Recognition times for 62 thousand English words:

Data from the English Crowdsourcing Project

Paweł Mandera [1], Emmanuel Keuleers [2], Marc Brysbaert [1]

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

[2] Department Cognitive Science and Artificial Intelligence, Tilburg University, The Netherlands

Keywords: megastudy, word recognition, lexical decision, crowdsourcing

Accepted for publication in *Behavior Research Methods*

Address: 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

We present a new dataset of English word recognition times for a total of 62 thousand words, called the English Crowdsourcing Project. The data were collected via an internet vocabulary test, in which more than one million people participated. The present dataset is limited to native English speakers. Participants were asked to indicate which words they knew. Their response times were registered, although at no point were the participants asked to respond as fast as possible. Still, the response times correlate around .75 with the response times of the English Lexicon Project for the shared words. Also results of virtual experiments indicate that the new response times are a valid addition to the English Lexicon Project. This not only means that we have useful response times for some 35 thousand extra words, but we now also have data on differences in response latencies as a function of education and age.

Research on word recognition has seen an interesting development in the last two decades. Whereas previously, word recognition was investigated in small-scale studies involving some 100 words divided over a factorial design with a few conditions and evaluated with analysis of variance, the new development consisted of collecting word processing times for thousands of words and analyzing them with regression analysis whenever a variable of interest is better represented continuously rather than categorically. Such studies are often called megastudies. Table 1 gives an overview of the megastudies available.

- -

Insert Table 1 about here

- -

Balota, Yap, Hutchison, & Cortese (2013) and Keuleers and Balota (2015) summarized the advantages of the megastudy approach. First, they listed the disadvantages of the factorial approach. These are:

- The difficulty to equate the stimuli in the conditions.
- The fact that many words with a shared feature are presented in a short experiment, which may give rise to context effects.
- The fact that continuous variables are categorized (e.g., divided into high vs. low).
- The fact that the study is limited to stimuli at the extremes of a word characteristic.
- The danger of experimenter bias when selecting words for the various conditions.

The disadvantages of the factorial design are less of an issue in the megastudy approach, because the various control variables can be entered in the regression analysis, participants see a random selection of words, continuous variables are not categorized, and there is no prior stimulus selection by the experimenter (for the last aspect, see also Liben-Nowell, Strand, Sharp, Wexler, & Woods, 2019). Additional advantages of the megastudy approach are:

- More power due to the large number of stimuli.
- The data can be used multiple times to address new questions.
- The relative importance of existing word characteristics can be assessed.
- The impact of a variable can be studied across the entire range.
- The strength of a new, theoretically important variable can be evaluated; the data can also be used to search for new variables.
- The quality of newly presented computational models can be evaluated.
- The quality of competing metrics (e.g., word frequency norms) can be compared.
- If the megastudy includes many participants in addition to many stimuli, individual differences can be studied.

The new possibilities can be illustrated with the English Lexicon project (Balota et al., 2007), consisting of lexical decision and naming times for over 40 thousand English words. In several studies, the dataset has been used to examine the relative importance of word features, such as frequency, length, similarity to other words, part of speech, age of acquisition, valence, arousal, concreteness, and letter bigrams (e.g., Brysbaert & Cortese, 2011; Kuperman, Estes, Brysbaert, & Warriner, 2014; Muncer, Knight, & Adams, 2014; New, Ferrand, Pallier, & Brysbaert, 2006; Schmalz & Mulatti, 2017; Yap & Balota, 2009). It has also been used to test new variables, such as OLD20 (Yarkoni, Balota, & Yap, 2008), the consonant-vowel structure of words (Chetail, Balota, Treiman, & Content, 2015), and word prevalence (Brysbaert, Mandera, McCormick, & Keuleers, 2019). It has been valuable to test mathematical models of word recognition and individual differences (Yap, Balota, Sibley, & Ratcliff, 2012), to understand how compound words are processed (Schmidtke, Kuperman, Gagné, & Spalding, 2016), to study the influence of semantic variables on word recognition (Connell & Lynnott, 2013), to find the best frequency measure for English words (Brysbaert & New, 2009; Gimenes & New, 2016; Herdağdelen & Marelli, 2017), to test new computational models (Norris & Kinoshita, 2012), and to predict word learning in speakers of English as a second language (Berger, Crossley, & Kyle, 2019).

To ensure the usefulness of the English Lexicon Project (ELP), it is important to check for converging evidence in other, independent sources. This motivated Keuleers, Lacey, Rastle, and Brysbaert (2012) to compile the British Lexicon Project (BLP), consisting of lexical decisions to 28,000 monosyllabic and disyllabic words. Other interesting additions were the collection of auditory lexical decision times (Goh, Yap, Lau, Ng, & Tan, 2016; Tucker, Brenner, Danielson, Kelley, Nenadić, & Sims, in press) and semantic decision times (Pexman, Heard, Lloyd, & Yap, 2017).

In the present article, we discuss the development of a new large English database of word processing times (there are large databases for other languages as well, as can be seen in Table 1). The present database is the result of a crowdsourcing project (Keuleers, Stevens, Mandera, & Brysbaert, 2015) that was not primarily set up to analyze response times. Because previous research showed that the collection of reaction times in a web browser can be accurate enough to be a useful method for behavioral research (Crump, McDonnell, & Gureckis, 2013; Reimers & Stewart, 2015), we will examine to what extent the response times from such a paradigm inform us about the ease of word recognition.

Method

Keuleers and Balota (2015) defined a crowdsourcing study as a study in which data are collected outside of the traditional, controlled laboratory settings. The English Crowdsourcing Project (ECP), which is presented here, is part of a series of internet-based vocabulary tests developed at Ghent University, in which participants have to indicate which of the presented stimuli they know as words. The vocabulary tests were started in 2013 in Dutch (Keuleers et al., 2015). The English test started in 2014 (Brysbaert, Stevens, Mandera, & Keuleers, 2016a) and is still running (available at http://vocabulary.ugent.be/). Its main goal was to get an idea of how well words are known in the population, a variable we called word prevalence (Brysbaert, Stevens, Mandera, & Keuleers, 2016b; Brysbaert et al., 2019; Keuleers et al., 2015).

The exact instructions of the ECP vocabulary test are: “In this test you get 100 letter sequences, some of which are existing English words (American spelling) and some of which are made-up nonwords. Indicate for each letter sequence whether it is a word you know or not. The test takes about 4 minutes and you can repeat it as often as you want (you will get new letter sequences each time). If you take part, you consent to your data being used for scientific analysis of word knowledge. Do not say yes to words you do not know, because yes-responses to nonwords are penalized heavily!”

Per test participants received 70 words and 30 nonwords. We expected average participants to know about 70% of the presented word, so we corrected for response bias by presenting around one third of the stimuli as non-words. To discourage guessing, participants were warned that they would be penalized if they responded “word” to nonword stimuli. At the end of the test, participants received an estimate of their vocabulary size, which was a big motivation for them to take part and to recommend the test to others. The presented estimate was computed by subtracting the percentage of word responses to nonwords (false alarms) from the percentage of word responses to words (hits).

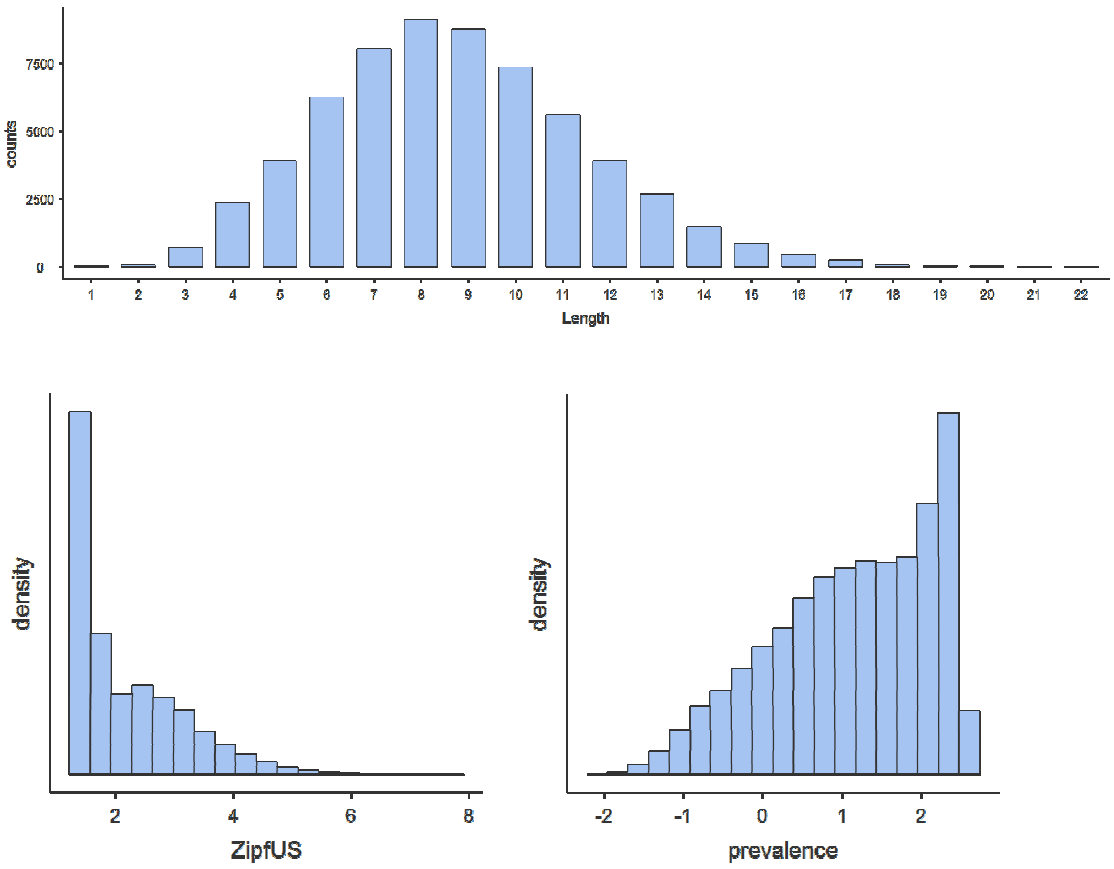
The yes/no format with guessing correction is an established form of vocabulary testing in the language proficiency literature (Ferré & Brysbaert, 2017; Harrington, & Carey, 2009; Lemhöfer & Broersma, 2012; Meara, & Buxton, 1987). However, in the ECP the presented words and nonwords were not fixed like in a regular vocabulary test.

