The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2

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

Ghent University

To be published in JEPHPP

Key word: word recognition, lexical decision, word prevalence, megastudies

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

Keuleers, Stevens, Mandera, and Brysbaert (2015) presented a new variable, word prevalence, defined as word knowledge in the population. Some words are known to more people than other. This is particularly true for low-frequency words (e.g., screenshot vs. scourage). In the present study, we examined the impact of the measure by collecting lexical decision times for 30,000 Dutch word lemmas of various lengths (the Dutch Lexicon Project 2). Word prevalence had the second highest correlation with lexical decision times (after word frequency): Words known by everyone in the population were responded to 100 ms faster than words known to only half of the population, even after controlling for word frequency, word length, age of acquisition, similarity to other words, and concreteness. Because word prevalence has rather low correlations with the existing measures (including word frequency), the unique variance it contributes to lexical decision times is higher than that of the other variables. We consider the reasons why word prevalence has an impact on word processing times and we argue that it is likely to be the most important new variable protecting researchers against experimenter bias in selecting stimulus materials.

The impact of word prevalence on lexical decision times:

Evidence from the Dutch Lexicon Project 2

Factorial designs vs. regression analysis of big datasets

Traditionally, word recognition researchers have been interested in the question whether a theoretically important variable has a statistically significant influence on word processing. This is typically done with a factorial design: Two samples of words are selected that differ as much as possible on the variable of interest and are matched on a series of other variables. Kousta, Vigliocco, Vinson, Andrews, and Del Campo (2011, Experiment 1), for instance, on the basis of their previous work reasoned that word concreteness is not exactly the same as word imageability. To test the hypothesis, they compiled a sample of 40 concrete words and a sample of 40 abstract words, matched on 12 variables (context availability, imageability, familiarity, age of acquisition, word frequency, number of letters, number of phonemes, number of syllables, mean positional bigram frequency, number of orthographic neighbors, mean neighbor frequency, and number of synsets). The words were presented in a lexical decision task and the authors measured whether there was a difference between the abstract words and the concrete words. After finding an advantage for abstract words (568 ms against 590 ms, $p_1 < .001$, $p_2 < .05$), the authors subsequently investigated why abstract words are easier to process than concrete words, against common intuition.

In addition to the above research tradition, in recent years word recognition researchers have become interested in the question how large the impact of various variables on word processing is and which of these variables are required as control variables in factorial experiments. This research line makes use of multiple regression techniques applied to large collections of behavioral data, called megastudies (e.g., Adelman, Sabatos-DeVito, Marquis, & Estes, 2014; Baayen, Feldman, & Schreuder, 2006; Balota et al., 2004, 2007; Chetail, Balota, Treiman, & Content, 2015; Cortese, McCarty, & Schock, 2015; Keuleers, Lacey, Rastle, & Brysbaert, 2012; Spieler & Balota, 1997). Since the influence of each variable is taken into account by including them as predictors in the regression analysis, word stimuli do not need to be matched on each potential confounding variable. The proponents of the regression approach see this as a more appropriate way to examine the factors affecting word processing times. Table 1 summarizes the main differences between factorial designs and

megastudies and the potential advantages of megastudies, as mentioned by Balota, Yap, Hutchison, and Cortese (2012), Brysbaert, Keuleers, and Mandera (2014), and Keuleers and Balota (2015).

- -

Insert Table 1 about here

- -

The megastudy approach was made possible by technical advances that allowed researchers to collect and analyze large-scale databases. In this respect, they are part of the big data movement currently seen in science at large (Boyd & Crawford, 2012; Howe et al., 2008; Wu, Zhu, Wu, & Ding, 2014). In language research, the movement has been most prominent in natural language processing (NLP) where billion word corpora are used to algorithmically derive word and text features (Cambria & White, 2014). In addition, increasingly large databases are collected that address typical psycholinguistic variables, such a word processing times and word features derived from human ratings.

Another motivation behind megastudies has been the realization that scientific progress not only relies on falsification but also on verification. For a long time, verification was the cornerstone of the scientific method, until Popper (1934, 1963) convincingly argued that falsification is a stronger test. However, the logical advantage of falsification does not imply that verification has no place in scientific research. Specifically related to word research, the concern has grown that the thousands of theory-driven, small-scale factorial studies testing specific hypotheses do not answer broad questions such as: Which word characteristics explain most of the variance in word recognition times? Are the measures of these characteristics good or can they be improved? Are the existing word features enough to explain all differences in word processing times or do we need more of them?

