that for monosyllabic words one subsyllabic segment (onset, nucleus, or coda) was substituted, for disyllabic words two segments were substituted, and so on. For instance, a nonword based on an existing trisyllabic word could be constructed by replacing an entire syllable, by replacing two segments in one syllable and one in another, or by replacing

one segment in each syllable. Substituted segments were always of the same length as the original ones and were chosen to cause minimal deviations in transitional frequency. In total, 20,653 nonwords were created and selected for inclusion in the test.

The final list of 73,500 stimuli (52,847 words and 20,653 nonwords) was randomly shuffled and split into 735 sublists of 100 stimuli. This larger proportion of words (72%) than nonwords (28%) allowed us to collect more responses to words and therefore to collect more data. Moreover, since the stimuli contained many low-frequency words or words only found in dictionary sources, if we had presented an equal amount of words and nonwords the effective proportion of known words would have been much smaller than 50% for a typical participant. In fact, for a participant in the study who knew about 71% of the words the proportion of known words was close to 50%.

### Procedure

The guiding principles for the design of the experiment were (a) nothing is mandatory, (b) the test must be self-motivating, (c) it must not take more than five minutes, and (d) users should be able to do the test from any device on which they can click on a link to the test. Of course, having just 100 items relies on many subjects to have a reasonable number of observations per item. Forty observations per word requires nearly 30,000

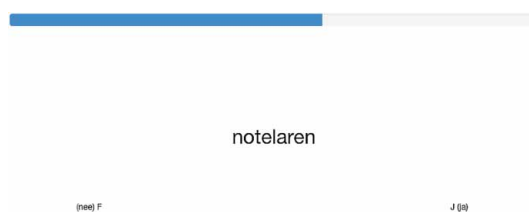
participants. Fortunately, we were able to collaborate with the Dutch Science Foundation (NWO) and two Dutch television broadcasters (NTR and VPRO) who—in the context of a television programme called *Groot Nationaal Onderzoek* (Big National Research)—have years of experience in bringing online scientific studies to a large audience. They ensured that the study received ample media attention. Participants were also able to share their score through social media channels Facebook and Twitter, or via e-mail. As a result, the initial recruitment of participants through traditional media channels could lead to the recruitment of other participants through social media.

To display the test correctly across a wide range of devices and web browsers, we made use of a light and responsive browser-display framework (“Bootstrap,” n.d.). As stated earlier, the test had tailored instructions and answering modalities for keyboard-centric devices such as desktops or laptops and touch-based devices such as smartphones and tablets. While we do not include it in our further analysis, it is interesting to know that about 27.37% of the test sessions were completed on touch devices (mobile phones and tablets). Figures 2 and 3 show the experiment screen layout for both types of devices.

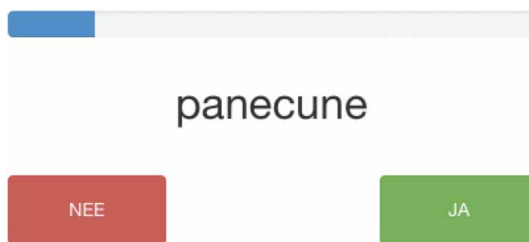
When arriving on the website, participants were greeted with a welcome message and an explanation of the goal of the test. Translated to English, the instructions were as follows:

Hello! In this test you will see 100 letter sequences, some of which are existing Dutch words and some of which are made-up nonwords. Indicate for each letter sequence whether it is a word you know or not by pressing the F or J key,  
 J: YES, I know this word  
 F: NO, I don't know this word  
 The test takes about 4 minutes and you can repeat the test as often as you want (you will get new letter sequences each time).  
 Advice! Do not say yes to words you do not know, because yes-responses to nonwords are penalized heavily!

The instructions were accompanied by an image of a keyboard with the index fingers resting on the keys corresponding to a *no* and a *yes* answer. For touch devices, the instructions were changed where necessary, and the image showed index



**Figure 2.** Experiment screen layout for keyboard devices, displaying the word “notelaren”. Left and right bottom corners contain reminders for the key mapping for No (Nee) and Yes (Ja) responses. To view this figure in colour, please visit the online version of this Journal.



**Figure 3.** Experiment screen layout for touch devices, displaying the nonword “panecune” and buttons for No (Nee) and Yes (Ja) responses. To view this figure in colour, please visit the online version of this Journal.

fingers resting on red and green buttons corresponding to no and yes answers.

After being presented with instructions, participants were asked to complete a small questionnaire with information on their age, gender, location, education level, mother tongue, number of other languages known, best other language, level of other language, and handedness.

Answering the questions was not required in order to proceed. Participants could also choose to share their geolocation. Depending on a combination of factors, including the browser, the operating system, and the device the participant used, this geolocation would give us a more detailed picture of latitude and longitude. As the geolocation data are not of immediate use for our analyses, they are not discussed further in this paper.

After going through the questionnaire screen, participants were presented with the experiment, which started with the instruction to place the fingers on the corresponding keys and to press the



space bar to start the experiment. Following this, 100 stimuli would be presented centred vertically and horizontally on the screen (see also Figures 2 and 3). There was no time-out. Participants were informed about the progress of the experiment by a blue progress bar in the top part of the screen.

After responding to all the stimuli, the score was presented to participants as a percentage estimate of their vocabulary knowledge. The score was calculated by subtracting the percentage of incorrectly accepted nonwords from the percentage of correctly recognized words. The score screen also gave participants the opportunity to examine their answers. Each of the word stimuli was linked to a definition on an external dictionary site (<http://encyclo.nl>). Participants were also able to leave free-form feedback for each of the stimuli. This option was most frequently used to report nonword stimuli for which participants suggested a definition. This enabled us to flag stimuli that did not appear in the dictionaries and wordlists that we used but were actually attested words. Most of the time they were inflected variants of existing words.

## Results

### *Variables influencing vocabulary size*

We first restricted the analysis to participants who had given complete or nearly complete profile data and who indicated that Dutch was their first language. Only the first session was included for any participant. After subtracting the percentage of false alarms (incorrectly accepted nonwords) from the percentage of hits (correctly accepted words), we obtained a score ranging from  $-100$  (all nonwords accepted, and all words rejected) to  $100$  (all nonwords rejected, and all words accepted).<sup>2</sup> For statistical analysis, these scores were transformed to logits. Potential outliers were identified using the boxplot criterion: Scores below or above the first and third quartile  $\pm 1.5$

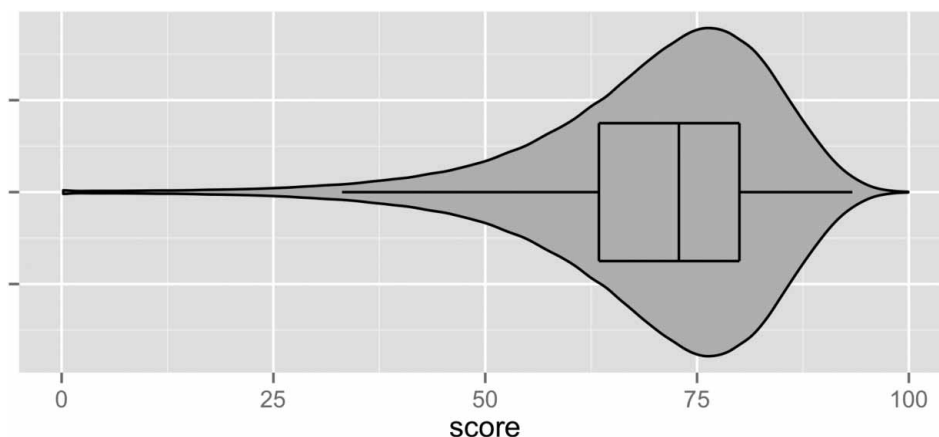
times the interquartile distance were removed. This applied to 2.32% of the data, leaving us with 303,937 completed sessions. For ease of interpretation, the logits were re-transformed to the original scale in the results reported in this paper and can be interpreted as the percentage of words known by each participant. In this analysis, we focus on the data of 278,590 participants who indicated that Dutch was their first language, who were between 12 and 72 years old, who had given complete profile data, and who had indicated growing up in Belgium (46.57%) or the Netherlands (53.43%).

Figure 4 shows the distribution of scores, illustrating that a wide range of scores was obtained and that the bulk of participants achieved a relatively high score. As a point of reference, the Dutch-speaking authors of this paper took the test multiple times and achieved scores ranging from 70% to 90%. Although it is impossible to be certain about the integrity of each individual participant, the sparseness of scores at the low end suggests that the large majority of participants took the test seriously.

For the analysis of vocabulary size, we used a linear regression model: Age was log transformed and was treated as a continuous variable; education was an ordered factor with four levels (no education/primary school, secondary school, bachelor, and master); number of foreign languages was treated as a continuous variable; L2 (best second language) was a factor with four levels (English, French, German, or other); L2 proficiency was an ordered factor with five levels (“I know a few words”, “I can have a simple conversation”, “I can read a simple book”, “I speak and read the language fluently”, “I am a native speaker”); location was a factor with two levels (Belgium, the Netherlands); gender was a factor with two levels (male, female); and, finally, handedness was also a factor with two levels (left, right).

As we were dealing with a large number of predictors and the large amount of data would lead to

<sup>2</sup>Although the current paper focuses on the accuracy measures obtained from our study, reaction times were also collected for each trial. For reference, we report the intraclass correlation coefficients (ICCs) as a measure of the reliability of the obtained reaction times. ICC(C, k), or the expected correlation for a repeat study on the average reaction times for individual words, was .999. ICC(C, 1), or the expected correlation of the average reaction times per word with those of a single new participant, was .168.