The words were selected from a set of 61,851[1] English words compiled over the years. These words included the lemmas and high frequency irregular word forms from the SUBTLEX databases, supplemented with stimuli from dictionaries and spelling checkers. Figure 1 shows the distributions of word length, word frequency, and word prevalence in the stimulus list. Word length varied from 1

[1] One stimulus (null) got lost in the various handlings of the database because Microsoft Excel automatically converts a number of words to other variable types. The same was true for the words false and true in the initial list; this mistake was corrected about halfway.

to 22 letters. Word frequency is expressed as Zipf-scores (Brysbaert, Mandera, & Keuleers, 2018), going from 1.29 (not present in the corpus) to 7.62 (the word “you”). Particularly interesting is the large number of words not observed in the SUBTLEX-US frequency list (or in most other frequency lists) but present in dictionaries and spelling checkers. Many of these are well known, even though they are rarely used in spoken or written language (such as mindfully, rollerblade, submissiveness, toolbar, jumpstart, freefall, touchable, …; see Brysbaert et al., 2019, for more information). Word prevalence ranges from less than -2 (a word unknown to virtually everybody) to over +2.33 (a word known by more than 99% of the population).

Figure 1: Overview of the word lengths, word frequencies, and word prevalence values present in the stimulus list.



The nonwords were selected from a list of 329,851 pseudowords generated with Wuggy (Keuleers & Brysbaert, 2010). They were constructed to be as similar as possible to the words on length and letter transition probabilities within and across syllables. Because the stimuli presented in the test were not fixed, participants could take the test more than once. Indeed, a few participants took several hundreds of tests over the years.

Specific to the ECP stimulus set is that the vast majority of words consist of uninflected lemma forms. This is different from BLP, where about half of the stimuli were inflected forms (the only inclusion criterion was monosyllabic or disyllabic words) and ELP, which consisted of all words observed in a corpus, including inflected forms and proper nouns (names of people and places).

Although the ECP task involves a yes/no decision, it is important to consider the differences with a traditional lexical decision task. First, at no point were participants told time is an issue. Second, participants were explicitly instructed to only indicate which words they knew and not to guess if they were unfamiliar with a sequence of letters. Participants did the test outside of a university setting and did it because they wanted to know their English proficiency level. Still, Harrington and Carey (2009) noticed that under these conditions the response times (RTs) can be informative. Because averaging over large numbers reduces the noise in the individual observations, the worth of RTs is expected to increase with the number of participants taking part.

Before the start of the test, participants were asked a few basic questions. These were: (1) what their native language was, (2) where they grew up, (3) what the highest degree was they obtained or were working towards, (4) their gender and age, and (5) how many languages they spoke in addition to English and their mother tongue, and (6) how good their knowledge of English was. Participants were not required to provide this information before they could take part, but the vast majority did.

Results and discussion

The data used in the present article are based on all the tests taken between January 2014 and September 2018. During that period we collected more than 142 million answers from 1.42 million experimental sessions.

For the analyses of the current paper, we used the following data pruning pipeline (run entirely before looking at the data; nothing was changed as a result of the analyses).[2]

1) We only took into account the word data. This reduced the dataset from 142 million to 99.5 million.
2) We only used the first 3 sessions from each IP-address, to make sure that no individual had an undue influence (some participants did hundreds of sessions). This reduced the dataset to 93.6 million observations.
3) We deleted the first 9 trials of each session, which were considered training trials, leaving us with 84.3 million observations.
4) RTs longer than 8000 ms were deleted, so that no dictionary consultation could take place. This reduced the dataset to 83.5 million observations.
5) Outliers were filtered out based on an adjusted boxplot method for positively skewed distributions (Hubert & Vandervieren, 2008) calculated separately for the words in each individual session, leaving 79.0 million observations.
6) Sessions with more yes-responses to nonwords than to words were omitted (often people pressing the wrong buttons), further reducing the dataset to 78.7 million data points.

[2] Readers who have doubts about the choices made (introduced in Mandera, 2016, Chapter 4) are invited to analyze the raw data.

7) Finally, only data from users with English as native language who answered the person-related questions were retained. This reduced the final dataset to 41.2 million observations coming from almost 700 thousand sessions.

For 47% of the sessions, responses were collected from a device with a touchscreen; in the other sessions, responses were given on a keyboard. In the touch interface, responses were made using virtual YES and NO buttons; in the keyboard interface, the "*F*" key was used for the no response and the "*J*" key for the yes response. [3]

About 60% of the participants grew up in the US, 22% in the UK, and the remaining 18% in other countries. All words had American spellings (e.g., labor, center, analyze).[4]

Per word there were on average 666 observations in the resulting subset of the data, going from a minimum of 190 to a maximum of 7,895. The reasons for these deviations are twofold. First, we received feedback from the users that our initial list contained too many non-existing adverbs (lucklessly, felinely) and non-existing nouns ending on –ness (gingerliness, gelatinousness). These were pruned, together with some other letter sequences that created confusion (such as compound words written as a single word – clairsentience, taylormade – and the letters of the alphabet). At that time we also entered new words we had come across since the start of the project, which explains why the minimum number of responses is only 190. The high maximum number of responses was due to two occasions on which the randomization algorithm blocked. As a result, the same sequence was presented repeatedly, until we were alerted to the problem. Because of these infelicities, cautious users may want to exclude entries with less than 316 observations (N = 2,544) or more than

[3] Average RT was 1161 ms for keyboard devices and 1258 for touch devices.

[4] We do not present separate data for US and UK participants, because all stimuli were presented in American spelling and the strength of the database lies in the high number of observations per stimulus.

1,000 observations (N = 140), although we do not think these RTs are problematic and we did not exclude them from the analyses presented here.

RTs were calculated on correct trials only. RTs were defined as the time interval between the presentation of the stimulus and the response of the participant. Overall accuracy was .78. Mean RT was 1,297 ms (SD over stimuli is 357). The mean standard deviation in RTs per stimulus across participants was 784 ms (SD over stimuli is 264). Both values are considerably higher than in laboratory based megastudies. For comparison: in the lexical decision part of ELP the mean RT for the words was 784 ms (SD = 135) and the mean standard deviation of the LDT latencies was 278 ms (SD = 92; Balota et al., 2007).

Correlations with data from other megastudies

A first way to measure the merit of the RTs in ECP is to correlate them with the RTs from other megastudies. The prime candidate, of course, is ELP, with its lexical decision RTs and naming latencies. Next is the British Lexicon Project (BLP) also providing lexical decision times. For both databases we used standardized RTs (zRTs), as they correlate more with word characteristics. There was no need to work with standardized RTs for ECP, as the correlation between raw RTs and zRT was r = .992. The reasons for the high correlation are the large number of observations per word (several hundred, compared to the 30-40 observations per word in ELP) and the fact that each participant added only a tiny fraction of the data. Raw RTs are easier to understand because they are closer to human intuitions and they retain individual differences in RT (but see below for some analyses with zRT).

We also excluded words that had an accuracy of less than .85 in ECP, as the RTs of these words are less trustworthy.[5] This left us with a total of 12,001 words for which we had RTs in all databases.

[5] Given that RTs are based on correct responses, the number of responses decrease as accuracy decreases. In addition, it can be assumed that in particular long RTs are missing as a result of no responses (of participants

Because of the design of BLP, the observations are limited to monosyllabic and disyllabic words (the words most often used in experimental research). Table 2 gives the correlations between the databases. As can be seen, for this particular dataset ECP correlates almost as much with ELP_{LDT} as BLP correlates with the same database. This is good news for the value of ECP.

Table 2: Correlations between the RTs of ECP, ELP, and BLP for the items in common that were generally known (N = 12,001). For ELP and BLP standardized RTs were used.

	ELP_{LDT}	ELP_{nam}	BLP
ECP	.75	.54	.76
ELP_{LDT}		.63	.77
ELP_{nam}			.55

A second way to examine the usefulness of the ECP RTs is to see how well they correlate with the RTs from other studies mentioned in Table 1 and, more importantly, how the correlations compare to those with ELP_{LDT} and BLP. Table 3 lists the findings for some classic datasets.

doubting whether they know the word but in the end deciding they do not). Accuracy of .85 corresponds to a prevalence of 1 in Figure 1. The data do not differ much from when a criterion of .75 is used (see below), but the criterion of .85 ensures that the correlation is valid for well-known words.

Table 3: Correlation of the ECP, ELP and BLP RT data with other datasets. For ELP and BLP, zRT values were used. Between brackets, the number of shared words.

	$N_{stimuli}$	ECP	ELP_{LDT}	BLP
Word naming				
Seidenberg & Waters (1989)	2,900	.24 (2,690)	.27 (2,649)	.24 (2,659)
Treiman et al. (1995)	1,327	.41 (1,306)	.37 (1,268)	.36 (1,287)
Spieler & Balota (1997)	2,428	.21 (2,417)	.31 (2,422)	.28 (2,408)
Balota & Spieler (1998)	2,428	.30 (2,417)	.39 (2,422)	.37 (2,408)
Kessler et al. (2002)	3,688	.29 (3,576)	.30 (3,257)	.30 (3,486)
Balota et al. (2007)	40,481	.74 (27,387)	.79 (40,468)	.61 (18,975)
Lexical decision				
Spieler & Balota (1997)	2,428	.65 (2,417)	.72 (2,422)	.72 (2,408)
Balota & Spieler (1998)	2,428	.58 (2,417)	.61 (2,422)	.60 (2,408)
Balota et al. (2007)	40,481	.79 (27,387)	- - -	.77 (18,973)
Keuleers et al. (2010)	28,730	.73 (16,294)	.77 (18,973)	- - -
Progressive demasking				
Lemhöfer et al. (2008) [a]	1.025	.39 (1025)	.40 (1025)	.44 (1024)
Semantic decision				
Pexman et al. (2017)	10,024	.42 (10,024)	.40 (9,211)	.32 (4,855)
Eye movements in reading (gaze duration)				
Pynte & Kennedy (2007)	9,271	.45 (5,555)	.45 (7,122)	.30 (4,710)
Cop et al. (2017)	5,012	.38 (3,782)	.48 (4,634)	.31 (3,306)
Auditory perceptual identification (accuracy)				
Liben-Nowell et al. (2019)	1,081	-.31 (1,062)	-.31 (1,081)	-.32 (1,071)

[a] Averaged over all participants (including L2 speakers)

As can be seen in Table 3, the ECP RTs correlated .79 with the standardized ELP lexical decision times and .73 with the BLP zRTs. These correlations can be considered as the bottom level of reliability for the dataset (based on convergent validity), indicating that some 75-80% of the variance in ECP times is systematic variance that can be explained by stimulus characteristics. As for the correlations with the other datasets, ECP seems to be slightly worse than ELP (in particular for short words) and on par with BLP.