Studies focusing on the statistical significance of small-scale effects are good for theory-driven, detailed questions but should be embedded in more general, descriptive information about the phenomenon under investigation, to make sure that the hypotheses tested apply to phenomena of a worthwhile size and are examined accurately. The wider information is provided by datasets that acknowledge the importance of exploration and verification in scientific research, in addition to falsification. Table 2 summarizes the strengths and the

weaknesses of falsification and verification in scientific progress (Brysbaert & Rastle, 2013; Ward, 1998).

- -

Insert Table 2 about here

- -

Analyses of megastudies are ideal for verification, but can also be used for hypothesis-driven falsification (Kang, Yap, Tse, & Kurby, 2011; Kuperman, 2015). With respect to Kousta et al.'s (2011) research question, for example, it is perfectly possible to draw the data for the selected stimulus set from a megastudy database. Moreover, it is possible to automatically draw random samples repeatedly according to the same criteria, to see how well the finding generalizes beyond the specific sample studied (Kuperman, 2015).[1]

A final advantage of megastudies is their statistical power. Many factorial designs include too few stimuli and participants (Vankov, Bowers, & Munafo, 2014), increasing the chances of spurious effects being published (Kühberger, Fritz, & Scherndl, 2014; Open Science Collaboration, 2015), of flexible data collection and analysis being applied to get 'promising' trends beyond the statistical significance threshold (Francis, 2012; Simmons, Nelson, & Simonsohn, 2011), and of demand characteristics playing a role in item selection (Forster, 2000; Kuperman, 2015). Megastudy data can also be used to check the generality of a newly found effect by means of virtual replications, and to check whether the stimuli across conditions were properly matched.

Variables affecting performance in lexical decision tasks

As indicated above, one of the aims of megastudies is to find out which variables affect word recognition and what their relative importance is. Word recognition megastudies thus far mainly involved lexical decision and word naming (there are two studies with progressive demasking: Ferrand et al., 2011; Lemhöfer, Dijkstra, Schriefers, Baayen, Grainger, &

[1] It is, of course, also possible to use regression analysis for falsification. This will be particularly interesting when the effect is reasonably large and the variable present for many stimuli in the megastudy data.

Zwitserlood, 2008). With respect to lexical decision (the task used in the present study), the following conclusions were reached.

First, there is good evidence that word frequency is the most important variable to predict lexical decision times and accuracy levels, at least if a good frequency measure is used. For all databases, word frequency correlates most with reaction times (RT). Yap and Balota (2009) reported that it accounted for over 40% of the variance in RT for monomorphemic words (both monosyllabic and multisyllabic) in the English Lexicon Project (see also Brysbaert & New, 2009). Ferrand et al. (2010) observed that 38% of the variance in RTs in the French Lexicon Project could be accounted for by frequency. Similar figures were reported by Keuleers, Diependaele, and Brysbaert (2010) for the Dutch Lexicon Project and by Keuleers et al. (2012) for the British Lexicon Project. The last two databases only include monosyllabic and disyllabic words.

A variable coming close to word frequency in the percentage of variance accounted for, is age-of-acquisition (AoA), the age at which the word was learned first. However, because AoA is highly correlated with word frequency ($r = -.63$ for 30,389 ratings collected by Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), the percentage of additional variance accounted for after word frequency has been partialed out, is typically in the order of 5%. Word familiarity also correlates highly with lexical decision times but it is uncertain whether it still explains unique variance once word frequency and AoA are taken into account (Brysbaert & Cortese, 2011; Ernestus & Cutler, 2015), at least in a typical sample of university students (Kuperman & Van Dyke, 2013).

Two more variables that seem to matter are the number of letters and syllables in a word: Word processing times are longer for long words and words containing several syllables. Word length accounts for 16% of the variance in English Lexicon Project (ELP), but for only 7% in the Dutch Lexicon Project (DLP) and the French Lexicon Project (FLP). The most likely reason for this difference is that in ELP all nonwords were made by changing a single letter in a word, whereas in DLP and FLP care was taken not to have a correlation between nonword length and word similarity. The percentages of variance accounted for by syllable length were 10% in ELP, 5% in DLP, and 2% in FLP. Because of the intercorrelations between the variables (short words tend to be of high frequency and acquired earlier), the percentage of additional variance accounted for by both variables is considerably smaller if they are entered after word frequency and AoA.