**Figure 4.** Violin plot of the distribution of scores. The violin is a mirrored density plot. Inside the violin, a boxplot of the scores is shown. Note that the critical values for the boxplot were calculated on the logit scale and then back-transformed to percentages, ensuring that all scores are bounded within the 0–100% interval.

**Table 1.** Analysis of variance table showing effects of predictors on vocabulary size

Variable	Sum of squares	df	F	p	$\eta^2$
(Intercept)	5384	1	19,030.34	.00	.0000
Age	2795	1	9880.57	.00	.1761
Education	271	3	319.62	.00	.0358
Number of foreign languages	583	1	2061.28	.00	.0073
L2	39	3	45.88	.00	.0062
L2 proficiency	62	4	54.55	.00	.0055
Location	342	1	1209.32	.00	.0041
Gender	29	1	102.54	.00	.0004
Handedness	0	1	0.09	.77	.0000
Age $\times$ L2 proficiency	35	4	31.13	.00	.0004
Age $\times$ Location	76	1	268.45	.00	.0010
Education $\times$ L2	47	9	18.51	.00	.0006
Education $\times$ L2 Proficiency	22	12	6.45	.00	.0003
L2 $\times$ L2 proficiency	31	12	9.16	.00	.0004
Residuals	78,797	278,536			

*Note:* Test score. The final column indicates the effect size ( $\eta^2$ ) for each term. L2 = second language.

significant effects with very small effect sizes, we used a pragmatic criterion to limit the terms in the model. First, we restricted the model to main effects and two-way and three-way interactions. Then we removed all three-way interactions that explained less than 0.02% of variance. None of the three-way interactions survived this step. We refitted the model with the remaining terms and applied the same criterion to the two-way

interactions. Here, five interactions remained: Age  $\times$  L2 Proficiency, Age  $\times$  Location, Education  $\times$  L2, Education  $\times$  L2 Proficiency, and L2  $\times$  L2 Proficiency. The final model consisted of all main effects and the remaining two-way interactions. Table 1 illustrates the results obtained for the terms in the final model. Except for handedness, all the terms explained a significant part of the variance.

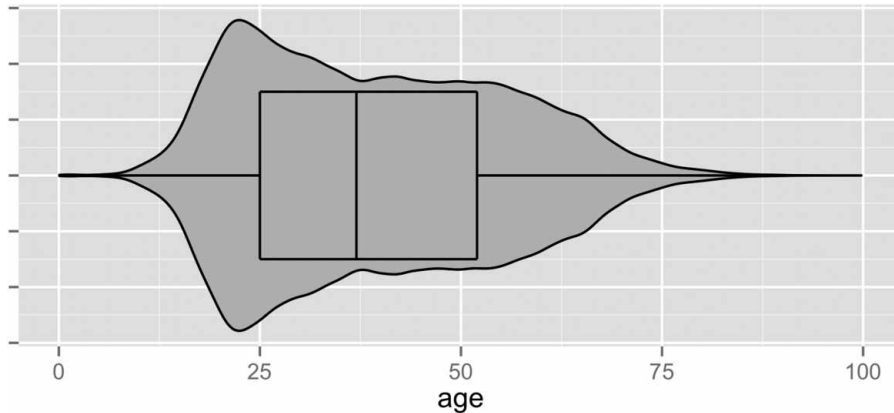


Figure 5. Distribution of age in the experiment. The outside shape is a mirrored density contour. Inside, a boxplot of the distribution is shown.

*Age.* Before discussing the effects of age, it is worthwhile to look at the distribution of the age of our participants. In typical psycholinguistic research, participants' age is rather homogeneous. Moreover, as the exchange of course credits for participation and proximity to researchers' laboratories presumably account for the bulk of participants in psychological experiments, a large part of psycholinguistic knowledge is derived from homogeneous groups of university students. In addition, it is often tacitly assumed that having a homogeneous group of participants leads to less variance in behavioural measures and therefore to greater statistical power. Figure 5 shows the distribution of age. While we cannot be sure that all participants report their age correctly, the distribution reflects participation by a very broad range of the population. The median age of participants was 37, a quarter of the participants was younger than 25, and a quarter was older than 52, showing that, at least in these circumstances, online experiments with free participation can reach a large and diversely aged number of participants.

With over 17% of variance in scores explained, age is by far the most important predictor of vocabulary size in our test ( $\eta^2 = .1761$ ). The evolution of score with age is illustrated in Figure 6<sup>3</sup> and is

consistent with the interpretation that vocabulary size increases with age (McCabe et al., 2010; Park et al., 2002). The large number of words and participants in the study, however, allows us to get a more fine-grained picture of the relation. At the age of 13, it can be estimated that the participants know about 55% of the words (about 29,000 words in our test), while participants over 70 know nearly 80% of the words in our test, or an increase of 14,000 words over 57 years.

*Education.* We found a clear effect of education, contributing to about 3.5% of the variance in individual scores ( $\eta^2 = .0358$ ). Figure 7 shows that each additional level of education accounted for a substantial increase in vocabulary size. The largest increase in vocabulary occurred between secondary education and bachelor education (3.73% or about 2000 words), followed by an increase from bachelor education to master education (2.88% or about 1500 words).

*Number of foreign languages.* Of the participants discussed here—those who indicated that Dutch was their best language—the majority responded that they knew two (34.2%) or three (38.62%) foreign languages. A smaller group indicated that they

<sup>3</sup>The plotted values are conditioned on the weighted mean for the discrete variables and on the true mean for the continuous variables. As such, the effect plots can be interpreted as if they resulted from a completely balanced design for all discrete variables (i.e., the same number of Dutch and Belgian participants, the same number of female and male participants, etc.)

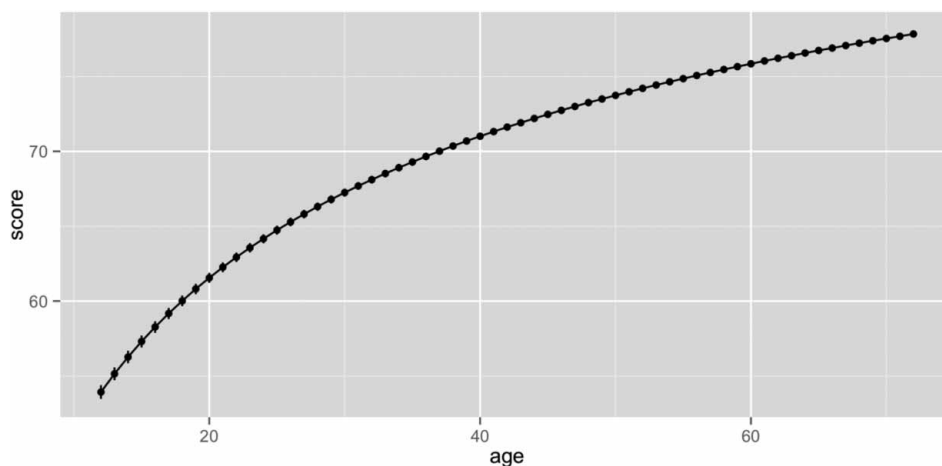


Figure 6. *Effect plot of age on vocabulary size.*

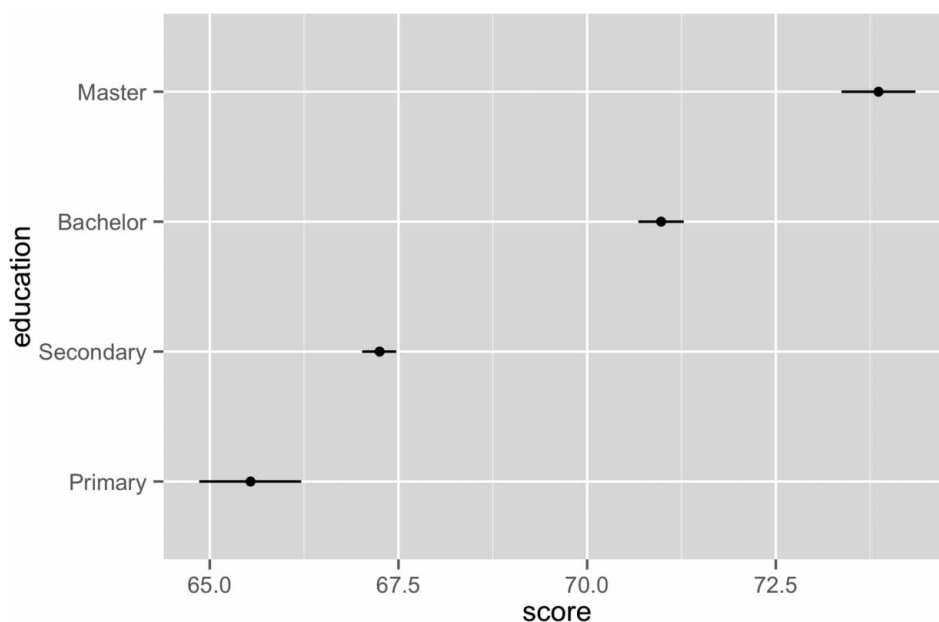


Figure 7. *Effect plot of education on vocabulary size.*

knew one (9.82%), four (12.28%), or five (3.12%) other languages. Response options indicating six languages or more were each chosen by less than 1% of participants. Less than 1% of participants indicated they knew no languages besides Dutch. While this survey question does not take into

account specific proficiency in other languages, these results indicate that, among the Dutch-speaking participants of our test, multilingualism is the norm. This is not surprising: Foreign language classes are ubiquitous from secondary education level onwards in Belgium and the