Variance accounted for by word characteristics

A third way to gauge the quality of the ECP dataset is to see how strongly RTs are influenced by word characteristics. In a recent article, Brysbaert et al. (2019) evaluated the contribution of seven variables on ELP zRTs.[6] They were:

- Word frequency (SUBTLEX-US; Brysbaert & New, 2009)
- Word length (in letters)
- Word length (in syllables)
- Number of morphemes (from Balota et al., 2007)
- Orthographic distance to other words (OLD from Balota et al., 2007)
- Phonological distance to other words (PLD from Balota et al., 2007)
- Age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012)
- Concreteness (Brysbaert, Warriner, & Kuperman, 2014)

Table 4 compares the regression analysis for the words in common between ELP and ECP (N = 18,305; the words dropped from the analyses in Table 3 were words for which we did not have information on all variables and words not recognized by 75% of the ELP participants, the criterion used by Brysbaert et al., 2019). For ease of comparison, regression weights are expressed as beta coefficients, meaning that the dependent and independent variables were standardized. Figures 2 and 3 give a graphical display of the effects.

[6] The word prevalence variable cannot be tested here, because it is based on the same dataset

Table 4: Outcome of regressions on the ELP LDT zRTs and the ECP RTs for the words in common (N = 18,305). Beta coefficients are given, which have the same meaning for both regressions. Predictors are centered, to ease the interpretation of the polynomials.

		ELP_{zLDT}	ECP
Word frequency		-.419 ***	-.586 ***
Word frequency squared		.067 ***	.199 ***
Word length (letters)		.124 ***	.154 ***
Word length (letters) squared		.137 ***	.135 ***
Number of syllables		.095 ***	.071 ***
Number of morphemes		-.035 ***	-.011
OLD		.134 ***	-.057 ***
PLD		.059 ***	.029 *
AoA		.186 ***	.131 ***
AoA squared		.068 ***	.101 ***
Concreteness		-.002	-.017 **
	R^2 =	.677	.597

*** p < .001, ** p < .01, * p < .05

Figure 2: Effects of the variables on the standardized ELP lexical decision times. First line: effects of word frequency and length in letters; second line: number of syllables and number of morphemes; third line: orthographic and phonological similarity to other words; last line: age of acquisition and concreteness.

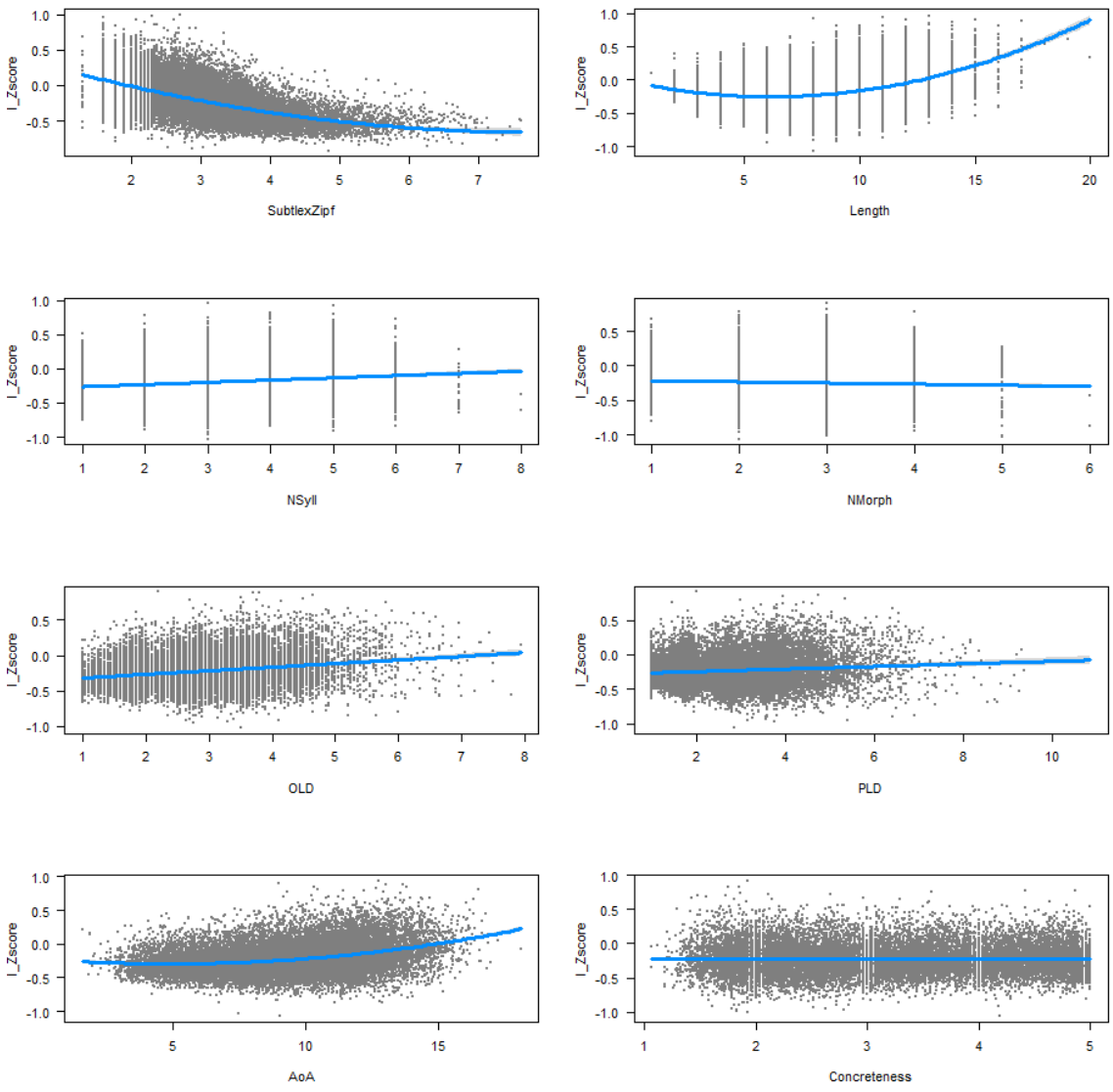
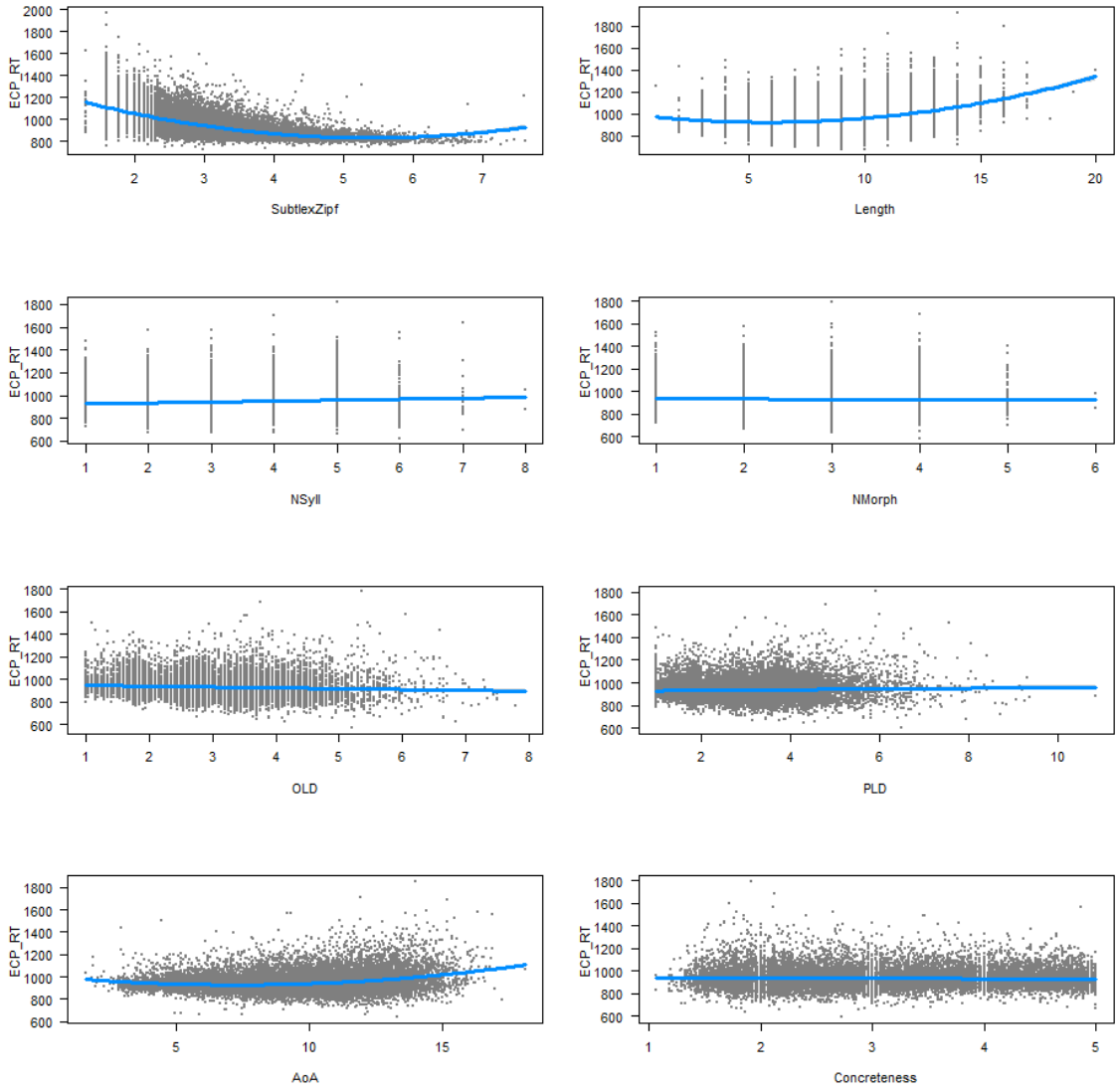


Figure 3: Effects of the variables on the ECP word recognition times. First line: effects of word frequency and length in letters; second line: number of syllables and number of morphemes; third line: orthographic and phonological similarity to other words; last line: Age of acquisition and concreteness.