A final variable that seems to account for a decent part of variance in lexical decision times is the similarity to other words: Words that are orthographically similar to other words are responded to faster in lexical decision than words with unique letter sequences. This effect is usually explained by assuming that lexical decisions are partly based on the overall activation in the mental lexicon: When the overall activation exceeds a certain level, a word response is initiated before the target word itself has been identified (Andrews, 1997; Grainger & Jacobs, 1996). The overall activation is higher when a target word activates many similar word forms than when it is a hermit. A measure that seems to capture most of the variance due to word similarity is the orthographic Levenshtein distance OLD20 (Yarkoni, Balota, & Yap, 2008). It accounts for 20% of the variance in ELP, against some 5% in DLP and FLP. Again, the most likely origin of this difference is the strong word length effect in ELP because of the nonwords used. OLD20 is highly correlated with word length and, therefore, the additional variance it explains after the other variables have been taken into account is considerably less. There is some discussion about whether the word length effect could be an OLD20 effect in disguise (Yarkoni et al., 2008). If that is the case, the length effect would disappear if entered after OLD20.

Variables that do not seem to account for much variance in lexical decision times, are related to the phonology of the words. The study addressing phonological variables in greatest detail is Yap and Balota (2009). They reported that word onset only accounts for .7% of variance in lexical decision times for the monomorphemic words in ELP, and that the stress pattern was no longer significant if entered after frequency, word length, and similarity to other words.

Semantic variables also predict a surprisingly small percentage of additional variance in lexical decision times, even though these effects are usually significant. Juhasz, Yap, Dicke, Taylor, and Gullick (2011) reported that the semantic variables imageability, sensory experience ratings (how much does the meaning of a word involve sensory experiences), and body-object interaction ratings together accounted for less than 1% of additional variance. To further examine the impact of semantic variables, we included the concreteness ratings recently collected by Brysbaert, Warriner, and Kuperman (2014) and looked at how much variance in the ELP data they accounted for after the other variables had been entered. Table 3 shows the results. The contribution of concreteness in this analysis (largely limited to

lemmas[2] for which there were concreteness ratings) was no longer significant (either when tested linearly or nonlinearly).

- -

Insert Table 3 about here

- -

Kuperman, Estes, Brysbaert, and Warriner (2014) reported that, for the ELP words with affective ratings, word valence accounted for an additional 2% of the variance (mainly because positive words are responded to faster) and arousal for .5% of the variance (highly arousing words are responded to more slowly). Other papers report that 1% extra variance may be accounted for by interactions between the variables (Cortese & Schock, 2013; Kuperman et al., 2014). In general, the effects of the non-frequency variables are smaller for high frequency words than for low frequency words.

All in all, it looks like some 63-64% of the variance in the ELP lexical decision times can be accounted for by the variables we have at present. No big gain is to be expected from other variables correlating with any of the just mentioned word characteristics. Indeed, none of other variables listed on the ELP website substantially improved the fit of the model reported in Table 3. This raises the question whether 64% represents the degree of variance in ELP that can be accounted for by word features, or whether we are missing some variable.

Word prevalence

Keuleers, Stevens, Mandera, and Brysbaert (2015) recently proposed a new variable, word prevalence, which may explain a decent share of the variance. They started from the observation that some low frequency words seem to be better known than predicted on the basis of their corpus frequency, as illustrated in Table 4. The table contains 10 English words with a frequency of 0 according to SUBTLEX-US (Brysbaert & New, 2009) and HAL (Balota et al., 2007). However, the words in the left column are words arguably known by many people, whereas the words in column 2 are unknown to many.

[2] A lemma is the uninflected, base form of a word. It is the form typically included in a dictionary.

- -

Insert Table 4 about here

- -

Although it is tempting to try to collect a language corpus that better reflects the difference in subjective frequency between the left and the right column of Table 4 (see the Discussion), Keuleers et al. (2015) opted for another strategy. They decided to use data from a very large group of participants who had answered whether they knew a word or not. The resulting percentages were seen as a measure of word prevalence, the degree of diffusion in the population.

When Keuleers et al. (2015) entered word prevalence as a predictor in a model trying to predict the lexical decision times of DLP, they observed that the variable added some 10% extra variance to the model (which further included word frequency, length, and OLD20). Unfortunately, because DLP only contains monosyllabic and disyllabic words (many of which were inflected), the new variable could only be tested on 7,800 words, most of which were short words.

In the study below, we repeat Keuleers et al. (2015) but this time with a new study in which lexical decision data were collected for 30,000 words of various lengths. As the word prevalence measure is most informative for lemmas, only such forms were included in the list. In addition, AoA and concreteness ratings were collected for all words (Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014), so that, for the first time, we can run a full analysis on the word prevalence measure and see how it affects the impact of the existing set of variables. Importantly, we were not interested in whether word prevalence would predict the accuracy rates for the words (which is a no-brainer), but whether it would predict the reaction times to the words known by the participants.

