Word prevalence norms for 62,000 English lemmas

Marc Brysbaert[1], Paweł Mandera[1], Samantha McCormick[2] & Emmanuel Keuleers[3]

[1] Department of Experimental Psychology, Ghent University

[2] Department of Psychology, University of Roehampton

[3] Department of Cognitive Science and Artificial Intelligence, Tilburg University

Keywords: Word prevalence, word frequency, word processing, megastudy

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 word prevalence data for 61,858 English words. Word prevalence refers to the number of people who know the word. The measure was obtained on the basis of an online crowdsourcing study involving over 220,000 people. Word prevalence data are useful to gauge the difficulty of words and, as such, are interesting to match stimulus materials in experimental conditions or to select stimulus materials for vocabulary tests. Word prevalence also predicts word processing times, over and above the effects of word frequency, word length, similarity to other words, and age of acquisition, in line with previous findings in the Dutch language.

Researchers working with word stimuli are taught to select words primarily on word frequency, word length, similarity to other words, and age of acquisition (e.g., Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, & Böhl, 2011). For instance, a researcher investigating the effect of emotional valence (negative, neutral, positive) on word processing efficiency would be expected to match the stimuli on those four variables.

In our work we gradually discovered that the above set of variables does not fully cover differences in word knowledge. This is particularly true for low frequency words. Some of these words are generally known (such as toolbar, screenshot, soulmate, uppercase, hoodie), whereas others are hardly known by anyone (e.g., scourage, thunk, whicker, or caudle). Furthermore, none of the other word variables that have been collected so far, seem to fully catch the differences in word knowledge.

For a long time, we hoped that improved word frequency measures would solve the problem, but so far this anticipation has not been met: Some words are much better known than expected from the frequency with which they are used in the corpora we have at our disposal to calculate word frequency measures. Subjective word familiarity ratings may be an alternative (Gernsbacher, 1984), but so far have not been collected for most of the words. In addition, such ratings can be criticized, because they are collected from a small number of people (who may be more or less familiar with some words for idiosyncratic reasons). In addition, there is a difference between how many people know a word and how familiar they are with the word. Some words score low on familiarity, yet are known to nearly everyone (such as basilisk, obelisk, oxymoron, debacle, emporium, and armadillo).

The variable that currently best seems to catch differences in word knowledge, is age of acquisition (AoA): Words that are not known to the raters get high AoA scores. Indeed, some researchers in natural language processing have started using AoA values as a proxy for word difficulty in addition to word frequency. However, this is not the common understanding of AoA, which is considered to be the order in which *known words* were acquired.

To solve the issue of differences in word knowledge unrelated to word frequency, we decided to directly ask people which words they knew. This was first done in Dutch (Brysbaert, Stevens, Mandera, and Keuleers, 2016a; Keuleers, Stevens, Mandera, & Brysbaert, 2015) and gave rise to a new word characteristic, which we called word prevalence. The variable refers to the percentage of people who indicate they know the word (in practice, the percentages are transformed to z-values; see below for more details). Word prevalence explained 6% extra variance in Dutch word processing times as measured with the lexical decision task. Even at the high end, it had an effect, as we observed a 20 ms difference in response times between words known to all participants and words known to only 99% of the participants (Brysbaert et al., 2016a).

The present article introduces the word prevalence measure for English and presents some of the initial analyses.

Method

Stimulus materials. The stimuli consisted of a list of 61,858 English words, collected over the years at the Center for Reading Research, Ghent University. The list is largely based on the SUBTLEX word frequencies we collected, combined with word lists from psycholinguistic experiments and word lists from freely available spelling checkers and dictionaries. The nonwords consisted of a list of 329,851 pseudowords generated by Wuggy (Keuleers & Brysbaert, 2010).

Participants and the vocabulary test used. For each vocabulary test, a random sample of 67 words and 33 nonwords was selected. For each letter string, participants had to indicate whether they knew the stimulus or not. At the end of the test, participants received information about their performance in the form of a vocabulary score based on the percentage of correctly identified words minus the percentage of nonwords identified as words. For instance, a participant who responded yes to 55 of the 67 words and to 2 of the 33 nonwords, received feedback that they knew 55/67 – 2/33 = 76% of the English vocabulary. Participants could do the test multiple times and always got a different sample of words and nonwords. The test was made available on a dedicated website (http://vocabulary.ugent.be/). Access to the test was unlimited. Participants were asked whether English was their native language, what their age and gender were, which country they came from, and which studies they had completed (see also Brysbaert, Stevens, Mandera, & Keuleers, 2016b; Keuleers et al., 2015). For the present purposes, we limited the analyses to the first three tests taken by native speakers of English from the USA and the UK.[1] All in all, we analyzed the data of 221,268 individuals who returned 265,346 sessions. Of these, 56% were completed by female participants and 44% by male participants.

Results

[1] Other countries with English as a native language did not (yet) produce enough observations to make reliable word prevalence estimates for them.

In the dataset we selected, each word was judged on average by 388 participants (282 from the USA and 106 from the UK). The percentages of people indicating they knew the word ranged from 2% (stotinka, adyta, kahikatea, gomuti, arseniuret, alsike, …) to 100% (…, you, young, yourself, zone, zoned). Figure 1 shows the distribution of percentages known. The vast majority of words were known to 90% or more of the participants.

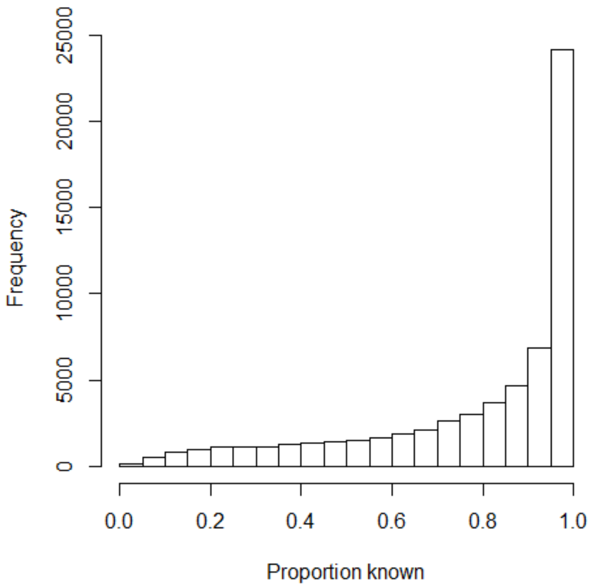

Figure 1: Distribution of the percentages of words known, showing that most words were known to 90% of the participants and more (see the rightmost two columns of the graph).