As can be seen in Table 4 and Figures 2 and 3, the effects of the word variables were quite comparable in the lexical decision part of ELP and ECP. High frequency words were responded to faster than low frequency words, except for the very high-frequency words, which are mostly function words (auxiliaries, conjunctions, determiners, particles, prepositions, pronouns). Function words do not seem to be expected in lexical decision experiments or vocabulary tests, possibly because they are rarely seen in isolation, or because of list-context effects, as the vast majority of stimuli presented in lexical decision tasks are content words. Indeed, the processing cost for these words is not seen in eye movement studies (Dirix, Brysbaert, & Duyck, in press).

Words of 6-8 letters were responded to faster than longer and shorter words; the effect was very much the same in ECP and ELP. Words with extra syllables were responded to more slowly and morphologically complex words were responded to more rapidly than expected on the basis of the other variables. These effects were stronger in ELP than in ECP. Also the similarity to other words tended to have a stronger effect in ELP than in ECP. Here we see the only contradiction between ELP and ECP: Whereas orthographic distance to other words hindered processing in ELP, it facilitated processing in ECP. Finally, the effects of age of acquisition (AoA) and concreteness were larger in ECP than in ELP.

All in all, variables related to the activation of representations in the mental lexicon (frequency, AoA, concreteness) were stronger in ECP than ELP. In contrast, variables related to the similarity with other words (morphology, orthographic and phonological similarity) tended to weigh more heavily in the speeded responses of ELP than in the unspeeded responses of ECP. Interestingly, words were responded to more slowly in ECP when they were orthographically similar to other words, whereas the reverse effect was observed in ELP. The ECP finding is in line with the hypothesis that it is more difficult to recognize a word when it resembles many other words. The ELP finding is in line with the proposal that speeded responses in a lexical decision task are not always based on individual word

recognition but can be based on the total degree of orthographic activation caused by the letter string (Grainger & Jacobs, 1996; Pollatsek, Perea, & Binder, 1999).

The regression accounted for 68% of the variance in ELP zRTs and 60% of the variance in ECP RTs. The correlation between ELP and ECP was .79 for the dataset. This is the same as for all words in common (Table 3), and means that we are still missing some 11-19% of the systematic variance in the datasets.

Virtual experiments

A final way to probe the value of ECP is to see whether we can replicate some classic studies with the dataset. This is done by extracting the RTs from ECP for the stimuli used in the original experiments and running analyses over items. Keuleers et al. (2012) ran a number of such virtual experiments with BLP. The first question they addressed was whether the word frequency effect could be replicated. Given that ECP has a stronger frequency effect than ELP, we would expect this to be the case. Table 5 shows the outcome. To ease the comparison, the ELP and BLP data are given as average RTs and not as zRTs.

Table 5: Virtual experiments on the frequency effect (if needed, British spellings were replaced with American spellings)

	Original	ECP	ELP	BLP
Monsell et al. (1989, Exp 1)				
High frequency, person	538	822	606	534
High frequency, thing	541	831	603	539
Medium frequency, person	553	853	636	571
Medium frequency, thing	570	855	638	565
Low frequency, person	639	971	741	648
Low frequency, thing	617	974	743	630
Effect of frequency	88**	146**	137**	102**
Effect of animacy	1	5	1	6
Frequency * animacy	$p < .01$	n.s.	n.s.	n.s.
Yap et al. (2008, Exp 1)				
High frequency words	557	823	590	531
Low frequency words	605	872	651	574
Frequency effect	48**	49**	61**	43**

The next variable Keuleers et al. (2012) investigated, was AoA. Given that the AoA effect was stronger in ECP than ELP, we again expect to replicate the findings. Table 6 shows the results. We indeed were able to replicate the published patterns. In particular for Gerhand and Barry (1999) the virtual experiment was closer to the original experiment than ELP and BLP, partly because there were several missing observations for the hardest condition in ELP and BLP.

Table 6: Virtual experiments on the AoA effect (if needed British spellings were replaced with American spellings)

	Original	ECP	ELP	BLP
Morrison & Ellis (1995, Exp 5)				
Early acquired	582	837	619	550
Late acquired	648	899	698	608
AoA effect	66**	62**	79**	58**
Gerhand & Barry (1999, Exp 1)				
Early AoA, high frequency	593	833	592	540
Early AoA, low frequency	621	857	673	584
Late AoA, high frequency	603	837	632	538
Late AoA, low frequency	730	957	711	623
AoA effect	59**	52**	39*	18
Frequency effect	77**	72**	80**	64**
AoA * Frequency	50**	48*	1	20

Another topic Keuleers et al. (2012) addressed, was orthographic neighborhood size. The first computational models suggested that words with many neighbors should take longer to process, because there is more competition between activated word forms. A series of lexical decision experiments pointed to facilitation, however, which Grainger and Jacobs (1996) explained by assuming that lexical decision responses can be based on the total activation in the mental lexicon. Words with many neighbors initially create more activation in the lexicon than words with few neighbors and this would lead to a “word” response before the target word is fully recognized.

Given that the OLD effect in ECP was opposite to the one observed in ELP, it is interesting to see what virtual experiments give for this variable. Table 7 shows the results for some classic studies. Remember that these all involved monosyllabic words, a very small subset of the words in ECP. Although the results of the virtual experiments are largely in line with those of the original studies

(including those of ECP), Table 7 is primordially a testimony of the weaknesses of the factorial design, as listed in the introduction. Most studies had too few stimuli to find anything significant in an analysis over stimuli, meaning that the differences could be due to one or two stimuli in one or the other condition. Overall, however, it looks like the effects of neighborhood size are facilitatory in lexical decision (in particular the number of body neighbors), and that inhibitory effects are largely due to the presence of a neighbor with a higher frequency (see also Chen & Mirman, 2012). In addition, neighbors are not limited to words of the same length, but include words with one letter omitted or added (Davis & Taft, 2005), as captured by the OLD and PLD measures. More importantly for the present discussion, the ECP findings are well in line with those of the other data for the monosyllabic words.

Table 7: Virtual experiments on the orthographic neighborhood effects (if needed British spellings were replaced with American spellings)

	Original	ECP	ELP	BLP
Andrews (1992, Experiment 1)				
High frequency, large N	586	833	592	539
High frequency, small N	570	822	589	536
Low frequency, large N	714	948	724	629
Low frequency, small N	757	988	759	661
Frequency effect	157**	140**	151**	107**
Effect N	-13	-14	-16	-14
Frequency * N	29**	25	19	17
Sears et al. (1995, Exp 1)				
High frequency, large N	509	827	585	538
High frequency, small N	528	819	590	532
Low frequency, large N	577	845	631	583
Low frequency, small N	587	861	656	564
Frequency effect	63**	30**	46**	38**
Effect N	-15*	-12	-15	12
Frequency * N	5	4	10	-6
Perea & Pollatsek (1998)				
High frequency neighbor	632	893	665	583
No HF neighbor	606	878	663	580
Effect HF neighbor	26*	15	2	0

Ziegler & Perry (1998)				
Many body neighbors (BN)	625	850	640	559
Few body neighbors	657	871	656	582
Effect body neighbors	-32*	-21*	-16	-23*
Large N controlled for BN	650	874	650	582
Small N controlled for BN	636	851	642	559
Effect N	14	23	8	23
Pollatsek et al. (1999, Exp1)				
Large N	579	872	643	573
Small N	605	891	661	569
N effect	-26*	-19	-18	4
Davis & Taft (2005, Exp 2)				
Deletion neighbor	641	895	694	599
Control	614	870	666	579
Effect	27**	25**	28*	20*

In a series of articles, Yates and colleagues argued that in particular phonological neighbors speed up lexical decisions (Yates, 2005, 2008; Yates, Locker, & Simpson, 2004). Table 8 looks at how well these findings replicate in ECP, ELP, and BLP. The basic finding of Yates et al. (2004) was replicated successfully with the stimuli selected by the authors, but the difference between two and three phonological neighbors (Yates, 2008) was less consistent. This agrees with Davis's (2010) argument that the main neighborhood size effect is between no neighbors and one neighbor (with higher frequency).

Table 8: The effect of phonological neighborhood size in published lexical decision experiments and in virtual experiments with the same stimuli

	Original	ECP	ELP	BLP
Yates et al. (2004, Exp 1)				
Large N	622	867	644	578
Small N	695	944	702	633
Effect	-73**	-77**	-58**	-55**
Yates (2005)				
Large N	656	854	628	567
Small N	729	891	686	610
Effect	-73**	-37*	-58**	-43**
Yates (2008)				
P = 2	659	866	653	575
P = 3	621	864	627	566
Effect	35*	2	26*	9

Another effect worth looking at is the influence of word ambiguity. Rodd, Gaskell, and Marslen-Wilson (2002) argued that ambiguity has two opposite effects. Words with unrelated meanings (e.g., can, second) have longer lexical decision times than unambiguous control controls, whereas words with related senses (uniform, burn) are responded to more rapidly than unambiguous control words. Table 9 shows that the facilitatory effect of multiple senses tends to be stronger than the inhibitory effect of multiple unrelated meanings, and that the effects seem to be clearer in ELP than in ECP, at least for the stimuli selected by Rodd et al. (2002).

Table 9: The effect of word ambiguity in published lexical decision experiments and in virtual experiments with the same stimuli

	Original	ECP	ELP	BLP
Rodd et al. (2002, Exp 2)				
Many meanings, few senses	587	856	650	572
Many meanings, many senses	578	849	617	559
One meaning, few senses	586	853	639	561
One meaning, many senses	567	838	609	551
Effect of meanings	6	7	9	9
Effect of senses	-14*	-11	-31**	-11*
Interaction	5	4	1	1

A final finding in lexical decision research we will look at is the size effect. Sereno, O'Donnell, and Sereno (2009) reported that participants respond faster to words representing big things (bed, truck, buffalo) than to matched words representing small things (cup, thumb, apricot). The authors related this finding to the importance of embodied cognition, a view according to which cognitive processing involves internal simulations of perceptual and motor processes (Barsalou, 2008; Fischer & Zwaan, 2008). Kang, Yap, Tse, and Kurby (2011), however, were unable to replicate the finding and, in addition, reported that the effect was absent in ELP. Table 10 gives the outcome of a virtual experiment in ECP, in addition to ELP. BLP could not be used, as nearly half of the stimulus materials

were longer than two syllables. As can be seen in Table 10, the size effect was not replicated in ECP either.