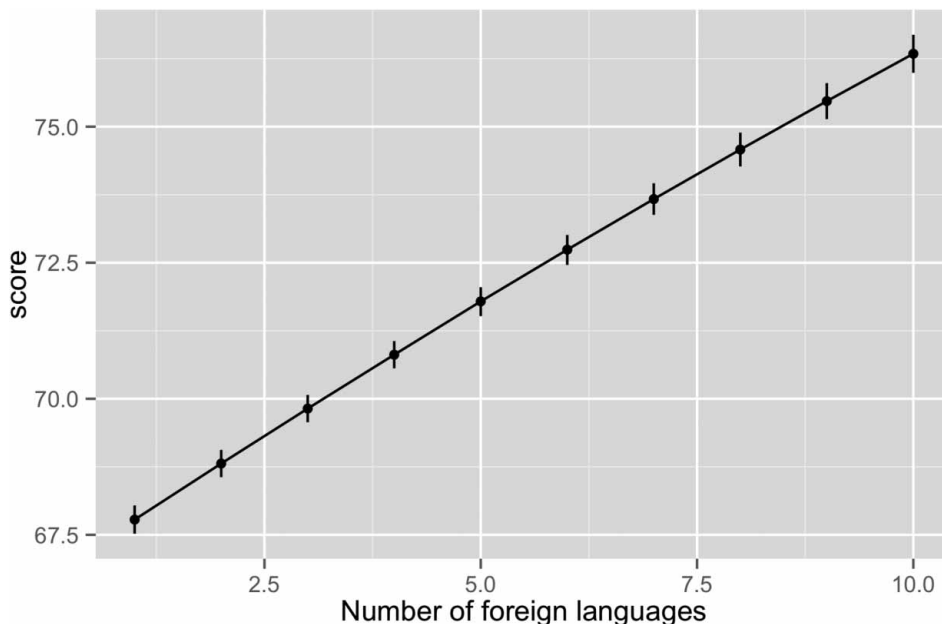


Figure 8. Effect plot of number of foreign languages on vocabulary size.

Netherlands and it can be assumed that participants with a university degree have an intermediate level of English, French, or German.

The number of foreign languages known accounted for under 1% of the variance in vocabulary size ( $\eta^2 = .0073$ ). Figure 8 shows that the estimated effect of knowing an extra language translates to about 500 words per extra language, with slowly diminishing returns.

The fact that scores keep increasing up to 10 languages is surprising, as we first assumed that choosing a very high number would be done jestingly by participants who were not very serious about the test and thus would also be expected to score low. However, we also inquired about the level of participants' best *other* language, with response options ranging from "I know a few words" to "I speak the language fluently". According to the criterion "I know a few words", participants responding five or six would not be very exceptional (e.g., for the authors of this paper, values would range from four to nine). Most participants who reported they knew at least one other language reported English as their best

other language (75.06%). After that, French (11.01%) and German (5.28%) were the most reported choices. Spanish, Frisian, and Italian were reported by less than 1% of the participants. No other language represented more than 0.1% of the answers to this question. About 5.76% of participants selected Dutch as their best second language although they had also said it was their first language. This supports the idea that some participants misinterpreted questions about *other* languages as questions about *all* languages.

*L2.* L2 explained less than 1% of the variance in score ( $\eta^2 = .0062$ ). Figure 9 shows that having English and French as a second language was the most beneficial for L1 vocabulary size. German came third, and other languages came last.

*L2 proficiency.* L2 proficiency explained about 0.5% of the variance in vocabulary size ( $\eta^2 = .0055$ ). Figure 10 shows that increased L2 proficiency had a positive effect on L1 vocabulary size, but that at the second-to-highest level of proficiency no further gains were made. Participants who

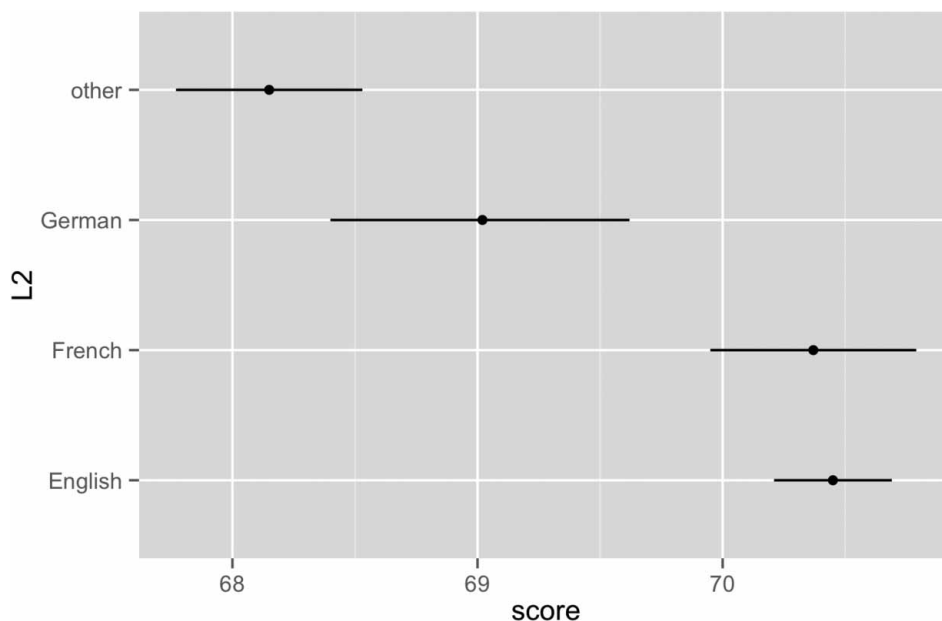


Figure 9. Effect plot showing the relation between second language (L2) and vocabulary size.

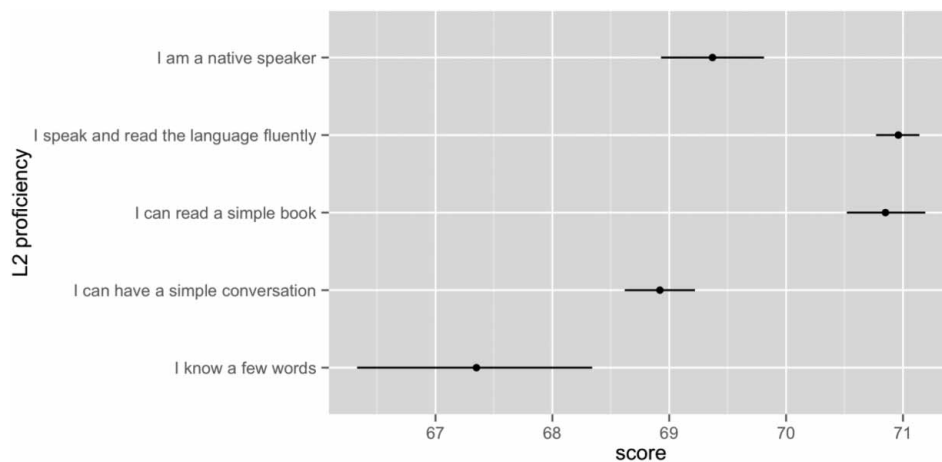


Figure 10. Effect plot showing the relation between second language (L2) proficiency and vocabulary size.

indicated native L2 proficiency had a smaller L1 vocabulary size than participants with an intermediate or high but not native level of L2 proficiency.

*Location.* Overall, Dutch participants scored about 1.5% higher than Belgian participants, accounting

for less than 0.5% of variance in our data ( $\eta^2 = .0055$ ).

*Gender and handedness.* Our analysis shows that male participants score on average about 0.5% higher than female participants, with a very small associated effect size ( $\eta^2 = .0004$ ).



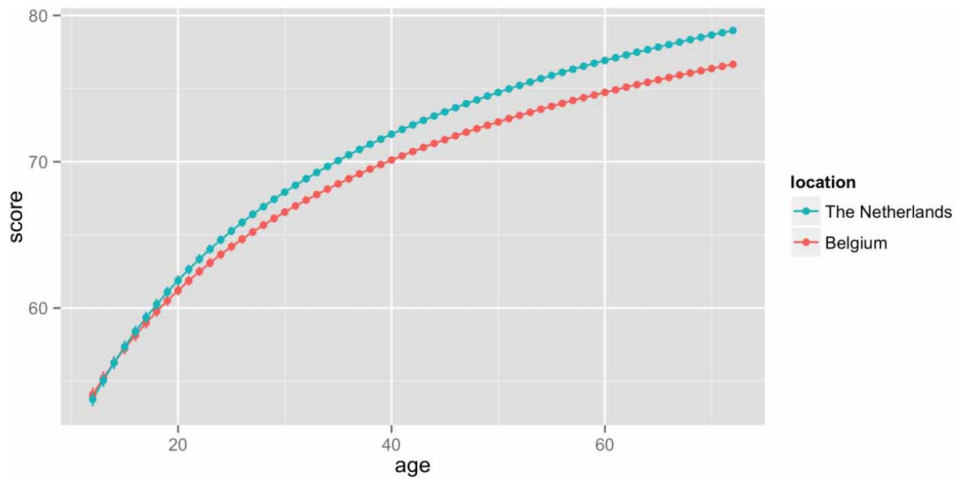


Figure 11. Interaction effect of age and location on vocabulary size (effect plot).