Because the distribution of percentages known is very right skewed and does not differentiate much between well-known words, it is useful to apply a probit transformation to the percentages (Brysbaert et al., 2016a). The probit function translates percentages known to z-values on the basis of the cumulative normal distribution. That is, a word known by 2.5% of the participants gets a word prevalence of -1.96; a word known by 97.5% of the participants gets a prevalence of +1.96. Because a percentage known of 0% would return a prevalence score of -∞ and a percentage known of 100% a prevalence score of +∞, the range was reduced to percentages known .5% (prevalence = - 2.576) and 99.5% (prevalence = +2.576).[2] Figure 2 shows the distribution of prevalence scores for the total list of words.

[2] The specific formula to use in Microsoft Excel was =NORM.INV(0.005+Pknown*0.99;0;1).

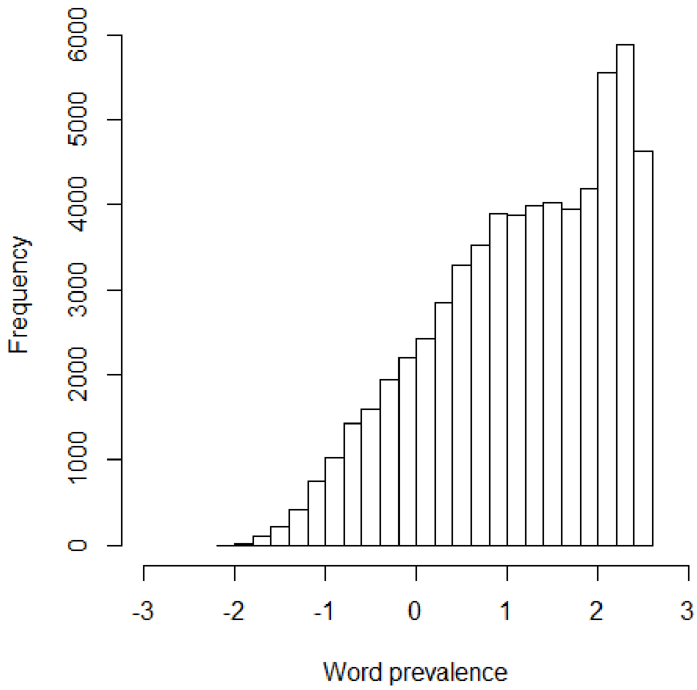

Figure 2: Distribution of word prevalence scores

Word prevalence has negative values for words known to less than 50% of the participants. This may be confusing at first sight, but is rather informative. All words with negative prevalence scores are uninteresting for experiments with RTs (because these words are not known well enough), but they are interesting for word learning experiments and experiments capitalizing on differences in accuracy.

Although the US word prevalence and the UK prevalence scores correlate r = .93 with each other, there are a few words that differ in prevalence between both countries, due to cultural differences. Table 1 gives a list of the extreme cases. If researchers want to collect or analyze data from one country only, it may be an idea to exclude the deviating words or to use country-specific word prevalence data.

- -

Insert Table 1 about here

- -

Similarly, although the word prevalence scores correlate r = .97 between men and women, some words deviate, as can be seen in Table 2. These tend to follow gender differences in interests (games, weapons and technical matters for males; food, clothing and flowers for females). The high correlations between the US and the UK measures and between males and females indicate that the reliability of the prevalence measure is very high (with .93 as the lower limit).

- -

Insert Table 2 about here

- -

Uses of the word prevalence measure

Word prevalence as a predictor variable

By its nature, word prevalence will be a good predictor of word difficulty. Experimenters interested in word processing times naturally want to avoid stimuli that are unknown to many of the participants. This can now easily be achieved, by only using words with percentage known of 95% and more (prevalence of 1.60 and more). Similarly, word prevalence can be used as an estimate of word difficulty for vocabulary tests. By ordering the words according to word prevalence (and word frequency) it is possible to delineate word difficulty bands, which can be used to select stimuli from.

Word prevalence is also likely to be of interest to natural language processing (NLP) researchers writing algorithms to gauge the difficulty of texts. At present, word frequency is used as a proxy of word difficulty (e.g., Benjamin, 2012; De Clercq & Hoste, 2016; Hancke, Vajjala, & Meurers, 2012). Word prevalence is likely to be a better measure, given that it does not completely reduce to differences in word frequency.

Finally, word prevalence can be used to predict differences in word processing efficiency. In recent years, researchers have started to collect reaction times (RTs) to thousands of words and tried to predict RTs on the basis of word characteristics. Table 3 gives an overview of the word characteristics included in the analyses and references to some of the articles involved.

- -

Insert Table 3 about here

- -

Although many variables have been examined, most of them account for less than 1% of the variance in word processing times, once the effects of word frequency, word length (letters), similarity to other words (OLD20), and age of acquisition are partialled out. Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, & Böhl (2011), for example, analyzed the lexical decision times provided by the English Lexicon Project (ELP; Balota et al. , 2007), using the 20+ word characteristics included in ELP as predictors. The three most important variables (word frequency, similarity to other words, and word length) together accounted for 40.5% of the variance. The remaining variables together accounted for only 2% extra variance. Indeed, our work over the last years has shown that the objective of explaining as much variance as possible in word processing times is better served by looking for improved word frequency measures than by searching for new variables or interactions between variables. At the same time, we do not appear to have found all possible sources of variation yet (see also Adelman, Marquis, Sabatos-DeVito, & Estes, 2013). The systematic variance to be accounted for in megastudies is typically larger than 80% (as estimated on the basis of the reliability of the scores).

To examine whether word prevalence is a variable that substantially increases the percentage of variance explained in word processing times, we repeated the analysis of Brysbaert et al. (2011) on the ELP lexical decision times and additionally included age of acquisition and word prevalence as predictors. The variables we included were:

- Word frequency based on the SUBTLEX-US corpus (Brysbaert & New, 2009) and expressed as Zipf scores (Brysbaert, Mandera, & Keuleers, 2018; Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The Zipf score is a standardized log-transformed measure of word frequency that is easy to understand (words with a Zipf score of 1-3 can be considered low-frequency words; words with a Zipf score of 4-7 can be considered high-frequency).
- Word length in number of letters.
- Number of orthographic neighbors (words formed by changing 1 letter; information obtained from ELP).
- Number of phonological neighbors (words formed by changing one phoneme; from ELP).
- Orthographic Levenshtein Distance (from ELP).
- Phonological Levenshtein Distance (from ELP).
- Number of phonemes (from ELP).
- Number of syllables (from ELP).
- Number of morphemes (from ELP).