Table 10: The effect of concept size in Sereno et al. (2009) and in virtual experiments with the same stimuli

	Original	Kang	ECP	ELP
Sereno et al. (2009)				
Small concepts	528	654	858	649
Large concepts	513	649	858	654
Effect	15**	5	0	-5

Education differences

Up to now we have discussed findings ECP has in common with ELP and BLP and seen that for these words ECP is a valid addition to the existing megastudies. However, the merit of ECP goes further. For a start, ECP offers data for 35 thousand words not covered by ELP, and for 50 thousand words not present in BLP. This substantially increases the resources available to researchers.

In addition, ECP includes more participants than the typical undergraduate students. Some participants had only finished high school, others had achieved a bachelor degree (often outside university), a master degree (at university), or a PhD degree. On average, we had 170 observations per word for participants who finished high school, 296 for participants with a bachelor degree, 125 for participants with a master degree, and 46 for participants with a PhD degree. Because of the small numbers in the last group, we limit the analysis to the first three groups.

Keuleers et al. (2015) and Brysbaert et al. (2016a) already discussed the number of words known as a function of education level. Participants with more education know more words than participants with less education. Interestingly, the differences were modest when the participants' age was taken into account and mainly originated during the study years, arguably because the participants then

were acquiring the academic vocabulary related to their studies and word use in higher education (Coxhead, 2000).

To compare the three education groups, we report the outcome of the regression analysis with the data discussed in Table 4 (N = 18,305). Two outcomes are given: First, the analysis with the unchanged regression weights, and then the analysis with the beta coefficients. The former tells us how the RTs differ between groups, the latter how the relative importance of the variables varies. Variables were centered, so that the intercept gives us the RT of the "middle" word. Interestingly, the ELP zRTs correlate highest with the participants who finished high school (r = .79), then with those who have bachelor degrees (r = .77), and lowest with the participants who have a master degree (r = .71). This is in line with the fact that most ELP participants were undergraduate students. On the other hand, the lower correlation with the master degree group is probably also to some extent due to the lower number of observations for this group (resulting in a lower reliability of the ECP RTs).

Table 11 shows the outcome of the analyses. Participants with less education responded more slowly as can be seen in the intercepts and tended to show a stronger effect of frequency, AoA and number of syllables. Participants with master degrees seem to be more willing to respond on the basis of total orthographic activation, given that the effect of OLD is stronger for them. Overall, however, the differences are small and do not seem to offset the smaller number of observations per word. In particular R^2 for the participants with a master degree has dropped considerably (R^2 = .48).

Table 11: Outcome of regression analyses for the three education groups of ECP (high school, bachelor, master) for the words in common with ELP (N = 18,305). Polynomial to the third degree used for word frequency (the addition of the third power accounted for less than .5% of variance explained but lowered the predicted RTs for the high frequency words). Predictors are centered to improve interpretation.

Regression weights

	ECP_{High}	ECP_{Bach}	ECP_{Mast}
Intercept	955 ***	916 ***	906 ***
Word frequency	-85 ***	-79 ***	-73 ***
Word frequency squared	24 ***	25 ***	25 ***
Word frequency third power	-3 ***	-3 ***	-4 ***
Word length (letters)	8 ***	8 ***	7 ***
Word length (letters) squared	2 ***	2 ***	2 ***
Number of syllables	14 ***	7 ***	6 ***
Number of morphemes	-5 ***	-1	1
OLD	-6 ***	-6 **	-9 ***
PLD	6 **	2	2
AoA	10 ***	6 ***	4 ***
AoA squared	2 ***	1 ***	1 ***
Concreteness	-4 ***	-1 *	-2 **
R^2 =	.600	.566	.484

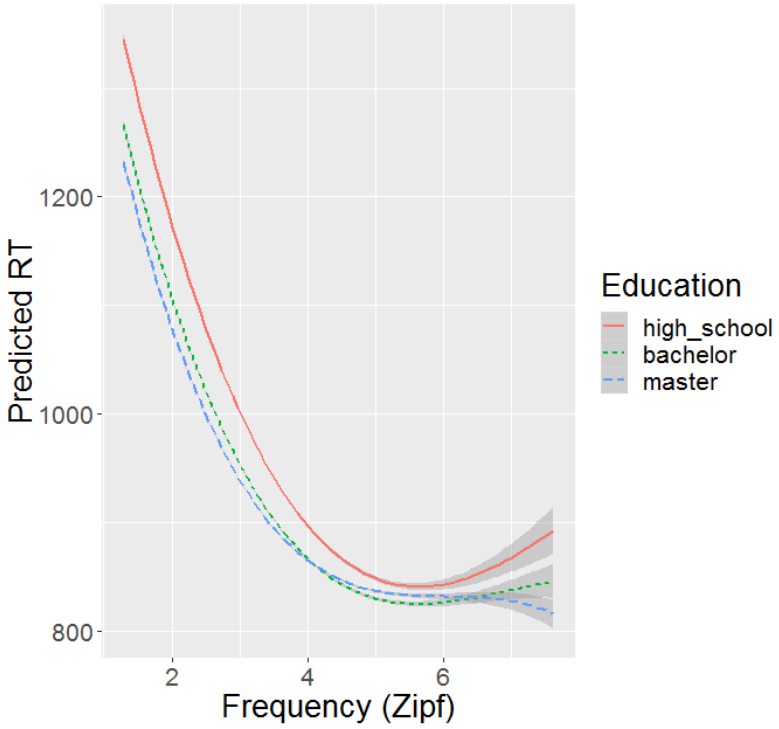
*** p < .001, ** p < .01, * p < .05

Beta coefficients

	ECP_{High}	ECP_{Bach}	ECP_{Mast}
Word frequency	-.49	-.54	-.52
Word frequency squared	.21	.25	.27
Word frequency third power	-.07	-.10	-.11
Word length (letters)	.12	.15	.14
Word length (letters) squared	.12	.13	.13
Number of syllables	.10	.06	.05
Number of morphemes	-.03	-.01	.00
OLD	-.04	-.05	-.07
PLD	.05	.02	.02
AoA	.18	.12	.08
AoA squared	.13	.09	.07
Concreteness	-.03	-.01	-.02

Figure 4 shows how the predicted RTs differ for the three education groups as a function of word frequency. This illustrates that the effect of education is particularly strong for low frequency words.

Figure 4: Predicted response times for the three education groups as a function of word frequency. Regressions included all variables mentioned in Table 11.



Age differences

Another variable we can look at, is the age group of the participants (Wulff et al., in press). Davies, Arnell, Birchenough, Grimmond, and Houlson (2017) reported that the effects of word frequency and AoA on lexical decision times become smaller with increasing age over adult life. At the same time, there was ageing-related response slowing, which could be attributed to decreasing efficiency of stimulus encoding and/or response execution in older age, but is also consistent with increased processing costs related to the accumulation of information learned over time (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014).

The decrease of the word frequency effect in older participants is expected on the basis of their longer exposure to the language. A number of publications indicate that the word frequency effect becomes smaller as participants are exposed to more language (Brysbaert, Lagrou, & Stevens, 2017; Brysbaert, Mandera, & Keuleers, 2018; Cop, Keuleers, Drieghe, & Duyck, 2015; Diependaele, Lemhöfer, & Brysbaert, 2013; Mainz, Shao, Brysbaert, & Meyer, 2017; Mandera, 2016, Chapter 4; Monaghan, Chang, Welbourne, & Brysbaert, 2017). This is expected on the basis of two models. First, connectionist models at a certain point show a decrease in the frequency effect, when overlearning takes place (Monaghan et al., 2017). Second, Mandera (2016, Chapter 4) showed that a decrease in the frequency effect as a function of practice is predicted if word learning follows a power law rather than an exponential law (Logan, 1988).

At the same time, exposure to language increases the vocabulary of a person. Healthy old participants indeed have a larger vocabulary than young adults (Verhaeghen, 2003) and vocabulary has been shown to have logarithmic growth over age (Keuleers, Stevens, Mandera, & Brysbaert, 2015). Particularly related to the ECP stimulus set, Brysbaert et al. (2016a) reported that a 60 year old person on average knows 6,000 lemmas more than a 20 year old person, or an increase of some three words per week.

In contrast to Davies et al. (2017) and our own work, Cohen-Shikora and Balota (2016) failed to find a decrease in the word frequency effect as a function of age. They administered three tasks (lexical decision, word naming, and animacy judgment) to 148 participants, ranging in age from 18 to 86 years. Each task consisted of responses to 400 words (in counterbalanced order). Only in word naming latencies was there a hint of a smaller word frequency effect in older participants than in younger participants. At the same time, the data of Cohen-Shikora and Balota (2016) replicated the core effects of the other studies: (1) Older participants were slower and more accurate than younger participants, (2) older participants had a larger vocabulary than younger participants, and (3) there was a negative correlation between vocabulary size and the word frequency effect. The analyses of

Cohen-Shikora and Balota (2016) were done on z-scores of RTs. Could this have made a difference, as z-scores not only eliminate differences in means but also equalize the standard deviations?

To compare age groups, we made a distinction between participants of 18-23 years (on average 104 observations per word), 24-29 (117 observations), 30-39 (150 observations), 40-49 (106 observations), and 50+ (124 observations). To see whether our age differences were in line with those of Spieler and Balota (1997; young participants) and Balota and Spieler (1998; old participants), we looked at the correlations with these datasets. For the young participants of Spieler and Balota (1997), the correlations with increasing age group were: .60, .59, .58, .52, and .50. For the old participants of Balota and Spieler (1998) the correlations were respectively: .51, .50, .53, .48, and .49. The pattern of result was as expected for the young participants, but not for the old participants. One reason may be that the old participants of Balota and Spieler had a mean age of 74 years, substantially older than the ECP participants. Another contributor probably is differences in the reliability of the word processing estimates in the various age groups.

Table 12 and the left panel of Figure 5 show the results of the regression analyses. They are in line with the observation of Davies et al. (2017) that the frequency and the AoA effect decrease over age. The OLD and PLD effects also seem to become smaller, in line with the observation that the older participants took some more time to respond. Finally, it looks like the effects of number of syllables and concreteness increase as adults grow older.

Table 12: Outcome of the regression analyses for the five age groups in ECP for the words in common with ELP (N = 18,305). Predictors are centered.