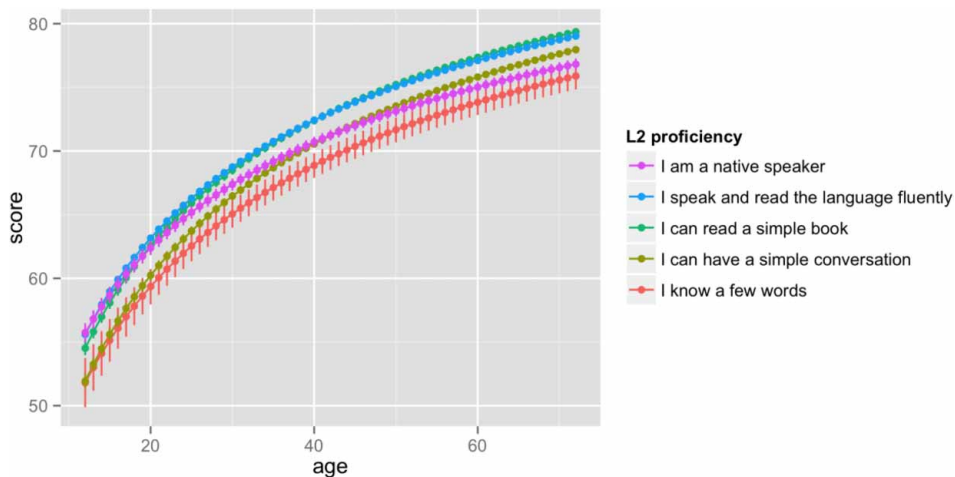


Figure 12. Interaction effect of age and second language (L2) proficiency on vocabulary size (effect plot).

Handedness did not explain any variance in vocabulary size.

*Age × Location.* The interaction effect of age and location (Figure 11) shows that where there is virtually no difference in score for the youngest Dutch speakers in Belgium and the Netherlands, older participants from the Netherlands score better than older Belgian participants, reaching a

gap of 4% for participants older than 70. The unique variance in vocabulary size accounted for by this interaction was about 0.1% ( $\eta^2 = .0010$ ).

*Age × L2 proficiency.* Figure 12 shows the interaction of age and L2 proficiency. The general trend is that differences due to proficiency get smaller as age increases. Of particular interest here are the participants who indicate that they

have native L2 proficiency. The rather strong advantage for younger participants having native L2 proficiency compared to their peers having only basic proficiency in their L2 is almost entirely lost for older participants. The effect size for this interaction was very small ( $\eta^2 = .0004$ ).

*Education  $\times$  L2 and Education  $\times$  L2 Proficiency.* Education interacts both with L2 and with L2 proficiency. Figure 13 shows that the effect of having a specific L2 gets less pronounced with increasing education. Figure 14 shows that L2 proficiency differentially affects scores by education level: The lower the level of education, the larger the advantage of knowing more than a few words in a second language. The effect sizes for both the interaction between education and L2 ( $\eta^2 = .0006$ ) and education and L2 proficiency were very small ( $\eta^2 = .0003$ ), both explaining less than 0.1% of the unique variance in scores.

*L2  $\times$  L2 Proficiency.* Finally, Figure 15 shows that L2 interacts with L2 proficiency in a relatively

straightforward manner. At the very basic level of proficiency, there is almost no difference in the effect of various second languages on vocabulary size. As proficiency grows, the differences become more pronounced and reach the same order as is seen in the main effect for L2: English and French have a larger effect on vocabulary size than German, which in turn has a larger effect on vocabulary size than other languages. Like the other interaction effects discussed, the effect size was very small ( $\eta^2 = .0004$ ).

### Prevalence

To build a measure of prevalence, we used the data of 190,771 participants (up to their third session) who indicated that they were living in Belgium, giving us about 250 observations per word. The reason for using only Belgian participants was that the data of the Dutch Lexicon Project, which was carried out at Ghent university, also came from Belgian participants. Rather than looking at accuracy measures or scores, we fitted an explanatory Rasch model (Doran, Bates, Bliese, &

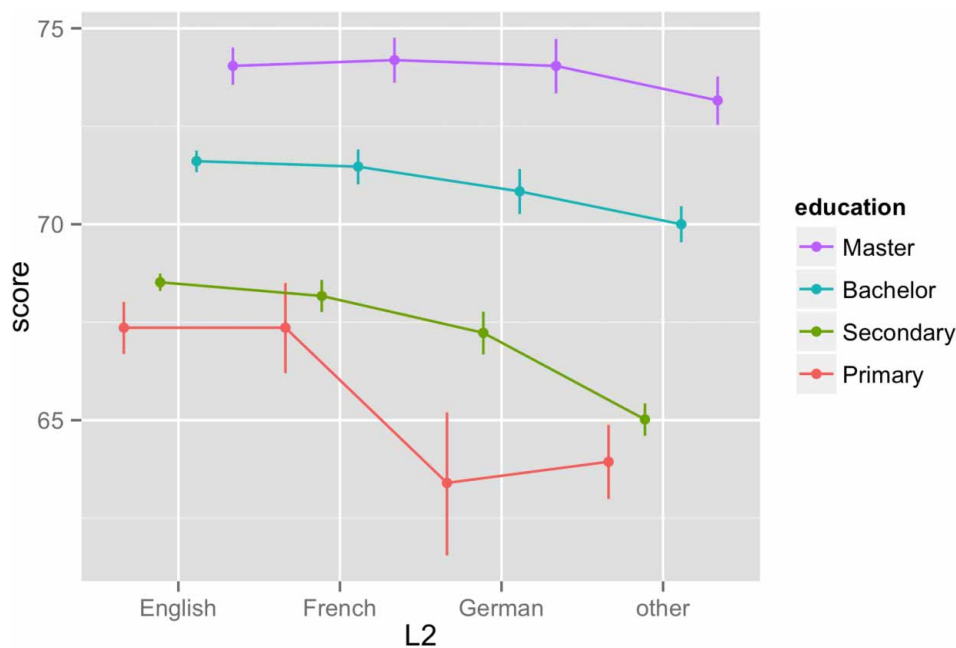


Figure 13. Interaction effect of education and second language (L2) on vocabulary size (effect plot).

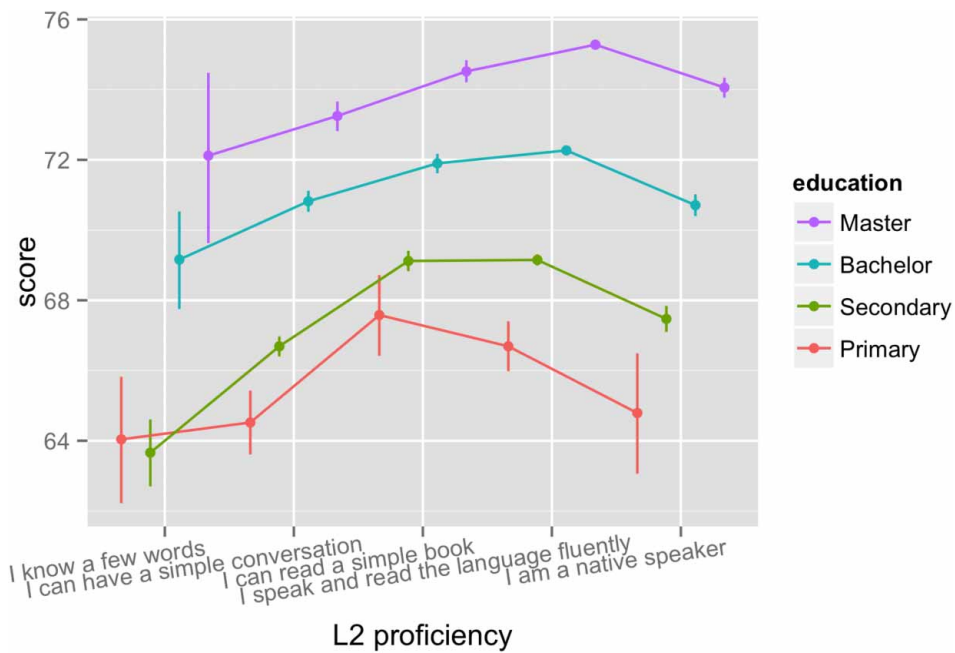


Figure 14. Interaction effect of education and second language (L2) proficiency on vocabulary size (effect plot).