- Age of acquisition (AoA; from Kuperman et al., 2012; lemma values applied to inflected forms).
- Word prevalence.

We took the prevalence of an inflected form to be the same as that of its lemma in case the inflected form was not in the database. As we were interested in RTs, only words with 75% accuracy or more in the ELP lexical decision task were included. In our analyses, we used the z-scores of participants' RTs, rather than their absolute RTs, which eliminates variance in RTs due to participants being faster or slower than average. The percentage of variance in RTs that can be accounted for is substantially higher for z-scores than for raw RTs (as shown below, where the percentages of variance accounted for are substantially higher than the 43% reported by Brysbaert et al., 2011). In total we had complete data for 25,661 words. We analyzed both the ELP lexical decision times and the ELP naming latencies. Table 4 shows the correlations between the variables.

- -

Insert Table 4 about here

- -

Table 4 illustrates the high correlations observed between the different word characteristics. In this respect, word prevalence comes out well because it is rather unrelated to the variables associated with word length. In addition, the correlation with frequency is rather limited ($r = .487$). This is higher than the value observed in the Dutch analyses of Brysbaert et al. (2016a), probably because the words from ELP were selected on the basis of a word frequency list. This means that known words with a frequency of 0 in the corpus were excluded.

One way to find the relevant predictors for the word processing times is to run a hierarchical regression analysis. As we are particularly interested in the added value of word prevalence, we first entered all the other variables and then word prevalence. To take into account non-linearities, the regression analysis included polynomials of the second degree for word frequency, word length, AoA, and prevalence. Because the number of phonological neighbors and the number of phonemes were highly correlated with other variables and did not alter the picture, they were left out of the analysis.

- -

Insert Table 5 about here

- -

When we entered all variables except for prevalence, we explained 66.2% of the variance in the z-values of the lexical decision times (Table 5). When prevalence was added, we explained 69.8% of the variance. Figure 3 shows the effects of the various variables.

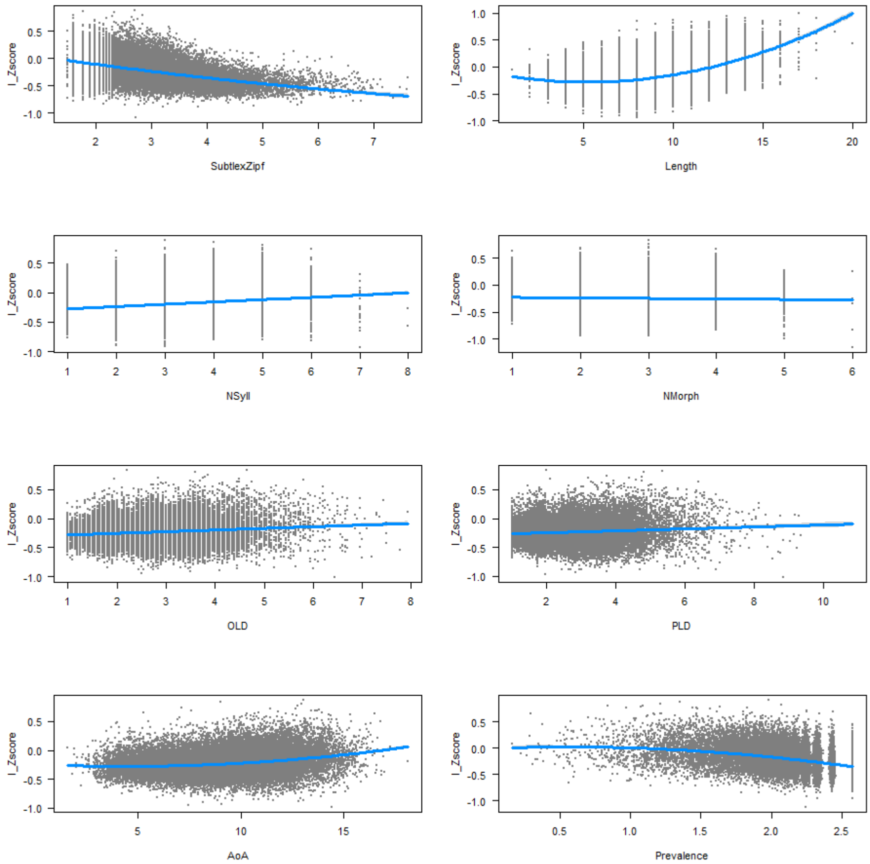

Figure 3: Effects of the variables on the standardized ELP lexical decision times. First line: effects of word frequency and length in letters; second line: number of syllables and number of morphemes; third line: orthographic and phonological similarity to other words; last line: Age of acquisition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

The results agree with what was found for Dutch. High frequency words are processed faster than low frequency words. Interestingly, when prevalence is added, the relation becomes linear, whereas before there was a floor effect for high frequency words. Words with 6-8 letters are responded to fastest. In addition, response times grow when the words contain more syllables, but tend to decrease for morphologically complex words when all the other variables are taken into account. Words that are similar in sound and spelling to many other words (i.e, words with low OLD and PLD values) are responded to faster. Words were responded to more slowly when they were acquired late. And finally, there is a robust effect of word prevalence. Interestingly, the effect is strongest at the high end, when all other variables have been accounted for. The effect is rather flat for words with a prevalence rate below 1.2 (which agrees with percentage known of 89%).

Table 5 and Figure 4 show that the effects were very similar for word naming, but that the contribution of word prevalence was smaller than for lexical decision times (though still highly significant).