Regression weights

	ECP_{18}	ECP_{24}	ECP_{30}	ECP_{40}	ECP_{50}
Intercept	895 ***	909 ***	919 ***	925 ***	975 ***
Word frequency	-85 ***	-84 ***	-81 ***	-74 ***	-68 ***
Word frequency ^{2}	22 ***	27 ***	28 ***	25 ***	23 ***
Word frequency ^{3}	-2 ***	-3 ***	-4 ***	-4 ***	-3 ***
Word length (letters)	8 ***	9 ***	9 ***	7 ***	5 ***
Word length (letters) ^{2}	2 ***	2 ***	2 ***	2 ***	3 ***
Number of syllables	-1	2	6 ***	13 ***	22 ***
Number morphemes	-3 *	-1	1	-2	-2
OLD	-9 ***	-8 ***	-7 **	-5 *	-2
PLD	6 ***	5 **	4 *	1	-2
AoA	9 ***	6 ***	6 ***	5 ***	5 ***
AoA squared	2 ***	2 ***	1 ***	1 ***	1 ***
Concreteness	2 *	-1	-3 ***	-4 ***	-5 ***
R^2 =	.513	.525	.536	.504	.510

*** p < .001, ** p < .01, * p < .05

Beta coefficients

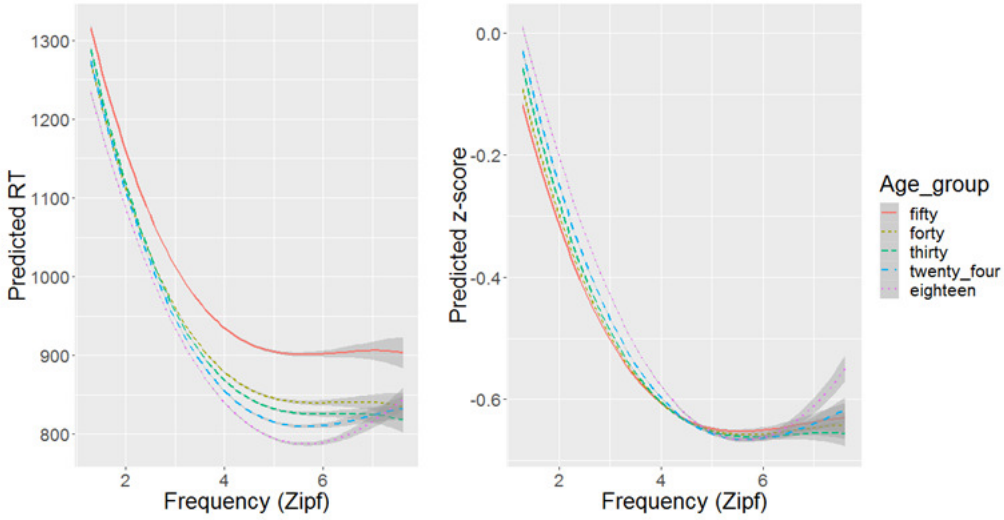
	ECP_{18}	ECP_{24}	ECP_{30}	ECP_{40}	ECP_{50}
Word frequency	-.54	-.53	-.52	-49	-.45
Word frequency ^{2}	.21	.24	27	.24	.23
Word frequency ^{3}	-.05	-.08	-.10	-.10	-.10
Word length (letters)	.13	.14	.15	.13	.10
Word length (letters) ^{2}	.10	.10	.12	.14	.16
Number of syllables	-.01	.02	.05	.11	.17
Number morphemes	-.02	-.00	.00	-.01	-.01
OLD	-.07	-.06	-.05	-.04	-.02
PLD	.05	.04	.03	.01	-.02
AoA	.18	.13	.11	.09	.09
AoA squared	.11	.09	.08	.08	.09
Concreteness	.01	-.00	-.02	-.03	-.04

The left panel of Figure 5 shows the predicted RTs for the five age groups as a function of word frequency. These point to longer response times for older participants. At the same time, because the cost for low frequency words is smaller for older participants, the age differences in RT are smallest for the low frequency words.

To make sure that our results did not rely on the use of raw RTs as the dependent variable, we also analyzed the standardized RTs. As can be seen in the right panel of Figure 5, the findings remained

the same. Because zRTs eliminate differences in average RTs, they more clearly illustrate the smaller frequency effect in older participants than in younger.

Figure 5: Predicted response times for the five age groups as a function of word frequency. Left panel: raw RTs; right panel: zRTs. Regressions included all variables mentioned in Table 12.



One reason for the difference in findings between Figure 5 and Cohen-Shikora and Balota (2016) could be that Cohen-Shikora and Balota (2016) were very careful to equate their groups on education level. It is possible that relatively more educated older people took part in our study than younger people (e.g., because they have more access to internet).[7] To test this possibility, we compared the age group of 24-30 years to the age group of 50-59 years for the participants with high school bachelor and master education, once with the data analyzed as in Figure 5 and once with equal weight given to the three education levels. The age group 24-29 was chosen because it is the youngest group for which master education is possible; the group 50-59 was chosen because it is the

[7] The authors thank one of the reviewers for the suggestion.

oldest group with comparable homogeneity. To optimize comparison with Cohen-Shikora and Balota (2016) we used zRT as dependent variable.

Figure 6: Comparison of the frequency effect in the age groups 24-30 and 50-59 years, when not controlled for possible differences in education level (left) and when controlled for such differences (right). Effects calculated on zRTs. Regressions included all variables mentioned in Table 12.

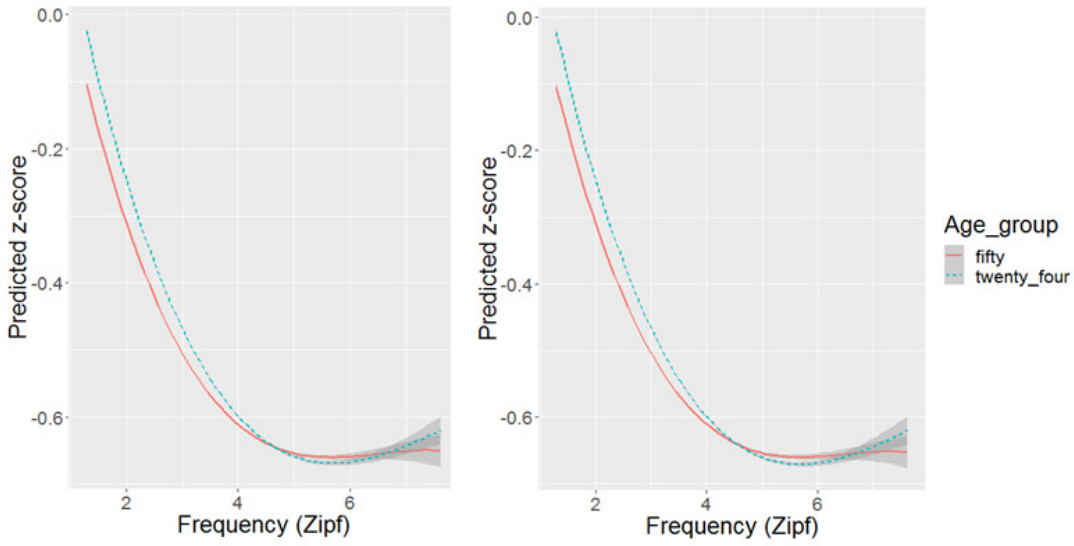


Figure 6 shows the outcome. There is no evidence that the smaller frequency effect in 50-59 group is due to differences in education level (something we did not see in the distribution of education levels in the two age groups either). So, our data agree more with those of Davies et al (2017) than those of Cohen-Shikora and Balota (2016). A further challenge for the interpretation of Cohen-Shikora and Balota (2016) is how to square the absence of a correlation between age and frequency with the presence of significant correlations between vocabulary size and frequency effect on the one hand and between vocabulary size and age on the other hand. To the defense of Cohen-Shikora and Balota (2016), their study is the only one to include several word processing tasks, a large group of

participants with ages above 60 years, and extensive attempts to match the groups of participants. On the negative side, they included words a more restricted frequency range than we did (Zipf scores going from roughly 2.0 to 5.2, with a mean of 3.6). This may have made it more difficult to see the interaction.

Conclusions

We present a new word dataset, the English Crowdsourcing Project (ECP), which is larger than all available datasets (Table 1). It is larger both in the number of words included and in the number and variety of participants taking part.

The dataset was collected by means of an internet vocabulary test, in which participants indicated which words they knew and which not. In order to discourage yes responses to unknown words, about one third of the stimuli were nonwords and participants were penalized if they said yes to these nonwords.

Although speed of responding was not mentioned as an evaluation criterion to the participants, the present analyses show that the response times correlate well with lexical decision times collected in laboratory settings. They are just some 250 ms longer. Surprisingly, the longer response times did not lead to larger effects in the virtual experiments. For all the experiments, the effect in ECP was comparable to the original effects and those in the English Lexicon Project (ELP) and the British Lexicon Project (BLP). This is unexpected, because often longer RTs are accompanied by larger differences between conditions (e.g., Table 8 of Keuleers et al., 2012). It suggests that the extra time in ECP is largely unrelated to word recognition and the decision processes (for a model including such a time delay, see Ratcliff, Gomez, & McKoon, 2004). Apparently, participants took some extra time to perceive the stimulus and give a response. In this respect, it is important to mention that RTs in a lexical decision task drop by some 100 ms in the first few hundred trials (Keuleers et al., 2010, 2012).

Given that most participants completed only 100 trials in ECP, this can explain some of the extra time taken to respond. Another contributor may be the software used to present the stimuli and collect the responses via the internet.

To some extent it is surprising that untimed answers to a vocabulary test resemble lexical decision times so well, when based on large numbers of observations. This testifies to the ecological validity of the lexical decision task, as very much the same results are obtained in an untimed vocabulary test outside of academia as on a speeded response task in the laboratory.

ECP is further interesting because a large range of people took part. Surprisingly, we found no large differences between education levels (Figure 4). Presumably this is due to the fact that only people interested in language and with easy access to internet took part in the test. There is evidence that the size of the frequency effect depends more on the amount of reading and language exposure than on the intelligence or the education level of the participants (Brysbaert, Lagrou, & Stevens, 2017). ECP does point to some interesting effects of age (or language exposure), however. The effects of frequency and age of acquisition seem to become smaller as adults grow older (see also Davies et al., 2017), whereas older people seem to be more affected by the meaning of the words (as indicated by the concreteness effect) and by the complexity of the word (the number of syllables). Further, targeted experiments will have to confirm these initial impressions. Such experiments could also try to include an even wider variety of participants.