Method

Participants

A total of 81 participants finished the experiment, 40 for the first list and 41 for the second list. They were recruited via the participant portal of the Department of Experimental Psychology

and via a Facebook group. Of the participants, 57 were female and 24 were male. Their age ranged between 16.6 and 72.1 years (mean 26, standard deviation 8.76). Seven other participants started the experiment but did not finish it. Participants who completed the experiment were paid 200 Euro for their effort.

Procedure

Each participant started with a one-hour startup session at the university. In that session, a reading questionnaire was filled in and the Lextale tests for Dutch, English, French and German were administered (Lemhöfer & Broersma, 2012; Brysbaert, 2013).

The participants brought their own laptop to the lab and we installed the software for the lexical decision task on it. The software was written in C using an updated version of the Tscope library (Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006), which runs on the latest versions of Windows and OS X. During the installation we made sure that the laptop could properly trace the vertical synchronization signal of the screen. Two participants had laptops that failed the test (early Windows Vista computers); they were given netbooks from the Department to use.

Once the software was installed, the participants did two practice blocks of 200 number decision trials. In that task, all two-digit number from 00 to 99 appeared. In the other half of the trials one digit was replaced by a letter, and the participants had to indicate whether the stimulus was a two-digit number or a letter-number combination. In the first block of practice trials the participants received trial-by-trial feedback about their response: On error trials the stimulus re-appeared on the screen in red and stayed on for an extra 1000 ms. After the first number block there was no trial-level feedback anymore; feedback was given at the end of the block. The practice blocks were used to make sure that participants clearly understood the response assignment and had the necessary familiarity with binary decision making.

After the two practice blocks the participants were sent home to run the experiment at their own speed in blocks of 500 trials. In each block there were pauses every 100 trials and feedback about mean reaction time and accuracy was provided at the end of the block. After each block the participants were asked whether they wanted to run a new block right away or later. If the accuracy of a block was below 75%, the participants were warned that they had to work more accurately. If the average reaction time of a block was more than 1000 ms, the

participants were asked to work faster. The instructions stressed that if they got one of the warnings for more than three blocks in a row the experiment would self-abort. Participants were asked to complete the experiment in a time period of four weeks, but this was no strict requirement (i.e., they easily got an extension if they had made good progress). The procedure was the same as in DLP (Keuleers et al., 2010), except for the fact that the participants were allowed to work at home.

Each trial began with two fixation bars above and below the center of the screen. After 500 ms the lines were replaced by the stimulus, which appeared in the gap between the lines and stayed on for 1500 ms or until the participant responded, whichever event occurred first.[3] After the response the screen was cleared and a new trial began 500 ms later.

All participants responded 'word' with their dominant hand and 'nonword' with the other hand. The response keys were the S and L keys of the computer keyboard.

All data were stored on the participants' computer. No intermediate results were sent over the internet. This meant that the participants could run the experiment whenever and wherever they wanted, even when no internet connection was available.

The trial-data were pre-randomized and stored in encrypted files (one file per block) on the participants' computers. The results were also stored in an encrypted file per block. This ensured that a) the participants could not see the stimulus list in advance, b) they could not tamper with their output file, and c) if something went wrong with the program a maximum of 500 trials were lost. The program kept track of the blocks the participant had finished and automatically presented the next one.

The participants were asked to send an email about their progress after every 15 blocks. After the last block was finished, they came back to the university to transfer their results and to receive payment.

[3] This procedure is preferred in our lab over a fixation cross, because the cross may mask the central letter. There was enough space between the lines so that the word could be presented without any visual superposition.

Stimuli

We started with a list of 29,601 words and for each word a matched nonword was generated using the Wuggy algorithm (Keuleers and Brysbaert, 2010; available from http://crr.ugent.be/Wuggy). This list was split randomly in two sublists, with the restriction that when a word was assigned to one list the matched nonword was assigned to the other list. In-between the making of the list and the test, we found 415 extra interesting words[4], which were distributed over both lists without matched nonwords. As a result, the first list contained 15,008 words and 14,800 nonwords and the second list contained 15,008 words and 14,801 nonwords.

Participants were assigned alternately to one of the lists. For each participant the words and nonwords were reordered at random, without any restrictions on the number of words/nonwords that could appear in consecution or per block of trials. The randomized lists were then cut in blocks of 500 trials each. The first two blocks of 500 trials were repeated at the end of the experiment, just before the final block of 308 (or 309) trials. All in all, participants completed 62 blocks.

Results

In line with an open data policy, all data discussed in this article are freely available on our website (http://crr.ugent.be/) and on the Open Science Framework website (https://osf.io/jfkvw/). There are two files with the raw data of DLP2 (one related to the stimuli; the other to the participants) and summary files (at the word level) for DLP2 and word prevalence.