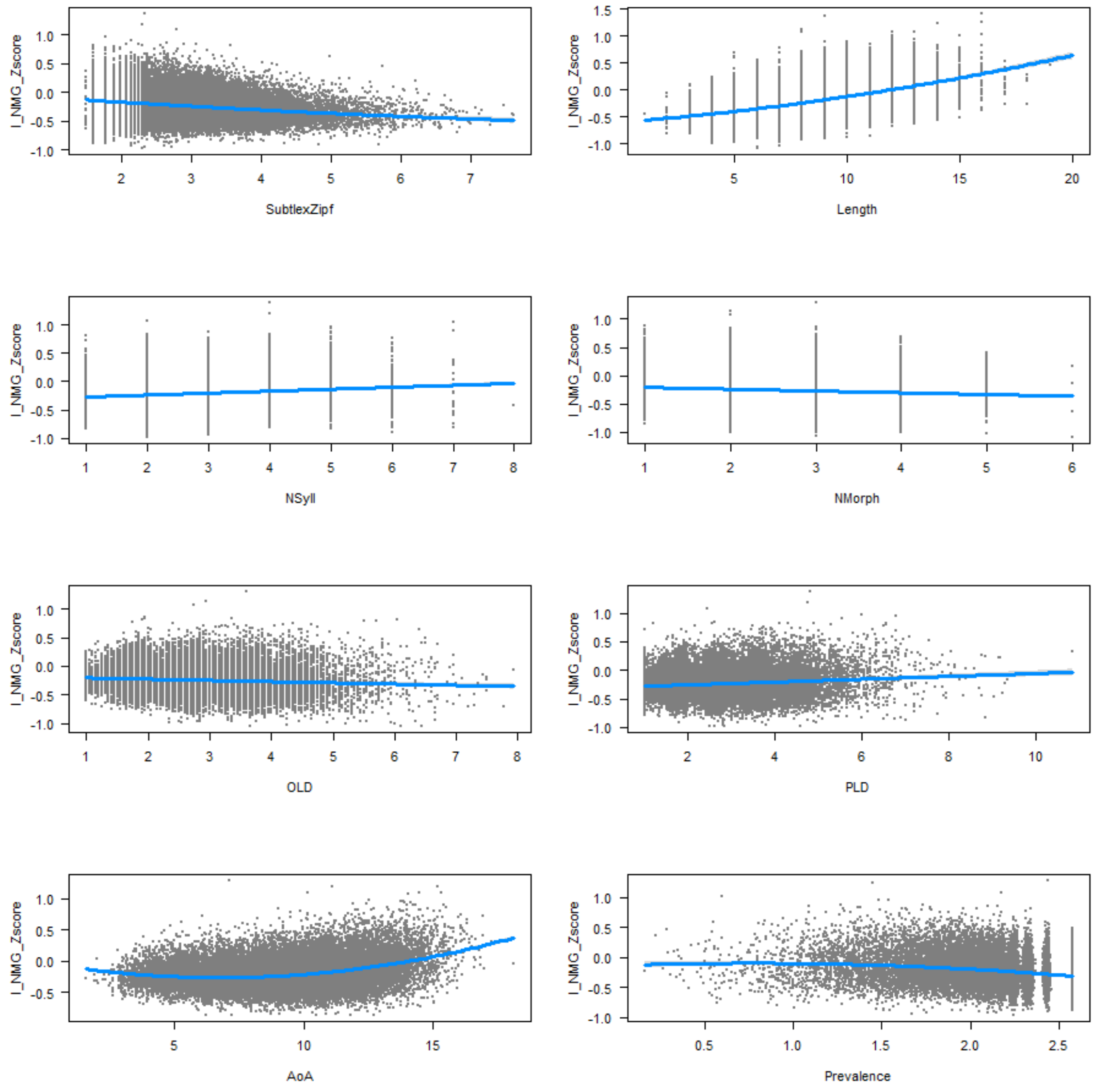

Figure 4: Effects of the variables on the standardized ELP naming times. First line: effects of word frequency and length in letters; second line: number of syllables and number of morphemes; third line: orthographic and phonological similarity to other words; last line: Age of acquisition and word prevalence.

Relation to other word characteristics

Kuperman et al. (2012) used a dataset of Cark and Paivio (2004) to gauge the relationship of AoA to 30+ other word features. We used the same dataset and added word prevalence to it, together with values from ELP, the concreteness ratings of Brysbaert, Warriner, & Kuperman (2014), and estimates of word valence, dominance and arousal collected by Hollis, Westbury, and Lefsrud (2016). Kuperman et al. (2012) found an 8-factor solution to best fit the data. We used the same structure, but in addition allowed the factors to intercorrelate (using the fa() function from the R package psych; Revelle, 2018). This resulted in a solution that was more straightforward to interpret. There were 907 words for which we had all measures. Table 6 shows the outcome.

- -

Insert Table 6 about here

- -

Word prevalence loads on the same factor as word accuracy of ELP and various ratings of familiarity. There is a second factor, word frequency, that is correlated r = .66 with the first factor. The other factors refer to the similarity to other words, word length, affect (valence and arousal), and gender ladenness of the words. The factor word frequency correlates with the factors similarity to other words (r = .41), length (r = -.34), and valence (r = .30). Similarity to other words in addition correlates with length (r = -.45). Valence also correlates with gender ladenness (r = .40). All other correlations are below r = .3 (absolute values).

All in all, when we analyze the word attributes collected by Clark and Paivio (2004) and add the ones collected since, we see that the features reduce to eight main word characteristics. Word prevalence loads on a factor together with other measures of word familiarity. The factor is correlated with word frequency as observed in various corpora. The word processing measures of ELP also load on the prevalence/familiarity factor, in line with the impact of prevalence on word processing times we saw above. On the other hand, the fact that the accuracy data of the ELP lexical decision experiment had the highest load raise the question to what extent the factor measures word knowledge or the decision element in the yes/no vocabulary task. If word prevalence is related to the decision component, one would expect it to only correlate with lexical decision times and not with word processing times in other tasks (e.g., eye movements in reading). To some extent, the worry is contradicted by the ELP naming data (Figure 4), but more research is clearly needed to establish to what extent word prevalence is a true word feature (independent variable) or a word processing characteristic (dependent variable). Notice that a similar question has been raised about word frequency: Whether it should be considered as an independent variable or a dependent variable (Baayen, Milin, & Ramscar, 2016). Other evidence that word prevalence is related to word knowledge (i.e., is an independent variable) can be found in the country and gender differences (Tables 1 and 2) and in the age differences observed (Brysbaert et al., 2016b; Keuleers et al., 2015). These seem more related to differences in word knowledge than in decision processes.

Word prevalence as a matching variable

In many studies, a new word attribute is the variable of interest. In such studies, the stimuli in the various conditions must be matched on word frequency, word length, orthographic similarity to other words, and age of acquisition. Even with this set of criteria, there is evidence that researchers can select stimuli in such a way that they increase the chances of observing the hypothesized effect (i.e., show an experimenter bias; Forster, 2000; Kuperman, 2015). We think word prevalence will be an important variable to correct for this bias. Table 7 shows words with different percentages known matched on frequency (Zipf = 1.59, meaning the words were observed only once in the SUBTLEX-US corpus of 51 million words). The various words clearly illustrate the danger of experimenter bias when word prevalence is not taken into account.