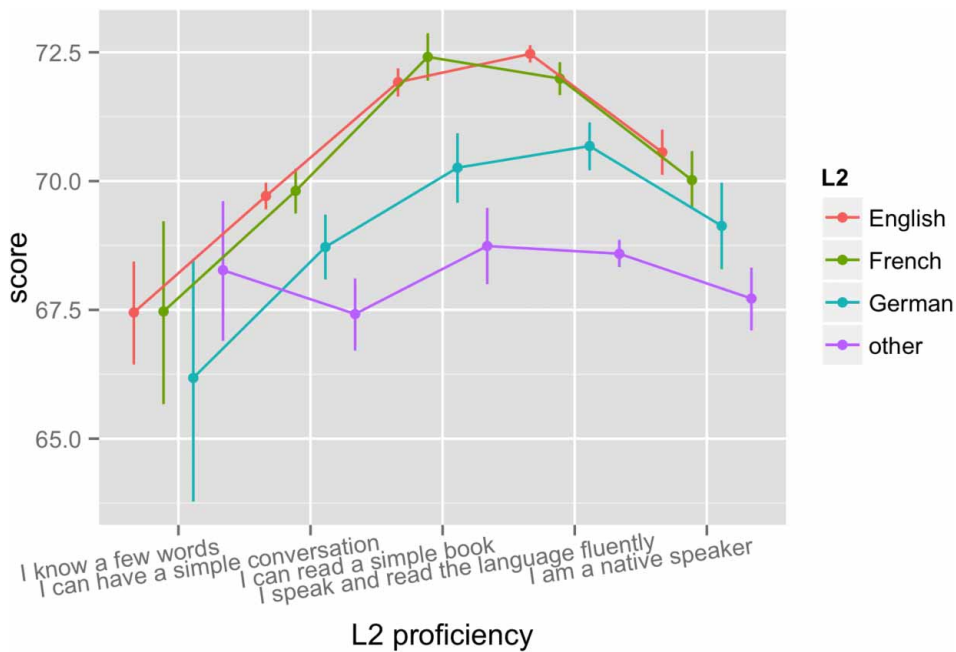
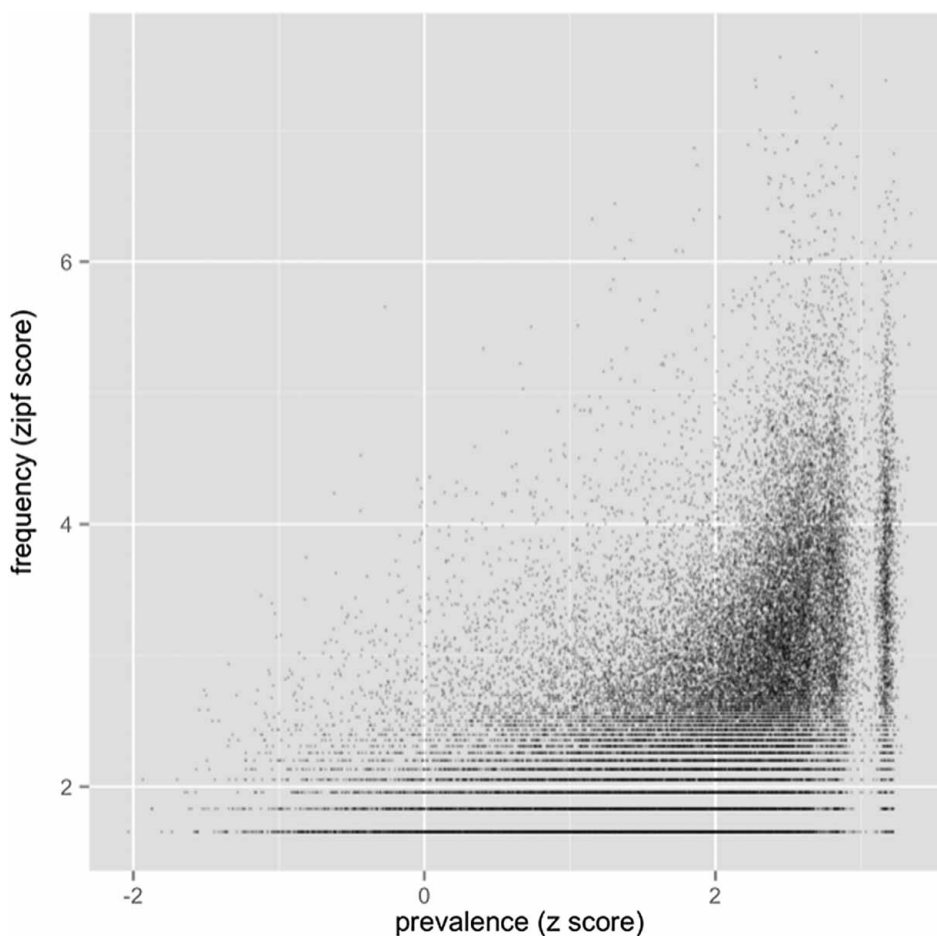


Figure 15. Interaction effect of second language (L2) and L2 proficiency on vocabulary size (effect plot).

Dowling, 2007). The standard Rasch model would include a global intercept and a random intercept per participant and per item, but to account for differences in word knowledge and decision bias we added a fixed effect of lexicality and a random effect of lexicality for participants. We considered the predicted item difficulty resulting from this model as an operationalization of its prevalence. Raw item difficulty is a  $z$ -score that can be easily transformed to proportion known using the cumulative normal distribution function. However, in our analysis we observed better fits using the  $z$ -scores.

The relationship between the frequency counts from SUBTLEX-NL (Keuleers, Brysbaert, & New, 2010) and prevalence is illustrated in Figure 16. The dark lines at the bottom half of the plot indicate words with singularly low frequencies over a large range of prevalence. The elongated cluster at the right side of the plot shows words with nearly full prevalence over large frequency ranges.

In order to examine the contribution of Dutch corpus word frequency (SUBTLEX-NL) and word prevalence on reaction times, we analysed the data from the 7885 items in the



**Figure 16.** *The relationship between frequency and prevalence. Word frequency is displayed as Zipf-score (Van Heuven et al., 2014). Higher  $z$ -scores indicate more prevalent words.*

Dutch Lexicon Project (DLP; Keuleers, Diependaele, et al., 2010) for which both frequency and prevalence were available. In single variable analyses, log word frequency explained about 36.13% of the variance in average standardized reaction times, and prevalence explained about 33.03% of the variance in reaction times. The correlation between prevalence and frequency was low (see Table 2), showing that the two measures are not simple mathematical transformations.

This was also made clear when both measures were considered in the same analysis. Frequency and prevalence jointly explained 51.37% of the variance in reaction times. The low correlation between the two variables suggested high unique contributions of each of them: Eta squared was .2231 for frequency and .1854 for prevalence.

Figure 17 further illustrates how the measures are complementary. The left panel shows that a linear model with frequency as a predictor underestimates slow reaction times. As in Figure 16, the dots forming dark lines at the top of the panel show words with singularly low frequency, which are spread over a wide range of reaction times. The right panel shows that a linear model with prevalence as a predictor overestimates fast reaction times. The dots forming a dark cluster at the bottom of the panel show words with a near-complete prevalence, which are predicted to lie in a small range of reaction times. The central panel shows that a linear model including both frequency and prevalence does not suffer from problems of over- and underestimation. Instead, in the range in which frequency underestimates reaction time, prevalence is a good

Table 2. Correlations between main predictors of lexical decision reaction time in the Dutch Lexicon Project

Measure	Frequency	Prevalence	OLD20	Length	Contextual diversity
Frequency	1.00	.35	-.34	-.37	.98
Prevalence	.35	1.00	.00	.07	.36
OLD20	-.34	.00	1.00	.74	-.34
Length	-.37	.07	.74	1.00	-.35
Contextual diversity	.98	.36	-.34	-.35	1.00

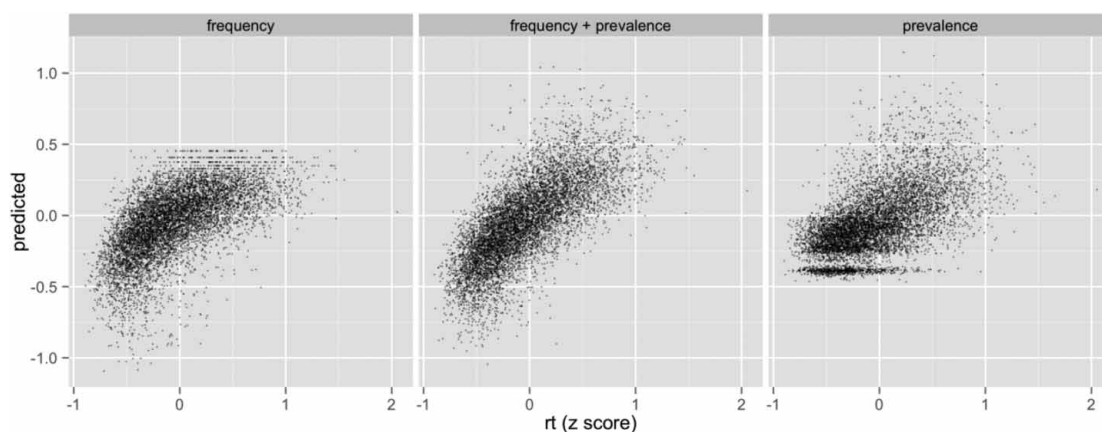


Figure 17. Predicted (y-axis) versus observed (x-axis) standardized lexical decision reaction times in the Dutch Lexicon project for three linear models. Each panel indicates the predictors used in the model. Left: frequency. Middle: frequency + prevalence. Right: prevalence. Lower numbers indicate faster reaction times.

**Table 3.** Regression table for linear model fitted on lexical decision reaction times of 7885 items in the Dutch Lexicon Project

Variable	Estimate	SE	t	$Pr(> t )$	$\eta^2$
(Intercept)	1.0268	0.0203	50.5173	.00	NA
Frequency	-0.1776	0.0040	-44.4968	.00	.1573
Prevalence	-0.3463	0.0067	-51.4988	.00	.2108
OLD20	-0.0123	0.0079	-1.5550	.12	.0002
Length	0.0294	0.0029	10.1843	.00	.0082

predictor, and in the range where prevalence overestimates reaction time, frequency is a good predictor.