Availability

The raw data and Excel files containing the most important information can be found at the Open Science Framework webpage https://osf.io/rpx87/ or on our website http://crr.ugent.be/. To facilitate analyses of the full dataset, we release a Python module for working with the raw data (available at https://github.com/pmandera/vocab-crowd).

The Excel files are included for a broader audience as their usage does not require programming skills. First, we have the master file containing the information calculated across all participants, called *English Crowdsourcing Project All Native Speakers*. Its outline is shown in Figure 6.

Figure 6: Outline of the ECP master file including RTs based on all native speakers

	A	B	C	D	E	F	G	H	I
1	**spelling**	**nobs**	**accuracy**	**prevalence**	**rt_mean**	**rt_std**	**zrt_mean**	**zrt_std**	**ZipfUS**
2	a	690	1.00	1.92	972	413	-0.43	0.55	7.31
3	aardvark	744	0.94	1.68	1087	593	-0.29	0.57	2.63
4	aardwolf	718	0.19	-0.79	1670	1162	0.64	1.13	1.29
5	abaca	627	0.21	-0.71	1863	1163	0.70	1.15	1.59
6	aback	672	0.86	1.08	1221	641	-0.05	0.69	2.50
7	abacus	663	0.91	1.43	1053	575	-0.29	0.67	2.41
8	abaft	637	0.22	-0.88	1344	844	0.16	0.92	1.77
9	abalone	668	0.66	0.50	1252	727	-0.08	0.71	2.72
10	abandon	605	1.00	2.43	934	551	-0.54	0.55	3.91
11	abandoned	650	1.00	2.58	972	594	-0.50	0.58	4.12
12	abandonee	688	0.66	0.41	1857	1125	0.76	1.11	1.29
13	abandoner	646	0.87	1.08	1439	1077	0.09	0.92	1.59
14	abandonment	746	0.99	2.18	1085	560	-0.32	0.61	2.99

Column A gives the word. Column B says how many observations there were for that word. Column C gives the response accuracy, indicating the number of observations on which the RTs are based. We would prefer users not to use the information of Column C for anything other than the analysis of RTs. In Brysbaert et al. (2019) we present the word prevalence measure, which is better than accuracy (even though it correlates .96 with the accuracies reported here). Word prevalence is given in Column D. Columns E to H contain the new information: the ECP RTs and their standard deviations across participants, plus the zRTs and their standard deviations. Finally, for the user's convenience, Column I includes the SUBTLEX-US frequencies expressed as Zipf values (Brysbaert et al., 2018).

In addition to the master file, we also have files with the data split per education level (ECP Education groups), per age (ECP Age groups), and per age * education level. Users who want other summary files, are invited to make them themselves on the basis of the raw data.

References

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(4),* 1037-1053.

Aguasvivas, J., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2018). SPALEX: A Spanish lexical decision database from a massive online data collection. *Frontiers in Psychology, 9, 2156*. doi: 10.3389/fpsyg.2018.02156.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 234–254.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133(2),* 283-316.

Balota, D. A. & Spieler, D. H. (1998). The utility of item level analyses in model evaluation: A reply to Seidenberg & Plaut (1998). *Psychological Science, 9(3),* 238-240.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2013). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual Word Recognition Volume 1: Models and methods, orthography and phonology* (pp. 90-115). New York, NY: Psychology Press.

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.

Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617–645.

Berger, C. M., Crossley, S. A., & Kyle, K. (2017). Using Native-Speaker Psycholinguistic Norms to Predict Lexical Proficiency and Development in Second-Language Production. *Applied Linguistics, 40(1)*, 22-42.

Brysbaert, M. & Cortese, M.J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology, 64*, 545-559.

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

Brysbaert, M., Mandera, P., McCormick, S.F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51(2),* 467-479.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*, 45-50.

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

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016a) How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology 7:1116.* doi: 10.3389/fpsyg.2016.01116.

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

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904-911.

Chang, Y. N., Hsu, C. H., Tsai, J. L., Chen, C. L., & Lee, C. Y. (2016). A psycholinguistic database for traditional Chinese character naming. *Behavior Research Methods, 48(1),* 112-122.

Chateau, D., & Jared, D. (2003). Spelling–sound consistency effects in disyllabic word naming. *Journal of Memory and Language, 48(2),* 255-280.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review, 119(2),* 417-430.

Chetail, F., Balota, D., Treiman, R., & Content, A. (2015). What can megastudies tell us about the orthographic structure of English words?. *The Quarterly Journal of Experimental Psychology, 68(8),* 1519-1540.

Cohen-Shikora, E. R., & Balota, D. A. (2016). Visual word recognition across the adult lifespan. *Psychology and Aging, 31(5)*, 488-502.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, *22*(5), 1216–1234.

Cohen-Shikora, E. R., Balota, D. A., Kapuria, A., & Yap, M. J. (2013). The past tense inflection project (PTIP): Speeded past tense inflections, imageability ratings, and past tense consistency measures for 2,200 verbs. *Behavior Research Methods, 45(1)*, 151-159.

Connell, L., & Lynott, D. (2014). I see/hear what you mean: Semantic activation in visual word recognition depends on perceptual attention. *Journal of Experimental Psychology: General, 143(2),* 527-533.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods, 49(2),* 602-615.

Cortese, M.J., Hacker, S., Schock, J. & Santo, J.B. (2015a). Is reading aloud performance in megastudies systematically influenced by the list context*? Quarterly Journal of Experimental Psychology, 68*, 1711-1722. doi: 10.1080/17470218.2014.974624

Cortese, M.J., Khanna, M.M., & Hacker, S. (2010) Recognition memory for 2,578 monosyllabic words. *Memory, 18*, 595-609. DOI: 10.1080/09658211.2010.493892.

Cortese, M.J., Khanna, M.M., Kopp, R., Santo, J.B, Preston, K.S., & Van Zuiden, T. (2017). Participants shift response deadlines based on list difficulty during reading aloud megastudies, *Memory & Cognition, 45*, 589-599.

Cortese, M.J., McCarty D.P., & Schock, J. (2015b). A mega recognition memory study of 2,897 disyllabic words. *Quarterly Journal of Experimental Psychology, 68,* 1489-1501. doi: 10.1080/17470218.2014.945096

Cortese, M. J., Yates, M., Schock, J., & Vilks, L. (2018). Examining word processing via a megastudy of conditional reading aloud. *Quarterly Journal of Experimental Psychology, 71(11)*, 2295-2313.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34(2),* 213-238.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE, 8(3), e57410*. http://doi.org/10.1371/journal.pone.0057410

Davies, R., Barbón, A., & Cuetos, F. (2013). Lexical and semantic age-of-acquisition effects on word naming in Spanish. *Memory & Cognition, 41(2),* 297-311.

Davies, R. A., Arnell, R., Birchenough, J. M., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(8),* 1298-1338.

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

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

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.

Dirix, N., Brysbaert, M., & Duyck, W. (in press). How well do word recognition measures correlate? Effects of language context and repeated presentations. *Behavior Research Methods*.

Dufau, S., Grainger, J., Midgley, K. J., & Holcomb, P. J. (2015). A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological Science, 26(12)*, 1887-1897.

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology, 68(8)*, 1469-1488.

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., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., ... & Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods, 50(3),* 1285-1307.

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.

Ferré, P., & Brysbaert, M. (2017). Can Lextale-Esp discriminate between groups of highly proficient Catalan-Spanish bilinguals with different language dominances? *Behavior Research Methods, 49,* 717-723.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology, 61,* 825–850.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods, 45(4),* 1182-1190.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language, 140*, 1-11.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2018) The Natural Stories Corpus. In *Proceedings of LREC 2018, Eleventh International Conference on Language Resources and Evaluation* (pp. 76—82). Miyazaki, Japan. Available at http://www.lrec-conf.org/proceedings/lrec2018/pdf/337.pdf.

Gerhand, S., & Barry, C. (1998). Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 267–283.

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods, 48(3)*, 963-972.

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M., & Tan, L. C. (2016). Semantic richness effects in spoken word recognition: A lexical decision and semantic categorization megastudy. *Frontiers in psychology, 7: 976*.

González-Nosti, M., Barbón, A., Rodríguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765 words. *Behavior Research Methods, 46(2)*, 517-525.

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

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System, 37(4)*, 614-626.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science, 41(4)*, 976-995.

Heyman, T., Van Akeren, L., Hutchison, K. A., & Storms, G. (2016). Filling the gaps: A speeded word fragment completion megastudy. *Behavior Research Methods, 48(4)*, 1508-1527.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis, 52(12)*, 5186-5201.

Husain, S., Vasishth, S., and Srinivasan, N. (2014). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research, 8(2)*, 1-12.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods, 45(4)*, 1099-1114.

Kang, S. H., Yap, M. J., Tse, C. S., & Kurby, C. A. (2011). Semantic size does not matter:"Bigger" words are not recognized faster. *The Quarterly Journal of Experimental Psychology, 64(6)*, 1041-1047.

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language, 47*, 145-171.

Keuleers, E & Balota, D.A. (2015) Megastudies, crowd-sourcing, and large datasets in psycholinguistics: An overview of recent developments, *The Quarterly Journal of Experimental Psychology. 68(8)*, 1457-1468.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*, 627-633.

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, M., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology, 68*, 1665-1692.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A.B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General, 143*, 1065-1081.

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

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., & Kliegl, R. (2019). Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods*.

Lee, C. Y., Hsu, C. H., Chang, Y. N., Chen, W. F., & Chao, P. C. (2015). The feedback consistency effect in Chinese character recognition: Evidence from a psycholinguistic norm. *Language and Linguistics, 16(4)*, 535-554.

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.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(1)*, 12-31.

Liben-Nowell, D., Strand, J., Sharp, A., Wexler, T., & Woods, K. (2019). The Danger of Testing by Selecting Controlled Subsets, with Applications to Spoken-Word Recognition. *Journal of Cognition, 2(1), 2*. DOI: http://doi.org/10.5334/joc.51

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39(2)*, 192-198.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95(4)*, 492-527.

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods, 50(2)*, 826-833.

Mainz, N., Shao, Z., Brysbaert, M., & Meyer, A. (2017). Vocabulary Knowledge Predicts Lexical Processing: Evidence from a Group of Participants with Diverse Educational Backgrounds. *Frontiers in Psychology, 8:1164*.