- -

Insert Table 7 about here

- -

As can be seen in Figures 3 and 4, matching words on prevalence is not only needed for words with very divergent prevalence scores, but also for words with high prevalence scores, something that cannot be achieved without the present dataset.

Word prevalence as a dependent variable

A final set of studies for which word prevalence will be interesting, relates to the question what causes differences in prevalence rates. As we have seen above, familiarity and word frequency are important variables, but not the only ones. Which other variables are involved?

The best way to answer this question is to examine the divergences between word prevalence and word frequency. Which words are more widely known than expected on the basis of their frequency, and which words are less well known than expected on the basis of their frequency? As for the former question, it is striking that many well-known words with low frequencies are morphologically complex words. The best known very low frequency words with a frequency of Zipf = 1.59 are "binocular, distinctively, reusable, gingerly, preconditioned, legalization, distinctiveness, inaccurately, localize, resize, pitfall, unsweetened, unsaturated, undersize, compulsiveness", all words derived from simpler stems. Another set of words with frequencies less than predicted, are words mainly used at a young age, such as grandma (AoA = 2.6 yrs; prevalence = 2.4, frequency = 4.7), potty (AoA = 2.7 yrs; prevalence = 1.9, frequency = 3.2), yummy (AoA = 2.9 yrs; prevalence = 2.1, frequency = 3.7), nap (AoA = 3.0 yrs; prevalence = 2.3, frequency = 4.1), or unicorn (AoA = 4.8 yrs; prevalence = 2.6, frequency = 3.4). Also words that denote utensils are often known more widely than expected on the basis of their frequency, such a hinge (AoA = 8.6 yrs; prevalence = 2.2, frequency = 2.2), sanitizer (AoA = 10.9 yrs; prevalence = 2.1, frequency = 1.6), or wiper (AoA = 8.4 yrs; prevalence = 2.3, frequency = 2.8).

Finally, the prevalence measure itself is likely to be of interest. One may want to investigate, for instance, to what extent prevalence scores depend on the way in which they were defined. Goulden, Nation, and Read (1990) presented students with 250 lemmas taken at random from a dictionary and tested them in the same way as we did (i.e., students had to indicate which words they knew). Students selected on average 80 words. Milton and Treffers-Daller (2013) used the same words but asked participants to give a synonym or explanation for each word they knew. Now students were correct on 45 words only. Two questions are important: (1) how strong is the correlation between both estimates of word knowledge, and (2) which measure best captures "word knowledge"?

As for the first question, Paul, Stallman, and O'Rourke (1990) reported high correlations between the yes/no test and tests involving interviews and multiple choice questions. Surprisingly, no other studies on this topic could be found with native speakers (there are more studies with second language speakers, which largely – but not always – confirm the finding that the yes/no test correlates well with other test formats). In addition, all studies looked at correlations across participants and not at correlations across items (given that the interest was in assessing the language proficiency of participants, not knowledge of individual words). In order to obtain some more information, we presented three existing English multiple-choice vocabulary tests to 248 first-year psychology students at a British University.[3] The three tests were Mill Hill Form 2 (Raven, 1958;

[3] We are grateful to Dr. Joe Levy for his help in designing and running the experiments.

34 words), the Shipley test (Shipley, 1940; 40 words), and a TOEFL test (Landauer & Dumais, 1997; 80 words). When we correlated the scores on the items with the word prevalence measures, we obtained a correlation of r = .69 (N = 174), which is shown in Figure 5.

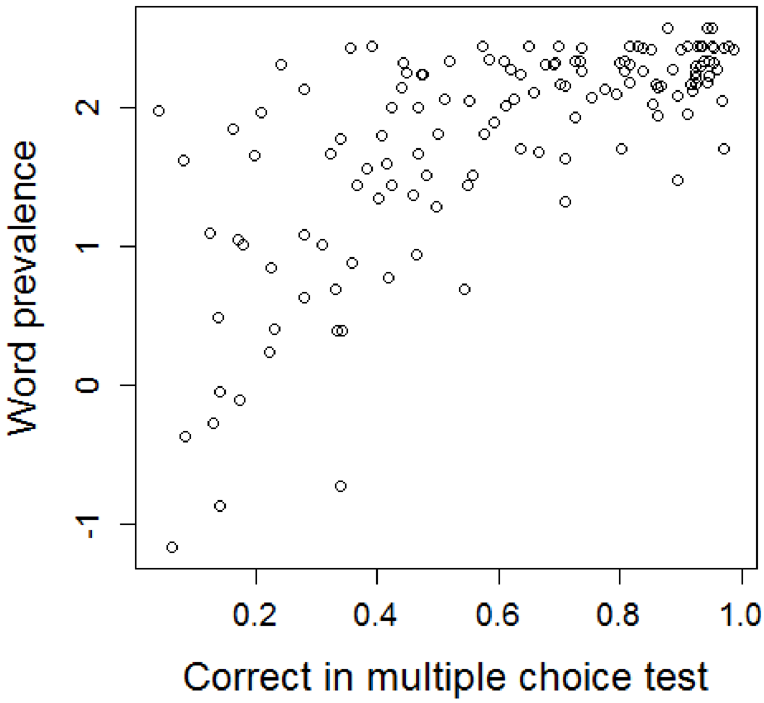

Figure 5: Correlation between word prevalence and probability correct on a multiple choice vocabulary test.

Because the correlation is lower than hoped for, we may want to have a look at the outliers. The upper left outlier (correct on MC test = .04, prevalence = 1.98) is the word “sultry”, an item from the Mill Hill test. According to our test, 98% of the people claim to know the word, whereas the Mill Hill test suggests that no-one really knows the meaning. If we look in a dictionary, “sultry” has two meanings: (1) hot and humid, and (2) displaying or arousing sexual desire. If we look at semantic vectors (Mandera, Keuleers, & Brysbaert, 2017), the closest synonyms are “breathy, steamy, songstress, hot, sexy, alluring, languid”. The first associates given by humans are “sexy, hot, humid, steamy, woman, warm, seductive” (De Deyne, Navarro, Perfors, & Storms, 2016). However, none of these words are among the options available in the Mill Hill: Participants have to choose between “instinctive, sulky, trivial, solid, severe, muggy”. No surprise then that no-one knows the intended meaning. Another word in the upper left corner comes from the Shipley test: “pristine” (correct on MC test = .24, prevalence = 2.3). The alternatives given in the Shipley test are “vain, sound, first, level”, rather than one of the expected associates “clean, pure, or perfect”. On the right side of Figure 5, we find the word “easygoing” from the TOEFL test (correct on MC test = .97, prevalence = 1.7). In all likelihood, the low prevalence score for this word reflects the fact that many people do not consider easygoing as a correct English spelling (they arguably prefer the two-word expression “easy going”). A similar reasoning explains why the word “impostor” is doing worse on word prevalence (1.5) than on the Shipley test (0.9). Currently, the preferred spelling of the word is “imposter” (which has a prevalence score of 2.2).