To further investigate the contributions of frequency and prevalence, we fitted a model with other typical predictors of lexical decision reaction time—namely, word length and orthographic neighbourhood density as measured by OLD20 (Yarkoni, Balota, & Yap, 2008). This increased the total explained variance to 52.35%. Table 3 shows that that estimates for the coefficients of both frequency and prevalence are negative. An increase in any of the two variables results in faster reaction times. The addition of length and neighbourhood density, which are more correlated with frequency than with prevalence, leads to prevalence becoming the most important predictor of reaction times in this model. The remaining effect of length is surprisingly small, and the effect of OLD20, which is correlated with length, becomes insignificant.

Contextual diversity (Adelman, Brown, & Quesada, 2006) was also considered, but due to extremely high collinearity with frequency (see Table 2), we included only frequency in the model. Fitting a model with contextual diversity instead of frequency did not change the pattern of results, except for the effect of OLD20, which remained significant.

For the same reason we did not include the quadratic trend of frequency in the model, although it is known to explain additional variance

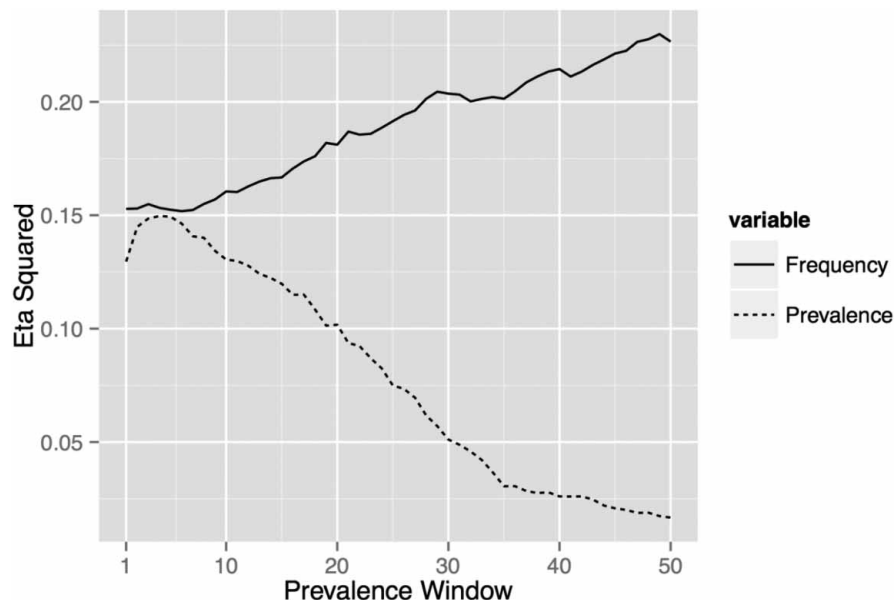
in lexical decision reaction times (e.g., Balota et al., 2004). This would predictably lead to a much smaller amount of uniquely accounted variance for frequency, which would make a comparison unfair.<sup>4</sup>

Including a term for the interaction between frequency and prevalence in the model resulted in a very small effect size for the interaction term ( $\eta^2 = .00077$ ,  $p = .00181$ ) and a very small increase in fit (an increase in  $R^2$  from .5235 to .5241), indicating that the effects of frequency and prevalence are largely additive.

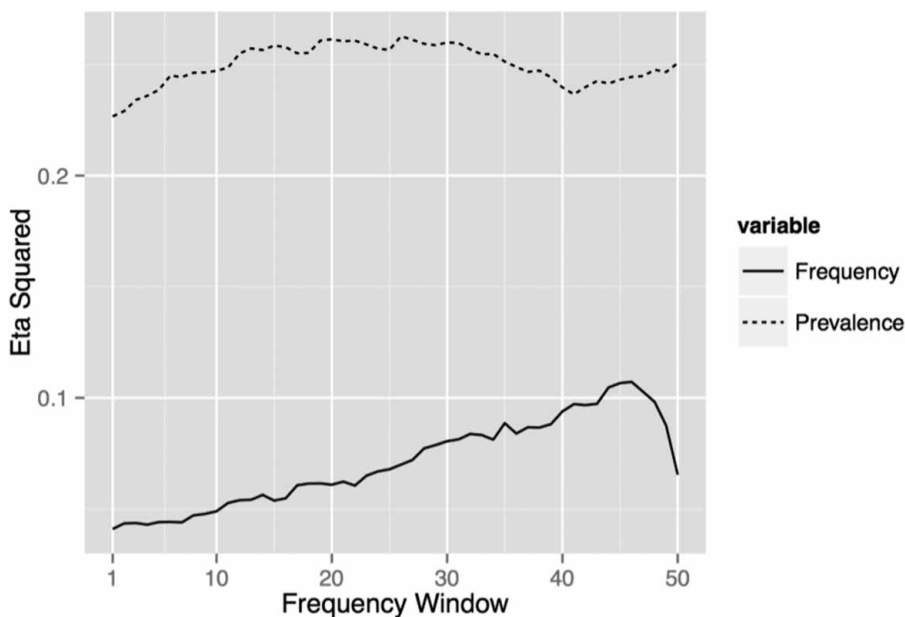
Finally, we repeated the analysis with the predictors shown in Table 3 in different ranges of the frequency variable and in different ranges of the prevalence variable. We used a moving window of half the range of the variable, sliding from the bottom to the top of the range in 50 increments. In other words, the first window covered the range from percentile 0 to percentile 50 of the distribution while the last window covered the range from percentile 50 to percentile 100 of the distribution. Percentiles were computed on the scale of the variables as they were used in the analysis. Figures 18 and 19 show the results of these analyses. In both analyses the effect size of the predictors in the ranges delimited by their own percentiles was smaller than the effect size of the unconstrained predictor. However, this was much more so the case for frequency than for prevalence. Figure 18 shows that the effect size for prevalence starts on almost equal footing with that of frequency in the lowest half of the prevalence

<sup>4</sup>For reference, the correlation between log frequency and its quadratic trend was .96. Adding the quadratic trend for log frequency to the model lead to an increase in  $R^2$  from .5235 to .5457. Eta squared for the quadratic trend was .0181. Eta squared for log frequency decreased from .1573 to .0652, and eta squared for prevalence increased from .2107 to .2194.





**Figure 18.** Effect sizes (*eta squared*) for frequency and prevalence on Dutch Lexicon Project (DLP) lexical decision reaction times in different ranges of the prevalence variable. The x-axis represents a moving window over half the range of the variable, sliding from the bottom to the top of the range in 50 increments. The first window covers the range from percentile 0 to percentile 50 of the prevalence distribution. The last window covers the range from percentile 50 to percentile 100 of the distribution. Each window contains between 3957 and 3975 observations.



**Figure 19.** Effect sizes (*eta squared*) for frequency and prevalence on Dutch Lexicon Project (DLP) lexical decision reaction times in different ranges of the frequency variable. The x-axis represents a moving window over half the range of the variable, sliding from the bottom to the top of the range in 50 increments. The first window covers the range from percentile 0 to percentile 50 of the frequency distribution. The last window covers the range from percentile 50 to percentile 100 of the distribution. Each window contains between 3958 and 4052 observations.

distribution and then gradually declines, with the effect size of frequency showing a commensurate increase. This shows that frequency becomes more important as prevalence increases.

Figure 19, on the other hand, shows that the effect size of prevalence is quite strong throughout the different ranges of the frequency distribution. The important observation here is that the effect size of frequency is very small in the lower half of the distribution and then shows a steady increase. The steep decline in the last frequency windows is probably due to high-frequency words with slightly longer reaction times (Keuleers, Diependaele, et al., 2010), which deviate from the linear trend for frequency. Including a quadratic trend for frequency in a model will tend to account for these observations, but, as explained above, the focus in the current analysis is on uniquely explained variance, and adding a highly collinear variable will mask the effect size for frequency.

## Discussion

We first focus on the effects of age and multilingualism on vocabulary size. Then, we discuss the relation between word prevalence and word frequency and the effects of word prevalence on word processing times.

### *Age*

Our results show that age is by far the most important variable in predicting vocabulary size. From the point of view that every day lived represents an opportunity for acquisition of vocabulary and that existing vocabulary is not forgotten, a steadily increasing vocabulary is not surprising. But how do we explain that, although vocabulary keeps increasing with age, the vocabulary growth rate decreases? A first naive interpretation would be that the capacity to learn new words also decreases with age. However, the relationship between age and vocabulary size that we found in our study is strikingly similar to the vocabulary growth curve predicted by Herdan's law, which we discussed in the introduction in connection with multilingualism. Assuming that Herdan's law is a good

explanation for the relation between age and vocabulary size implies that the slowing of vocabulary growth with age is not linked to a decreasing capacity to learn new words but is the result of the ever smaller probability of encountering an unlearned word in the environment. After the fact it is perhaps unsurprising that the vocabulary growth curve of people is similar to the vocabulary growth observed in text corpora. Ramscar, Hendrix, Shaoul, Milin, and Baayen (2014) proposed a theoretical model of the evolution of vocabulary size over age in which they show that simulated speakers display a vocabulary growth that is remarkably similar to the one we find empirically. However, we know of no earlier empirical demonstration of this for humans.