Reaction times and accuracy in DLP2

Block 60 and 61 were exact replications of blocks 1 and 2. For the accuracy analysis we used the results of blocks 1 to 59 + block 62 (the first occurrence of each item); for the reaction time analyses we used the results of blocks 3 to 62. As can be seen in Figure 1, this reduced the impact of the learning effect on RTs, which was particularly strong in the first two blocks.

[4] Making an "exhaustive" list of words of interest for psycholinguistic research is a frustrating activity.

The average correlation between block 1 and block 60 was 0.25, that between block 2 and block 61 was also 0.25, which is higher than the overall reliability (The ICC(C,1) reliability was 0.16). These correlations show that the RTs in both blocks were interchangeable except for the overall longer RTs in the early blocks (see Cortese, Hacker, Schock, & Santo, 2015, for a similar finding). [5]

- -

Insert Figure 1 about here

- -

Reliability of the DLP2 measures

Reliability analyses were run on the reaction times (RTs) and the accuracy scores. For the RT analyses, some basic cleaning was done. First, items with accuracies below 65% were removed. For the remaining items, all trials with error responses were removed. Finally, outlier RT cutoffs were computed per block, participant and lexicality using the adjusted boxplot method (Hubert & Vandervieren, 2008). The adjusted boxplot is similar to the regular boxplot, but acknowledges the fact that RT distributions are skewed by computing boundaries that are not symmetrical around the median.

To find out which dependent variable is best, we started by computing several RT averages per item: a) on the untransformed RTs or z-transformed RTs per participant per block; b) on RT in milliseconds or the response rate per second (1000/RT); c) on the observed RTs only or on RTs with the missing data estimated using an imputation method that adjusts for item mean, participant mean and the overall variance in the data. The latter matrix is interesting, because it allows all types of analyses (factor analyses, component analyses, item analyses, …), which require full datasets.

To get an idea of the stability of the RT averages, the reliability was estimated using ICC(C,k) – the consistency type Intraclass Correlation Coefficient for the average of k participants (Shrout & Fleis, 1979, McGraw & Wong, 1996). Table 5 shows the results (as well as those of a similar analysis of the DLP data).

---

[5] Another way to correct for practice effects, is to work with z-scores per block (Balota et al., 2007; Keuleers et al., 2010).

- -

Insert Table 5 about here

- -

The results show that the reliability of the two item lists was about equal to that of DLP. This means that the reliability of the study was not attenuated by the fact that the participants did the task at home instead of in a laboratory room (which is interesting information for further megastudies). The second finding is that reliability was higher for z-scores: Removing between-block and between-participant variance from the RTs increases the reliability of the mean. The third finding is that the reliability of the response rate was higher than the reliability of the RTs. Table 5 does not include the results of the imputed data, as imputation did not affect the ICC (this was the aim of the imputation method: imputing the data without artificially increasing the ICC estimate).

Another way to estimate the degree of systematic variance in the database is to correlate the dependent variables of the present study with those of other studies. There were 7,530 items in common between DLP and DLP2. Table 6 shows the correlations between both datasets, together with the intraclass correlations calculated for the set of shared words. The observation that the correlation between tests is lower than the reliabilities of the tests themselves indicates that some of the systematic variance is experiment specific. Two elements that may have led to experiment-specific variance are that DLP only included short words and had inflected word forms on nearly half of the trials. The nonwords of DLP also shared these characteristics.

- -

Insert Table 6 about here

- -

Another dataset available for Dutch is BALDEY (Ernestus & Cutler, 2015). This contains auditory lexical decision times for 2,780 words of which 1,160 were in common with DLP2 (the other words in BALDEY nearly all were inflected forms). The correlations between DLP2 and BALDEY for the raw RTs and the accuracies are shown in Table 7. They show that in particular the correlation between RTs is limited, meaning that there are interesting differences between auditory and visual lexical decision times to be investigated.

- -

Insert Table 7 about here

- -

Predictors of the DLP2 lexical decision times

Now that we have assessed the usefulness of the DLP2 data, we can address the core question of the present study: To what extent does word prevalence account for variance in lexical decision times?[6] Although the reliability of the standardized RTs is higher, we will work with the raw RTs, because this variable is easier for the reader to relate to. The findings we report generalize to the other response time measures.[7]

For each word we had the following predictors:

- **Word frequency**: This is a word frequency measure based on subtitles, similar to the SUBTLEX-NL word form frequency published by Keuleers, Brysbaert, and New (2010). The new measure, however, is based on a larger corpus (120 million words instead of 41 million) and has also been cleaned better for optical character recognition errors and other spelling errors present in the corpus. This frequency measure correlates slightly better with the DLP2 RTs and accuracies, and, therefore, is preferred over the original SUBTLEX-NL measure, also because the larger database makes the estimate of the dominant part of speech more reliable. Word frequencies are expressed in Zipf values (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). They equal log10(frequency per billion words) and have the following interpretation: Zipf values 1-3 = low frequency (i.e., equal to or less than 1 per million words), Zipf values 4-7 = high frequency (i.e., equal to or more than 10 per million words). The problem of zero frequencies in the SUBTLEX-NL word list (i.e, assigning a frequency to words that were presented in the experiment,

[6] A similar question for the accuracy data is less interesting, because inherently there is a high correlation between word prevalence and word accuracy (r = .75 for the words in common between DLP, DLP2, and the crowdsourcing study), as word prevalence measures how many people know a word.

[7] The percentages of variance accounted for were 53.6% for zRT, 52.9% for rate, and 56.5% for zrate.

but not in the SUBTLEX-NL corpus) was dealt with by using Laplace smoothing (i.e., by using observed frequency plus one and changing the denominator to take into account the extra unobserved frequencies; Brysbaert & Diependaele, 2013).

- **Word length**: The number of letters in the word.
- **Syllable length**: The number of syllables in the word.
- **Age-of-acquisition**: Age at which the word is thought to have been learned (based on ratings provided by university students; Brysbaert et al., 2014).
- **Concreteness**: Degree to the word refers to an experience-based entity. Measured with a 5-point Likert scale (Brysbaert et al., 2014).
- **Old20**: The similarity to other words, calculated on the basis of the Celex vocabulary (Yarkoni et al., 2008; Keuleers, 2015).
- **PoS**: Dominant part of speech of the word (is the word most often used as a verb, a noun, an adjective, an adverb, a function word, or a number word).
- **Word prevalence**: Measure of the number of people who knew the word in a crowdsourcing study involving over 100 thousand participants from Flanders, the Dutch-speaking half of Belgium (Keuleers et al., 2015). Expressed as probit values. These are the z-values corresponding to the percentages known when the latter are considered as the cumulative probability function of the standardized normal distribution. So, a word known by 2.5% of the participants gets a word prevalence value of -1.96; a word known by 97.5% of the participants gets a word prevalence value of +1.96; and so on. The percentages known were based on an item response theory analysis (calculated by means of a mixed effects analysis) in order to correct for response biases in the participants (in particular the bias to say they knew the word when this was not true; the bias was measured by taking into account the 'acceptance rate' for nonwords).

Table 8 shows the intercorrelations between the various predictor variables and the RTs from DLP2. As in Table 3, the analysis was limited to the words that were recognized by at least 67% of the DLP2 participants (N = 24,560). After word frequency, word prevalence had the highest correlation with RT. At the same time, the correlation between word frequency and word prevalence is moderate (r = .51), indicating that both variables measure partially different aspects of word knowledge. As can be seen, word prevalence accounted for an extra 6% of the variance in RTs even after all the other variables had been introduced. Also the dominant part of speech (PoS) made a difference, particularly because responses to the

function words and the number words were slower than predicted on the basis of the other features (see below).

- -

Insert Table 8 about here

- -

An interesting statistic to gauge the specific contribution of each predictor, independent of the other predictors, is to see how much extra variance it accounts for after all the other variables have been entered.[8] As can be seen in Table 9, word prevalence resulted in the highest R^2 change. This analysis also shows that the number of syllables no longer seems to have an effect, once all other variables are taken into account.

- -

Insert Table 9 about here

- -

To further test the robustness of the findings, we also ran cross-validation regression tests (using the validate() function of the rms package in R, run with 1000 bootstrap replications). All findings (including the percentages of variance accounted for) were replicated, in line with the observation that overfitting is unlikely for the present analysis given the small number of predictors and the large number of observations.[9]

---

[8] The maximum variance each predictor accounts for is easily obtained by squaring the correlation between the predictor and the RT in Table 8. The true contribution of each dependent variables lies between the minimum and the maximum depending on the order in which the variables are entered in the hierarchical regression analysis. For instance, word frequency accounts for 33.0%, 3.7% or somewhere in-between depending on whether it is entered first, last or in-between in the hierarchical regression.

[9] Another way to test the robustness of the effects is to make use of the variability in the data and to run mixed effects models on the complete dataset. This, however, increases the number of data lines from 24K to almost 1M. We aborted the lme4() R program after 10 days of running, but were able to run the analysis on the new package developed by Bates in Julia (https://github.com/dmbates/MixedModels.jl, available at 05/19/2015). The results were comparable.