Mandera, P. (2016). *Psycholinguistics on a large scale: combining text corpora, megastudies, and distributional semantics to investigate human language processing*. Ghent University: Unpublished PhD thesis. Available at: http://crr.ugent.be/papers/pmandera-disseration-2016.pdf.

Meara, P. M., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*, 142–154.

Monaghan, P., Chang, Y. N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language, 93*, 1-21.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General, 118*, 43–71.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word-frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 116–133.

Mousikou, P., Sadat, J., Lucas, R., & Rastle, K. (2017). Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *Journal of Memory and Language, 93*, 169-192.

Muncer, S. J., Knight, D., & Adams, J. W. (2014). Bigram frequency, number of syllables and morphemes and their effects on lexical decision and word naming. *Journal of Psycholinguistic Research, 43(3)*, 241-254.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*, 45-52.

Norris, D., & Kinoshita, S. (2012). Reading through a noisy channel: Why there's nothing special about the perception of orthography. *Psychological Review, 119(3)*, 517-545.

Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 24(3)*, 767-779.

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior Research Methods, 49(2)*, 407-417.

Pollatsek, A., Perea, M., & Binder, K. S. (1999). The effects of" neighborhood size" in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 25(4)*, 1142-1158.

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword reading: Comparing dual-route cascaded and connectionist dual-process models with human data. *Journal of Experimental Psychology: Human Perception and Performance, 38(5)*, 1268-1288.

Pynte, J., & Kennedy, A. (2006). An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. *Vision Research, 46(22)*, 3786-3801.

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in AdobeFlash and HTML5/JavaScript Web experiments. *Behavior Research Methods, 47(2)*, 309–327. http://doi.org/10.3758/s13428-014-0471-1

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.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111(1)*, 159-182.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46*, 245–266.

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor. *The Mental Lexicon, 12(2)*, 263-282.

Schmidtke, D., Kuperman, V., Gagné, C. L., & Spalding, T. L. (2016). Competition between conceptual relations affects compound recognition: the role of entropy. *Psychonomic Bulletin & Review, 23(2)*, 556-570.

Schröter, P., & Schroeder, S. (2017). The Developmental Lexicon Project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods, 49(6)*, 2183-2203.

Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood size and neighborhood frequency effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 876–900.

Seidenberg, M.S., & Waters, G.S. (1989). Word recognition and naming: A mega study. *Bulletin of the Psychonomic Society, 27*, 489.

Sereno, S. C., O'Donnell, P. J., & Sereno, M. E. (2009). Short article: Size matters: Bigger is faster. *Quarterly Journal of Experimental Psychology, 62(6)*, 1115-1122.

Soares, A. P., Lages, A., Silva, A., Comesaña, M., Sousa, I., Pinheiro, A. P., & Perea, M. (2019). Psycholinguistic variables in visual word recognition and pronunciation of European Portuguese words: a mega-study approach. Language, Cognition and Neuroscience.

Spieler D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8(6),* 411-416.

Sze, W. P., Liow, S. J. R., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods, 46(1)*, 263-273.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General, 124*, 107-136.

Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. W. F., Wang, S., & Chen, H. C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods, 50(5)*, 1763-1777.

Tse, C. S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods, 49(4)*, 1503-1519.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*.

Verhaeghen, P. (2003). Aging and vocabulary score: A meta-analysis. *Psychology and Aging, 18(2)*, 332-339.

Winsler, K., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2018). An electrophysiological megastudy of spoken word recognition. *Language, Cognition and Neuroscience, 33(8)*, 1063-1082.

Wulff, D. U., De Deyne, S., Jones, M. N., Austerweil, J. L., Baayen, R. H., Balota, D., … Mata, R. (in press). New Perspectives on the Aging Lexicon. *Trends in Cognitive Sciences*.

Yap, M.J., & Balota, D.A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*, 502-529.

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.

Yap, M. J., Balota, D. A., Tse, C. S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT

distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 495–513.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods, 42(4),* 992-1003.

Yarkoni, T., Balota, D.A., & Yap, M.J. (2008). Moving Beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971-979.

Yates, M. (2005). Phonological neighbors speed visual word processing: Evidence from multiple tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1385–1397.

Yates, M. (2009). Phonological neighborhood spread facilitates lexical decisions. *Quarterly Journal of Experimental Psychology, 62*, 1304–1314.

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review, 11*, 452–457.

Ziegler, J. C., & Perry, C. (1998). No more problems in Coltheart's neighborhood: Resolving neighborhood conflicts in the lexical decision task. *Cognition, 68(2)*, B53-B62.

Table 1: Word processing megastudies published so far, listed in chronological order for the various languages tested (limited to studies with 900 word types or more). An extended and updated table is available on http://crr.ugent.be/programs-data/megastudy-data-available with links to the studies and the data.

Name	Nstimuli	Presentation	Task	Reference
Chinese				
- No name given	2,423	Visual	Naming	Liu et al. (2007)
- Chinese Lexicon Project	2,500	Visual	Lexical decision	Sze et al. (2014)
- No Name given	3,423	Visual	Lexical decision	Lee et al. (2015)
- No name given	3,314	Visual	Naming	Chang et al. (2016)
- Chinese Lexicon Project	25,286	Visual	Lexical decision	Tse et al. (2017)
- MELD-SCH	12,578	Visual	Lexical decision	Tsang et al. (2018)
Dutch				
- Dutch Lexicon Project	14,089	Visual	Lexical decision	Keuleers et al. (2010)
- Baldey	2,780	Auditory	Lexical decision	Ernestus & Cutler (2015)
- Dutch Lexicon Project 2	30,016	Visual	Lexical decision	Brysbaert et al. (2016)
- Filling the gaps	8,240	Visual	Fragment completion	Heyman et al. (2016)
- GECO	5,575	Visual	Eye movements	Cop et al. (2017)
English				
- Mega study	2,900	Visual	Naming	Seidenberg & Waters (1989)
- No name given	1,327	Visual	Naming	Treiman et al. (1995)
- Word naming corpora	2,428	Visual	Naming (young adults)	Spieler & Balota (1997)
- Word naming corpora	2,428	Visual	Naming (old adults)	Balota & Spieler (1998)
- No name given	3,688	Visual	Naming	Kessler et al. (2002)
- No name given	1,000	Visual	Naming	Chateau & Jared (2003)
- Lexical decision corpora	2,428	Visual	Lexical decision (young adults)	Balota et al. (2004)
- Lexical decision corpora	2,428	Visual	Lexical decision (old adults)	Balota et al. (2004)
- Dundee corpus	9,776	Visual	Eye movements	Pynte & Kennedy (2006)
- English Lexicon Project	40,481	Visual	Lexical decision	Balota et al. (2007)
- English Lexicon Project	40,481	Visual	Naming	Balota et al. (2007)
- No name given	1,025	Visual	Progressive demasking	Lemhöfer et al. (2008)
- Recognition memory	2,578	Visual	Recognition	Cortese et al. (2010)
- British Lexicon Project	28,730	Visual	Lexical decision	Keuleers et al. (2012)
- Nonword naming	1,475	Visual	Naming	Pritchard et al. (2012)
- No name given	2,820	Visual	Naming	Adelman et al. (2013)
- Past tense inflection project	2,200	Visual	Word generation	Cohen-Shikora et al. (2013)
- No name given	1,524	Visual	Eye movements	Frank et al. (2013)

- No name given	1,524	Visual	Self-paced reading	Frank et al. (2013)
- Semantic priming project	1,661	Visual	Lexical decision	Hutchison et al. (2013)
- Semantic priming project	1,661	Visual	Naming	Hutchison et al. (2013)
- No name given	2,614	Visual	Naming	Cortese et al. (2015a)
- EEG study	960	Visual	Go/no-go	Dufau et al. (2015)
- EEG study	1,524	Visual	Reading	Frank et al. (2015)
- GECO	5,012	Visual	Eye movements	Cop et al. (2017)
- Response deadline	2,500	Visual	Naming	Cortese et al. (2017)
- Nonword naming	915	Visual	Naming	Mousikou et al. (2017)
- Calgary semantic decision project	10,000	Visual	Semantic decision	Pexman et al. (2017)
- Provo	1,197	Visual	Eye movements	Luke & Christianson (2018)
- Conditional naming	2,145	Visual	Naming	Cortese et al. (2018)
- Natural stories corpus	2,332	Visual	Self-paced reading	Futrell et al. (2018)
- ERP study	960	Auditory	Go/no-go	Winsler et al. (2018)
- SWR1081	1,081	Auditory	Perceptual identification	Liben-Nowell et al. (2019)
- MALD	26,793	Auditory	Lexical decision	Tucker et al. (2019)
French				
- Dundee corpus	11,321	Visual	Eye movements	Pynte & Kennedy (2006)
- French Lexicon Project	38,840	Visual	Lexical decision	Ferrand et al. (2010)
- Chronolex	1,482	Visual	Lexical decision	Ferrand et al. (2011)
- Chronolex	1,482	Visual	Naming	Ferrand et al. (2011)
- Chronolex	1,482	Visual	Progressive demasking	Ferrand et al. (2011)
- Megalex	28,466	Visual	Lexical decision	Ferrand et al. (2018)
- Megalex	17,876	Auditory	Lexical decision	Ferrand et al. (2018)
German				
- Potsdam sentence corpus	≈1,000	Visual	Eye movements	Kliegl et al. (2006)
- Developmental Lexicon Project	1,152	Visual	Lexical decision	Schröter & Schroeder (2017)
- Developmental Lexicon Project	1,152	Visual	Naming	Schröter & Schroeder (2017)
Hindi				
- No name given	≈1,000	Visual	Eye movements	Husein et al. (2015)
Malay				
- Malay Lexicon Project	9,592	Visual	Lexical decision	Yap et al. (2010)
Portuguese				
- No name given	1,920	Visual	Lexical decision	Soares et al. (2019)
- No name given	1,920	Visual	Naming	Soares et al. (2019)

Russian				
- Russian Sentence Corpus	≈1,000	Visual	Eye movements	Laurinavichyute et al. (2019)
Spanish				
- No name given	2,764	Visual	Naming	Davies et al. (2013)
- No name given	2,765	Visual	Lexical decision	González-Nosti et al. (2014)
- Spalex	45,389	Visual	Recognition	Aguasvivas et al. (2018)