An alternative hypothesis for the effect of age on vocabulary is connected with the speed-accuracy trade-off, or, in other words, the observation that faster responses lead to more errors. Reaction time is often found to increase with age, and so it is possible that the slower reaction times associated with increasing age would lead to fewer errors. A cursory analysis of our data shows that while there is a small increase of reaction time with age, the average accuracy per session in our test increases rather than decreases with speed, and this increase is not different for different age groups. Although the explanation that higher scores are caused by a speed-accuracy trade-off could be explored more in depth, we think that it would at most account for a small part of the results.

A related alternative explanation comes from research suggesting that conscientiousness increases with age (Srivastava, John, Gosling, & Potter, 2003). This could imply that older participants guess less and may therefore score higher. A brief analysis of our data shows that the number of false alarms (word responses to non-words) indeed decreases with age. However, the question is whether this influences the estimated vocabulary sizes for more versus less conscientious participants. The penalty for guessing in our calculation of participants' score only corrects for the estimated increase in score due to correctly guessed responses. Moreover, we confirmed that the raw accuracies, without guessing penalty,

show a similar increase with age as the corrected scores do. Still, if participants are partly answering randomly rather than guessing, their estimated vocabulary will indeed be too low. We believe this explanation could be explored further, but that it would only explain a small part of the association between age and vocabulary size. In addition, it is worth noting that increased conscientiousness does not necessarily lead to lower scores, as more conscientious participants are presumably less risk-prone, giving fewer false alarms and fewer correct responses.

In our opinion, the most promising explanatory framework comes from considering the lexicon as a dynamic system.<sup>5</sup> In this view, the dynamics of the lexical decision task necessarily change with age or, more precisely, with an increasing vocabulary (Ramsar, Hendrix, Love, & Baayen, 2013). For a participant who does not know a particular word, a nonword response to that item is a correct response. As a consequence, the proportion of words to nonwords changes from this perspective, and presumably so do associated response biases, which could explain the decreased number of false alarms for older participants. In addition, regarding the lexicon as a dynamic system raises some further possible explanations. In models of categorization such as SUSTAIN (Love, Medin, & Gureckis, 2004), an increase in the number of lexical discriminations in the model results in each lexical category developing a more discriminated and less general underlying representation. This implies that as the lexicon grows, it becomes easier to integrate new words. Not only would vocabulary increase with age as predicted by Herdan's law, but the acquisition of new vocabulary should also become easier. We believe this view complements our interpretation.

#### *Age and location*

Our results also show a striking interaction of age and location: Participants from the Netherlands have a larger vocabulary size than participants from Belgium, and this advantage increases with age (it is nearly absent for the youngest

participants). The explanation that seems most likely to us has its roots in language policy. In the Netherlands, Dutch is the dominant language, while in Belgium the linguistic situation is far more complex. De Caluwé (2012) explains that when Belgium was created in 1830, French was the only official language. The majority of the population in Flanders, however, spoke dialects related to Dutch. Because Belgium did not have a standardization process for these dialects, the standard language spoken in the Netherlands was adopted as the official standard in the early twentieth century. Language propaganda encouraging the use of standard Dutch started only in the 1960s and 1970s, at which point few speakers mastered the standard language. An increasing use of standard Dutch in Belgium from that time on would certainly explain why differences in vocabulary score with Dutch speakers of the Netherlands decrease with age. Additionally, due to the later introduction of standard Dutch in Belgium and perhaps also due to the large difference in number of speakers, Dutch as spoken in the Netherlands has a larger influence on what is codified as standard language. De Caluwé (2012) notes that until 2005 the authoritative Dutch dictionary Van Dale (Boon & Geeraerts, 2005) marked variants spoken primarily in Belgium explicitly as "Dutch in Belgium" while variants spoken primarily in the Netherlands were labelled "standard use".

#### *Multilingualism*

Our results can contribute to the knowledge about multilingual language processing. As a point of departure we took a core assumption of the weaker links theory in the bilingualism literature, namely that "bilinguals divide frequency-of-use between two languages" (Gollan et al., 2008, p. 787). Carrying this idea from usage to exposure implies that the amount of word tokens one is exposed to is independent of the number of languages one is exposed to. Since the number of new word types one discovers depends on the number of tokens one is exposed to, exposure to

<sup>5</sup>We thank Michael Ramsar for suggesting this interpretation.

other languages should lead to decreased L1 vocabulary.

Despite this, knowledge of another language could indirectly contribute to L1 vocabulary if it shares vocabulary with L1. We saw that the shared vocabulary between Dutch and English, French, or German is remarkably large (Schepens et al., 2012). We also suggested a consequence of Herdan's law could further reduce the effects of decreased exposure to L1. Because the vocabulary growth rate is larger in L2 than in L1, being exposed to an L2 leads to a relatively small vocabulary loss in L1 and a relatively large vocabulary gain in L2. The larger the differences in proficiency between L1 and L2, the larger this effect will be. The small decrease in L1 vocabulary because of exposure to other languages may be outweighed by the indirect acquisition of L1 vocabulary through L2.

Our results show that indirect vocabulary acquisition greatly outweighs vocabulary loss through decreased L1 exposure. First, vocabulary size increases with number of languages that a participant reports to know. Second, participants who reported that their best L2 was English, French, or German (languages that have the largest vocabulary in common with Dutch) had a larger Dutch vocabulary than participants who reported to have another best L2. However, although German has the largest shared vocabulary with Dutch, participants who reported French and English as their L2 had a higher Dutch vocabulary size than participants with German as their L2. A tentative explanation is that German and Dutch are so closely related that exposure to German vocabulary rarely leads to *new* vocabulary in Dutch. In contrast, Dutch borrows a lot of specialized vocabulary, much of it related to science, from English and French. It is less likely that participants already know these words in Dutch.

As a further indication that Herdan's law is a likely account for human vocabulary growth, our results show that L1 vocabulary size does not monotonously increase with L2 proficiency. Native L2 proficiency is associated with a smaller L1 vocabulary size than high but non-native L2 proficiency. This is highly indicative that exposure plays a role.

Although participants with native L2 proficiency probably have more vocabulary in L2, their greater exposure to L2 implies a smaller L1 exposure that also leads to a smaller L1 vocabulary size relative to participants with high, but non-native, proficiency. Note, however, that participants with a native L2 still have a higher L1 vocabulary size than participants with low level of L2.

The interaction between L2 and L2 proficiency shows that the above pattern is less clear for L2s that have a smaller shared vocabulary with Dutch. In particular, participants who indicated that they had the lowest proficiency level in an L2 other than English, French, or German had a relatively high L1 vocabulary size compared to the other proficiency levels. It should be noted that this is an atypical situation, as this implies an even lower level of proficiency in the languages they are most likely exposed to through education or media. This result lacks a clear explanation.

The interaction effect of age with L2 proficiency seems to show itself in different vocabulary growth rates for different levels of proficiency. This is most striking for participants with native L2 proficiency. Younger participants with native L2 proficiency start out with an L1 vocabulary size at the high end of the range for their age whereas older participants with native L2 proficiency score near the lower end for their age. This suggests that L1 vocabulary growth is slower for participants who are balanced in their L1 and L2 proficiency than for nonbalanced participants. Since native L2 is acquired earlier than non-native L2, we could also be witnessing the effect of decreased L1 exposure earlier. In contrast, the vocabulary growth rate looks strongest for participants who indicated the next-to-lowest level of proficiency in their L2. The gap with participants with a higher level of L2 proficiency is considerably narrower for older participants. It seems that that the large initial advantage of L2 exposure on L1 vocabulary size becomes less pronounced in time, probably through increased L1 exposure, although this is not seen for the participants who indicated the lowest level of L2 proficiency, who always score at the low end of the range for their age.

L2 and L2 proficiency also interact with education. The higher the level of education, the smaller the advantage of knowing a specific second language and being moderately proficient in a second language. A tentative explanation for this result is that the shared vocabulary that can be acquired via exposure to specific other languages can also be acquired through education, making the effects of L2 on vocabulary size less pronounced for higher education levels.

There remains a possibility that the results we are seeing are not, or not only, a consequence of the interplay of the factors described above. Two additional explanations may be considered. First, the assumption that exposure is independent of the number of languages could be wrong. In that case, to explain the results, participants who are exposed to foreign languages would be exposed to more tokens in L1 than participants who are only exposed to L1. Second, the language richness in L1 (the ratio of types per tokens) could be larger for participants who are exposed to more languages, thereby requiring less exposure to reach a larger vocabulary. Unfortunately, the current study does not allow us to address these other explanations.

### *Prevalence*

In this paper, we introduced *word prevalence* as a measure of word occurrence. Prevalence is based on the idea that a word that is known by more language users also has more potential speakers. Therefore, the prevalence of a word among language users can be taken as an estimate of its occurrence. We showed that prevalence and word frequency, as calculated from corpora, are complementary measures. They have a low correlation, and each seems to be a good estimator of occurrence, but possibly in different ranges. Prevalence is a more fine grained measure for rarer words, whereas word frequency more precisely measures differences between words that are widely known.

Through the online vocabulary test, we were able to collect prevalence norms for over 50,000 Dutch words. Using these norms to predict lexical decision reaction times resulted in a large increase in explained variance, which was not accounted for by word frequency, word length, or

neighbourhood density. In our analysis, prevalence was the most important predictor of visual word recognition times, suggesting that the measure should be included in any analysis where corpus word frequency is considered.