Interactions between the predictors

To test Cortese & Schock's (2013) finding of an interaction between AoA and word frequency (the AoA effect was stronger for low frequency words than for high frequency words), we included all two-way interactions in the regression model. Given that number of syllables did not have a significant effect, it was left out of consideration. Table 10 shows the results of the analysis.

- -

Insert Table 10 about here

- -

A first thing to notice in Table 10 is that the addition of the two-way interactions only adds 3.6% of the variance explained with no interaction term explaining more than .5%. This indicates that the interaction effects do not change the main effects dramatically (i.e., there is no evidence for cross-over interactions).

Figure 2 shows the interactions between AoA and the other predictors. In these figures it can be seen that although the AoA effect exists as a general main effect, it is stronger for words with medium frequency, long words, concrete words and words of low prevalence.

- -

Insert Figure 2 about here

- -

Figure 3 shows the interactions with the new variable, word prevalence. As can be seen, the word prevalence effect is particularly strong for words of medium frequency, short words, late acquired words, and words not resembling many other words (i.e., with high OLD20 values).

- -

Insert Figure 3 about here

- -

Discussion

Word prevalence

This paper investigated the usefulness of a new variable, word prevalence, for understanding word processing times in lexical decision. The variable measures the number of people who indicate they know the word in a vocabulary test (Keuleers et al., 2015). The reason why we thought this variable may be of importance is illustrated in Table 4. The table contains 10 English words that have a frequency of 0 according to SUBTLEX-US (Brysbaert & New, 2009) and HAL (Balota et al., 2007). However, the words in the left column are words that are known to many people, whereas the words in column 2 are unknown to many.

The data of the present study clearly show that word prevalence has a strong impact on lexical decision times, which cannot be accounted for by other known variables (Tables 8 and 9). In our view, it is also the variable responsible for the potential of experimenter bias in word selection (Forster, 2000; Kuperman, 2015). When one has to select words matched on (low) frequency, there is always the temptation of selecting words with a higher prevalence if one does not believe in the frequency effect, or words with a lower prevalence if one hopes to obtain a strong frequency effect.

For a good understanding of the word prevalence measure, it important to appreciate that the impact of variable is not limited to known vs. unknown words.[10] Figure 4 shows the impacts of the various predictors when the analysis is limited to the main effects. This graphs shows that prevalence, together with word frequency, has the strongest impact on RTs (over 100 ms). The other important aspect of the figure, however, is that the effect is linear across the entire range of the variable. Given that word prevalence is expressed as probit values, it is good to repeat that a value of 1 means that the word is known by 84.1% of the people, a value of 2 means that 97.7% know the word, and a value of 3 means that 99.9% of the people know the word. So, a big part of the word prevalence effect is present in words that are quite well known overall. Indeed, most words in the list we tested have a prevalence scores of more than

[10] Many colleagues informally told us that they never present words they do not know themselves. In addition, they always try to gauge the percentage of students who are likely to know the word. Unfortunately, this does not correct for the difference of 50 ms between the words known by 95% of the population vs. the words known by 99.9% of the population (Figure 4).

2, as shown in Figure 5 (remember that we only included words with 67% correct responses in the lexical decision task).

- -

Insert Figures 4 and 5 about here

- -

The easiest interpretation of the word prevalence effect is that it has the same origin as the word frequency effect: Words known by many people are likely to be produced regularly and, hence, likely to be encountered more often than words known by few people. According to this interpretation, the word prevalence measure corrects for gaps in the corpus materials on which the word frequency measures are based. Indeed, corpora are rarely representative for all language registers people use. In this respect, word prevalence has the same function as subjective word familiarity, proposed by Gernsbacher (1984) as a way to counteract the lack of validity of word frequency measures based on word counts. In the years since Gernsbacher (1984), it was hoped that better frequency measures based on larger corpora would take away the need for a subjective familiarity measure, but thus far this has not panned out. Even with the best word frequency measure, some words remain better known than other words of the same frequency. This is what word prevalence corrects for. It is important, however, to realize that word prevalence is not the same as word familiarity or subjective frequency. If all people know an unfamiliar word, this word has a high prevalence but a low familiarity rating. This is the case for words like *slush, oppress, conundrum, copious, alchemist, aboriginal, and plethora* (familiarity ratings based on Coltheart, 1981, and Nusbaum, Pisoni, & Davis, 1984). Subjective ratings of word familiarity seem to be a combination of frequency and prevalence. They correlate much more with word frequency than word prevalence does (e.g., r = .84 in Brysbaert & Cortese, 2011).