The deviations between the multiple-choice scores and word prevalence bring us to the second question: Which measure best captures “word knowledge”? As we have seen, answers to multiple choice questions (the most frequent way of testing vocabulary) not only depend on knowledge of the target word but also on the alternatives presented. If they test a rare (or outdated) meaning of a word, they can easily lead to low scores for that word (remember that test makers are not interested in the scores on individual items; they are interested in average scores of individuals). On the other hand, word prevalence scores are affected by the spelling of the word and only give information about the most familiar meaning. Which is the “best” way of testing word knowledge? Although one might be tempted to think that deeper knowledge is better, it may be that hazy knowledge is what we use most of the time when we are reading text or hearing discourse. Indeed, it might be argued that no person, except for specialized lexicographers, know the full meaning of the words they are using (Anderson & Freebody, 1981). Still, it would be good to have more information on the relationship between results based on the yes/no format used here and other test formats. In particular, correlations over items are important.

Availability

We made an Excel file with the Pknown and Prevalence values for the 61,858 words tested. Most words are lemmas (i.e., without inflections). An exception was made for common irregular forms (e.g., lice, went, wept, were) and nouns that had a different meaning in plural than in singular (glasses, aliens). The file also includes the SUBTLEX-US word frequencies, expressed as Zipf scores. Figure 6 gives a snapshot of the file.

The file further contains sheets with the differences between UK and US respondents, and between male and female respondents, so that readers can make use of this information if they want to do so.

Finally, we make the databases available that were used for the various analyses reported in the present article, so that readers can check them and, if desired, improve on them. These files are available as supplementary materials and can also be found at https://osf.io/g4xrt/.

	A	B	C	D	E
1	**Word**	**Pknown**	**Nobs**	**Prevalence**	**FreqZipfUS**
2	a	0.98	438	1.917	7.309
3	aardvark	0.96	434	1.684	2.634
4	aardwolf	0.21	428	-0.788	1.292
5	abaca	0.24	396	-0.706	1.593
6	aback	0.86	343	1.077	2.496
7	abacus	0.93	401	1.428	2.406
8	abaft	0.19	363	-0.876	1.769
9	abalone	0.69	383	0.496	2.723
10	abandon	1.00	378	2.427	3.909
11	abandoned	1.00	401	2.576	4.124
12	abandonee	0.66	362	0.409	1.292
13	abandoner	0.86	404	1.081	1.593
14	abandonment	0.99	419	2.185	2.991
15	abase	0.75	420	0.667	1.894

Figure 6: Snapshot of the data file with word prevalences, available as supplementary materials or at https://osf.io/g4xrt/.

References

Adelman, J. S., & Brown, G. D. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review, 14(3)*, 455-459.

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17(9)*, 814-823.

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.

Anderson, R.C., & Freebody, P. (1981). Vocabulary knowledge. In Guthrie, J. (Ed.), *Reading comprehension and education* (pp. 77-117). Newark, DE: International Reading Association.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology, 30(11)*, 1174-1220.

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

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24(1)*, 63-88.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*, 412-424.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science, 27*. DOI: 10.1177/0963721417727521

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44(4),* 991-997.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016a). 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., Stevens, M., Mandera, P., & Keuleers, E. (2016b) 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., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904-911.

Coltheart, M. (1981), The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition, 125(3),* 452-465.

Cortese, M. J., Hacker, S., Schock, J., & Santo, J. B. (2015). Is reading-aloud performance in megastudies systematically influenced by the list context? *Quarterly Journal of Experimental Psychology, 68(8)*, 1711-1722.

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*. https://doi.org/10.1177/1747021817741269

De Clercq, O., & Hoste, V. (2016). All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics, 42(3),* 457-490.

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General, 145(9)*, 1228.

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. *Quarterly Journal of Experimental Psychology, 68(8)*, 1469-1488.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., ... & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from Chronolex. *Frontiers in psychology, 2: 306*.

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*. https://doi.org/10.3758/s13428-017-0943-1

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition, 28(7)*, 1109-1115.

Gernsbacher, M. A. (1984). Resolving 20 Years of Inconsistent Interactions Between Lexical Familiarity and Orthography, Concreteness, and Polysemy. *Journal of Experimental Psychology. General, 113(2)*, 256–281.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11(4)*, 341-363.

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability Classification for German using Lexical, Syntactic, and Morphological Features. *Proceedings of COLING 2012*, 1063-1080.

Hollis, G., Westbury, C., & Lefsrud, L. (2016). Extrapolating Human Judgments from Skip-gram Vector Representations of Word Meaning. *The Quarterly Journal of Experimental Psychology, 70(8)*, 1603-1619.

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

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *Quarterly Journal of Experimental Psychology, 68(8)*, 1693-1710.

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(3)*, 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.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104(2)*, 211-240.

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

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57-78

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review, 4(1)*, 151-172.

Paul, P. V., Stallman, A. C. and O'Rourke, J. P. (1990). *Using three test formats to assess good and poor readers' word knowledge.* Technical Report No. 509, Center for the Study of Reading, University of Illinois at Urbana-Champaign, IL.

Raven, J. C. (1958). *Guide to using the Mill Hill Vocabulary Scale with the Progressive Matrices Scales.* Oxford, England: H. K. Lewis & Co.

Revelle, W. (2018). Package 'psych'. Available on May 29, 2018 at https://cran.r-project.org/web/packages/psych/psych.pdf.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology, 9(2)*, 371-377.

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.

Sze, W. P., Yap, M. J., & Rickard Liow, S. J. (2015). The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. *Quarterly Journal of Experimental Psychology, 68(8)*, 1541-1570.

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*. https://doi.org/10.3758/s13428-017-0944-0

Tse, C. S., & Yap, M. J. (2018). The role of lexical variables in the visual recognition of two-character *Chinese compound words: A megastudy analysis. Quarterly Journal of Experimental Psychology, 71.* https://doi.org/10.1177/1747021817738965.