A first possible objection to the prevalence measure is that the relatively low correlation between prevalence and corpus word frequency could be an artefact of the scale on which prevalence is measured and that a transformation of prevalence would make it more correlated with corpus word frequency. As explained in the Results section, we already compared some transformations and noted improved fits using the raw item difficulties ( $z$ -scores) obtained from the Rasch model instead of difficulty scores ranging from 0 to 1, which are the result of transforming the  $z$ -scores with the cumulative normal distribution function. We should note that using  $z$ -scores, which makes prevalence less right-skewed, not only improved fits on reaction times but also increased the correlation between frequency and prevalence from .21 to .35. To further examine the possibility of a better transformation, we estimated the optimal Box–Cox transformation for prevalence (starting from item difficulties, as these are positive) in relation to corpus word frequency. The best transformed prevalence measure had a correlation with word frequency of .3522 (instead of .3470) and had a correlation of .97 with the  $z$ -scores we reported in our analysis. Re-fitting the model shown in Table 3 using the Box–Cox transformed prevalence measure increased explained variance by 1.88% and resulted in an increase of 2.7% in the unique contribution of prevalence and a decrease of 1.17% in the unique contribution of corpus word frequency. The very high correlation between our original measure with the one obtained using estimation of the optimal Box–Cox transform shows that we were using a near-optimal measure in our analysis. In addition, the  $z$ -scores are easy to interpret and are defined a priori, whereas the scale of the measure obtained via Box–Cox estimation is arbitrary and probably overfits the data as a result of the parameter search.

A second possible objection to our analysis of word prevalence is that the word/nonword task used to collect the prevalence data is essentially the

same as the lexical decision task used in the Dutch Lexicon Project. It is common in a lexical decision study to find that both accuracy and reaction time are predicted by word frequency and that accuracy and reaction time are correlated, with items that are on average more accurate also leading to faster response times. We agree that one should be highly cautious in using lexical decision accuracies from one experiment to predict lexical decision times of the same experiment. However, we do not think it is prohibited to use another lexical decision task as a survey to gather prevalence data, and we believe it is highly likely that prevalence data gathered in another fashion would lead to very similar results. Still, we admit that more research is necessary to validate the prevalence measure and, specifically, to determine the degree in which it is task dependent. Since the word/nonword task is very efficient for collecting large amounts of data, it should be easier to show that prevalence data collected using a lexical decision paradigm also predicts visual word recognition latencies using another visual word recognition paradigm that does not have a decision component, such as word naming. While large naming data collections do not exist for Dutch, the English Lexicon Project data collected by Balota et al. (2007) do contain average naming latencies for over 40,000 words, making it an ideal target for the validation of a prevalence measure for English words.

Another question concerns the possible interaction of frequency and prevalence. We reported finding only a very small interaction effect on reaction times, which, compared to the size of the main effects, demonstrates that the effects are principally additive. This indicates that prevalence and word frequency do not give widely conflicting estimates but that together they give more precise estimates of word occurrence.

Our analysis of the performance of frequency and prevalence measures showed that, as conjectured, they are complementary. The unique contribution of corpus word frequencies increased in higher frequency bands, and the unique contribution of prevalence decreased in higher prevalence bands, while the unique contribution of frequency showed an almost mirrored increase. Somewhat

unexpectedly, prevalence turned out to have a more stable than expected contribution throughout the different frequency bands, showing that even small variations in highly prevalent words are relevant to lexical processing.

Finally, we come back to the precept that measures of occurrence should match participants' experience as closely as possible. As we mentioned in the introduction, frequency measures can be biased towards a specific location (Van Heuven et al., 2014) or a specific age group (Balota et al., 2004), and this bias is apparent in behavioural data. Since the demographic data collected in studies like the current one can be very fine grained, it should be possible to examine whether prevalence data for participants of a certain age group, location, education, gender, and possibly many more variables also better predict behavioural data from a matching cohort. We believe this as an important matter for future research.

## CONCLUSION

A first analysis of the dataset collected in this study allowed us to gain more insight in the variables influencing vocabulary size and in the influence of prevalence on lexical processing. In particular, the effects of age and multilingualism were theoretically revealing and yield valuable information for research in those areas. Our analyses suggest that the simple relationship between types and tokens made explicit by Herdan's law could drive differences in vocabulary size caused by the accumulation of linguistic experiences throughout life and in multiple languages. In addition, our results suggest that, for multilinguals, in most cases the increase in L1 vocabulary size through shared vocabulary outweighs the loss caused by decreased exposure to L1.

Using this dataset, we were able to identify prevalence as one of the most important predictors of lexical decision processing times in the Dutch Lexicon Project. Although the measure requires more investigation, in particular with respect to its use in other tasks, these results show that it may be a theoretically important variable in



psycholinguistic research. In this paper, we suggested a very simple pathway for the manner in which prevalence works: The more people know a word, the greater its probability of occurrence. However, this is conjecture, and future research may be able to bring other more sophisticated explanations to why the measure has such a strong association with word processing times.

The current study demonstrated that it is possible to design psycholinguistic research that has mass appeal and that is able to recruit hundreds of thousands of participants from a heterogeneous population. At the same time, we should keep in mind that a population taking part in an online test is also biased, in this case probably towards more highly educated internet users who are interested in testing their vocabulary. Compared to crowdsourcing platforms like Amazon Mechanical Turk, however, our study had two advantages. First, with viral recruitment through social media, the study probably reaches a wider population than subjects who sign up to do tasks on a crowdsourcing platform. Second, remunerating participants to do tasks on a crowdsourcing platform (pay-to-participate) is only economically reasonable up to a certain number of participants and also puts a bound on the number of participants than can be recruited. An incentive like a vocabulary score (motivate-to-participate) allows for recruitment of a much larger group and does not put a bound on the number of people who can participate. A limitation of the current approach is that it is not always possible to design a study where there can be an incentive that is powerful enough to motivate large groups of participants. A second limitation is that there can be an enormous amount of participants at one moment in time. The testing infrastructure must be able to handle such a large influx, and any errors in the test will be exacerbated.

As we showed, psycholinguistic research can greatly benefit from this type of research because questions regarding the effects of individual variation in language processing can be studied in a much larger population. In addition, the collection of word knowledge data can be used to derive new objective measures, such as prevalence, which are of theoretical and methodological importance.

## Supplemental material

The prevalence values for Dutch words used in this article can be downloaded from (<http://crr.ugent.be/prevalence/>).

## ORCID

Emmanuel Keuleers  <http://orcid.org/0000-0001-7304-7107>

## REFERENCES

- Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. doi:10.1111/j.1467-9280.2006.01787.x
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (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., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.
- Boon, T. den, & Geeraerts, D. (2005). *Van Dale: groot woordenboek van de Nederlandse taal*. Utrecht: Van Dale Lexicografie.
- Bootstrap. (n.d.). Retrieved May 26, 2014, from <http://getbootstrap.com/>
- Brysaert, M., & New, B. (2009). Moving beyond Kučera 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. doi:10.3758/BRM.41.4.977
- Brysaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.
- De Caluwe, J. (2012). Dutch in Belgium: facing multilingualism in a context of regional monolingualism and standard language ideology. In M. Hüning, U. Vogl, & O. Moliner (Eds.), *Standard languages and multilingualism in European history* (Vol. 1, pp. 259–282). Philadelphia: John Benjamins

- Publishing Company. Retrieved from <http://benjamins.com/catalog/mdm.1>
- Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, 20(2), 1–18.
- Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., ... Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE*, 6(9), e24974. doi:10.1371/journal.pone.0024974
- Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, 5(8), 509–524. doi:10.1111/j.1749-818X.2011.00295.x
- Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, 58(3), 787–814. doi:10.1016/j.jml.2007.07.001
- Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague: Mouton.
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3), 627–633. doi:10.3758/BRM.42.3.627
- 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), 643–650. doi: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 syllabic words and nonwords. *Frontiers in Psychology*, 1. doi:10.3389/fpsyg.2010.00174
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990. doi:10.3758/s13428-012-0210-4
- 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. doi:10.1037/a0030859
- 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/s13428-011-0146-0
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111(2), 309–332. doi:10.1037/0033-295X.111.2.309
- McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology*, 24(2), 222–243. doi:10.1037/a0017619
- Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317(5834), 82–82. doi:10.1126/science.1139940
- Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9–13.
- Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, 17(2), 299–320. doi:10.1037//0882-7974.17.2.299
- Ramscar, M., Hendrix, P., Love, B., & Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, 8(3), 450–481.
- Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, 6(1), 5–42. doi:10.1111/tops.12078
- Raven, J. C. (1965). *Guide to using the Mill Hill vocabulary test with progressive matrices*. London: HK Lewis.
- Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, 15(1), 157–166. doi:10.1017/S1366728910000623
- Singh-Manoux, A., Kivimaki, M., Glymour, M. M., Elbaz, A., Berr, C., Ebmeier, K. P., ... Dugravot, A. (2012). Timing of onset of cognitive decline: Results from Whitehall II prospective cohort study. *BMJ*, 344, d7622–d7622. doi:10.1136/bmj.d7622
- Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, 29(4), 471–488. doi:10.1177/0265532211433033
- Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, 84, 1041–1053.

- Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. doi:10.1080/17470218.2013.850521
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. doi:10.3758/s13428-012-0314-x
- 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. doi:10.1037/a0024177
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. doi:10.3758/PBR.15.5.971
- Zachary, R. A., & Shipley, W. C. (1986). *Shipley institute of living scale: Revised manual*. Los Angeles, CA: Western Psychological Services.