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.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176-1190.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60(4)*, 502-529.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review, 18(4)*, 742-750.

Table 1: Words much better known in the US than in the UK (left) and vice versa (right)

Word	Pus	Puk
manicotti	0.90	0.16
ziti	0.81	0.08
tilapia	0.93	0.20
garbanzo	0.92	0.21
kabob	0.98	0.28
kwanza	0.90	0.22
crawdad	0.86	0.20
hibachi	0.90	0.26
sandlot	0.95	0.32
acetaminophen	0.93	0.33
tamale	0.91	0.32
kielbasa	0.84	0.24
conniption	0.76	0.17
chigger	0.80	0.22
tomatillo	0.80	0.22
provolone	0.97	0.40
albuterol	0.74	0.16
staph	0.85	0.28
goober	0.97	0.40
luau	0.83	0.26

Word	Pus	Puk
tippex	0.07	0.91
biro	0.16	0.99
tombola	0.17	0.97
chipolata	0.16	0.94
dodgem	0.18	0.95
yob	0.21	0.98
gazump	0.05	0.82
abseil	0.14	0.89
naff	0.19	0.94
kerbside	0.23	0.98
plaice	0.16	0.91
judder	0.19	0.94
chiropody	0.19	0.94
korma	0.21	0.95
bolshy	0.11	0.85
quango	0.08	0.82
pelmet	0.11	0.85
brolly	0.24	0.96
chaffinch	0.12	0.85
escalope	0.19	0.91

Table 2: Words better known by males than by females (left) and vice versa (right)

Word	P_Male	P_Female
howitzer	0.84	0.53
thermistor	0.48	0.17
azimuth	0.58	0.27
femtosecond	0.47	0.15
milliamp	0.69	0.37
aileron	0.55	0.22
servo	0.61	0.28
degauss	0.59	0.26
boson	0.76	0.44
checksum	0.58	0.25
piezoelectricity	0.51	0.18
gauss	0.64	0.31
katana	0.80	0.47
shemale	0.88	0.54
neodymium	0.56	0.21
yakuza	0.69	0.32
teraflop	0.58	0.22
strafe	0.83	0.46
parsec	0.83	0.44
bushido	0.60	0.21

Word	P_Male	P_Female
peplum	0.13	0.64
tulle	0.27	0.77
chignon	0.24	0.72
bandeau	0.35	0.81
freesia	0.27	0.72
chenille	0.34	0.76
kohl	0.36	0.77
verbena	0.30	0.70
doula	0.21	0.59
ruche	0.18	0.55
espadrille	0.36	0.73
damask	0.43	0.80
jacquard	0.39	0.74
whipstitch	0.37	0.71
boucle	0.16	0.50
taffeta	0.53	0.87
sateen	0.38	0.72
chambray	0.43	0.77
pessary	0.19	0.53
voile	0.34	0.68

Table 3: Variables investigated in word processing megastudies that correlate with response times. For each variable an exemplary study is given in which the variable was examined (LDT = Lexical Decision Time).

	Chinese LDT	Chinese Naming	Dutch LDT	English LDT	English Naming	French LDT	French Naming	German LDT	German Naming
Word frequency	√ [18]	√ [14]	√ [4]	√ [20]	√ [20]	√ [11]	√ [10]	√ [15]	√ [15]
Word length (N letters)	√ [17]		√ [4]	√ [20]	√ [20]	√ [11]	√ [10]	√ [15]	√ [15]
Age of acquisition	√ [16]	√ [14]	√ [4]	√ [7]	√ [7]	√ [10]	√ [10]		
Concreteness/imageability	√ [16]	√ [14]	√ [4]	√ [8]	√ [7]	√ [10]	√ [10]		
Orthographic similarity to other words			√ [4]	√ [20]	√ [20]	√ [11]	√ [10]	√ [15]	√ [15]
Phonological similarity to other words	√ [19]				√ [2]	√ [10]	√ [10]		
Word length (N phonemes)				√ [20]	√ [20]	√ [11]	√ [10]		
First phoneme				√ [20]	√ [20]	√ [10]	√ [10]		
Visual complexity	√ [18]	√ [14]		√ [8]					
Semantic richness	√ [16]	√ [14]		√ [21]					
Contextual diversity	√ [18]			√ [1]	√ [1]				
Phonological consistency	√ [19]			√ [20]	√ [20]				
Word length (N syllables)			√ [4]	√ [20]	√ [20]				
Phonological uniqueness point			√ [9]			√ [11]			
Part of speech			√ [4]	√ [3]					
Homophone density	√ [19]	√ [14]							
Valence and arousal				√ [13]	√ [13]				
Number of senses				√ [20]	√ [20]				
Semantic neighborhood size				√ [20]	√ [20]				
Perceptual strength				√ [5]	√ [5]				
Sensory experience				√ [12]	√ [12]				
Stress pattern				√ [12]	√ [12]				
Orthographic uniqueness point						√ [11]			

Semantic transparency	√ [18]		
Pronunciation ambiguity	√ [17]		
Bigram frequency		√ [8]	
Consonant vowel proportion		√ [8]	
List context			√ [6]

[1] Adelman et al. (2006), [2] Adelman & Brown (2007), [3] Brysbaert et al. (2012), [4] Brysbaert et al. (2016), [5] Connell & Lynott (2012), [6] Cortese et al. (2015), [7] Cortese et al. (2018), [8] Dufau et al. (2015), [9] Ernestus & Cutler (2015), [10] Ferrand et al. (2011), [11] Ferrand et al. (2018), [12] Juhasz & Yap (2013), [13] Kuperman et al. (2014), [14] Liu et al. (2007), [15] Schröter & Schroeder (2017), [16] Sze et al. (2015), [17] Tsang et al., (2018), [18] Tse et al. (2017), [19] Tse & Yap (2018), [20] Yap & Balota (2009), [21] Yap et al. (2011)

Table 4: Correlations between the ELP variables and word prevalence (N = 25,661)

	Zipf	Ortho_N	Phono_N	OLD	PLD	NPhon	NSyll	NMorph	AoA	Preval	I_Zscore	I_NMG_Zscore
Length	-0.471	-0.570	-0.574	0.869	0.841	0.916	0.830	0.696	0.476	-0.150	0.654	0.627
SubtlexZipf		0.374	0.408	-0.443	-0.445	-0.451	-0.386	-0.427	-0.561	0.487	-0.649	-0.522
Ortho_N			0.810	-0.592	-0.536	-0.531	-0.495	-0.363	-0.380	0.128	-0.374	-0.379
Phono_N				-0.564	-0.580	-0.586	-0.522	-0.390	-0.393	0.128	-0.383	-0.375
OLD					0.912	0.817	0.738	0.542	0.471	-0.230	0.647	0.587
PLD						0.872	0.792	0.567	0.491	-0.224	0.650	0.599
NPhon							0.860	0.664	0.509	-0.136	0.636	0.629
NSyll								0.606	0.516	-0.151	0.614	0.591
NMorph									0.308	-0.065	0.458	0.411
AoA										-0.425	0.603	0.560
Prevalence											-0.512	-0.392
I_Zscore												0.753

Zipf = log word frequency based on SUBTLEX-US (Brysbaert & New, 2009), AoA = age of acquisition (Kuperman, Warriner, & Brysbaert, 2012), I_Zscore = RT in the ELP lexical decision task, I_NMG_Zscore = RT in the ELP naming task. All other variables are explained in the text and come from the ELP website (Balota et al., 2007).

Table 5: Variance explained in the ELP data

	R^2
Lexical decision times	
Frequency + Length + AoA + Nsyll + Nmorph + OLD + PLD	.662
Frequency + Length + AoA + Nsyll + Nmorph + OLD + PLD + Prevalence	.698
Naming latencies	
Frequency + Length + AoA + Nsyll + Nmorph + OLD + PLD	.539
Frequency + Length + AoA + Nsyll + Nmorph + OLD + PLD + Prevalence	.552

Table 6: Outcome of factor analysis on the word features collected by Clark & Paivio (2004; N = 907). Also included are features from ELP (Balota et al., 2007), Subtlex (Brysbaert & New, 2009), AoA (Kuperman et al., 2012), concreteness (Brysbaert et al., 2014), and valence, dominance, arousal (Hollis et al., 2016). The analysis shows that word prevalence loads on the same factor as familiarity (the prevalence factor is correlated r = .65 with the second factor – frequency). The second last column indicates how much of the variance in the variable is explained by their factor loading; the last column indicates the proportion of variance not explained by the factors.

	Prevalence	Frequency	Concreteness	Similarity	Length	Valence	Arousal	Gender	h2	u2
Lexical decision accuracy	0.92								0.77	0.23
Familiarity rating 2	0.90								0.95	0.05
Word prevalence	0.84								0.74	0.26
Context availability	0.84								0.90	0.10
Ease of definition (estimation)	0.67		0.36						0.80	0.20
Pronounceability	0.63								0.72	0.28
Lexical decision time	-0.54				0.34				0.80	0.20
Naming time	-0.53				0.38				0.64	0.36
Familiarity rating 1	0.52	0.42							0.81	0.19
Frequency SUBTLEX-US		0.97							0.99	0.01
Contextual diversity SUBTLEX-US		0.96							0.99	0.01
Kucera-Francis frequency		0.86							0.76	0.24
Frequency HAL		0.76							0.82	0.18
Thorndike-Lorge frequency		0.66							0.71	0.29
Word availability (used in dictionary)		0.64							0.58	0.42
Word availability (produced as associate)		0.58							0.72	0.28
Numer of childhood dictionaries		0.58	0.44						0.70	0.30
Concreteness rating Brysbaert			0.91						0.91	0.09
Concreteness rating			0.91						0.90	0.10
Imaginability rating			0.91						0.88	0.12
Imaginability rating 2	0.42		0.67						0.89	0.11
Meaningfulness (number of associates produced)			0.62						0.59	0.41
Age of acquisition			-0.45						0.75	0.25
age of acquisition Kuperman	-0.33		-0.39						0.76	0.24
Ambiguity word (estimation)			-0.37						0.31	0.69
Number of rhyming words				0.85					0.76	0.24
Number of similar looking words				0.80					0.70	0.30
Number of similar sounding words				0.79					0.85	0.15
Number orthographic neighbors				0.77					0.72	0.28
Number phonological neighbors				0.72					0.72	0.28
Words with same initial sounds				0.56	0.42				0.38	0.62
Number of letters					0.82				0.90	0.10
Number of phonemes					0.78				0.89	0.11
Number of syllables					0.65				0.75	0.25
Number of morphemes					0.64				0.56	0.44
PLD20				-0.39	0.62				0.83	0.17
OLD20				-0.47	0.55				0.83	0.17
Words with same initial letters				0.45	0.52				0.28	0.72
High frequency words starting with same letters				0.36	0.45				0.18	0.82
Mean Bigram frequency					0.41				0.19	0.81
Valence						0.92			0.88	0.12
Goodness						0.91			0.85	0.15
Dominance						0.88			0.83	0.17
Pleasantness						0.84			0.82	0.18
Deviation of goodness from mean							0.85		0.75	0.25
Deviation of pleasantness from mean							0.82		0.64	0.36
Emotionality							0.80		0.78	0.22
Arousal							0.63		0.57	0.43
Associates to words (estimation)	0.31						0.34		0.50	0.50
Gender ladenness rating 2								0.98	0.95	0.05
Gender ladenness rating 1								0.97	0.96	0.04
SS loadings	6.80	6.96	4.97	5.23	4.83	3.54	3.05	2.05		
Proportion Var	0.13	0.14	0.10	0.10	0.09	0.07	0.06	0.04		
Cumulative Var	0.13	0.27	0.37	0.47	0.56	0.63	0.69	0.73		

Table 7: Twenty very low frequency words with various prevalence scores, illustrating the danger of experimenter bias if words are selected on the basis of frequency only

Word	Pknown	Prevalence	FreqZipfUS
zarzuela	0.09	-1.32	1.59
cleek	0.13	-1.10	1.59
fovea	0.21	-0.80	1.59
motet	0.25	-0.66	1.59
cantle	0.30	-0.51	1.59
jackleg	0.35	-0.38	1.59
scenarist	0.40	-0.26	1.59
ropy	0.45	-0.11	1.59
snaffle	0.51	0.01	1.59
ablate	0.55	0.12	1.59
karting	0.60	0.25	1.59
lionize	0.66	0.39	1.59
maraud	0.70	0.52	1.59
bluesy	0.75	0.66	1.59
endomorph	0.80	0.83	1.59
inundation	0.85	1.04	1.59
straggle	0.90	1.27	1.59
bullish	0.95	1.62	1.59
dishearten	0.98	1.99	1.59
binocular	1.00	2.45	1.59