A second origin of the word prevalence effect is that it seems to correct for the familiarity with objects many people know from daily life, but which rarely figure in language corpora. These are words such as *ladle, washing machine, eyeliner, hinge, …*. Indeed, there may be a whole class of words with underestimated frequencies of usage, because word frequency measures are entirely based on written materials, and implements are rarely talked or written about except for conversations related to their use.

Third, a substantial number of low-frequency, high-prevalence words seem to be loan words from other languages (this may be stronger in Dutch than in English). Many scientific low-frequency words come from English, French, or German. As a result, people with knowledge of these languages are likely to understand the words, even though they are infrequent in Dutch.

Finally, many of the words with discrepant word frequency and word prevalence measures are compound words and derived words that are low in frequency of occurrence but that are perfectly understandable because they can be decomposed. Words like *distinctively, antioxidant, microbiologist, antivirus, reusable, legalization, unsaturated, relenting, and preconditioned* all have high prevalence but low frequency (and familiarity). Indeed, an interesting use of the word prevalence measures will be to examine to what extent morphological complex words inherit the prevalence of their base words, and which variables influence the generalization.

If we accept that word prevalence compensates for poor estimates of word frequency, the variable can be integrated into existing computational models of word recognition by modifying the mechanisms currently in use for the word frequency effect. We broadly distinguish between two types of models: static models mimicking end-state functioning, and learning models. In the former type, the effect of word frequency is captured by increasing the activation levels of word representations as a function of word frequency. For instance, in the DRC model of visual word recognition (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) units in the orthographic lexicon are frequency-sensitive: The activation of high-frequency words rises more quickly than the activation of the low-frequency words, because the unit of increase depends on the log frequency of the word. In such a model, the simplest way to integrate word prevalence would be to make the unit increase dependent on both word frequency and word prevalence. Applied to the words of Table 4, the well-known words from the left column would have a higher increase unit than the less-known words from the right column. In models that learn, the frequency effect is often mimicked by presenting the words unequal times to the model. For instance, in Seidenberg and McClelland's (1989) PDP model, the probability that a word was presented to the network for training during an epoch was a logarithmic function of its written frequency. Again, the most straightforward way to incorporate word prevalence in this model would be to make the presentation probability of the word dependent on both frequency and prevalence.

The considerations so far suggest that word prevalence has the same origin as word frequency and, therefore, can be integrated in our theories without much impact on our thinking about how word recognition is achieved. However, there may be another, theoretical reason why word frequency is not a perfect estimate of word knowledge. The word frequency effect is based on the assumption that each encounter with a word has equal impact. A word encountered twice is better mastered than a word encountered once, and less well than a word encountered three times. This is not always true. Goodman, Dale, and Li (2008), for instance observed that although closed class function words (articles, determiners, prepositions) are among the most frequently used by parents in their speech to babies and toddlers, these are not picked up by the children until they are over two years old. Only when Goodman et al. (2008) made a distinction between several word types (among which nouns, verbs, and adjectives) did they find a relationship between word frequency in child-directed speech and the order of word acquisition. Similarly, children usually need only one story or movie about a unicorn to remember the word for the rest of their lives, even though its frequency is very low. So, some words are learned much more easily than other (sometimes called one-shot learning or fast mapping; Borovsky, Elman, & Kutas, 2012).[11] These are words that fit in an existing memory schema and enrich it (Coutanche & Thompson-Schill, 2014; Tinkham, 1997; van Kesteren, Beul, Takashima, Henson, Ruiter, & Fernandez, 2013), that have a personal (emotional?) value for the listener, and that have a perceived utility (Breton & Robertson, 2014; Fischer & Born, 2009). Other words are hardly picked up, even though they are encountered multiple times. Typical examples are words that are not central to the understanding of the ongoing events (e.g., specific names of plants, animals, or tools). This dynamic aspect of human learning is completely overlooked by the word frequency effect: Not all words require many encounters to be acquired and (apparently) to be processed rapidly.

The most plausible way to integrate a frequency-independent effect of word prevalence in computational models of word processing is via the semantic route. It is well known that the semantic system is heavily involved in word reading (Taylor, Duff, Woollams, Monaghan, & Ricketts, 2015), although this part of human cognition has not yet been integrated in computational models of word recognition (see Harm & Seidenberg, 2004, for a first, simplified attempt). Words differ in their semantic richness and arguably also in the ease with

[11] Such fast learning is needed, given that school children learn over 2,000 new lemmas per year (Anderson, & Nagy, 1993).

which they activate meanings. There is a difference between encountering the word ‘squall’ in a list of weather phenomena in school and being in the middle of a squall when you are driving from Hamilton to London Ontario on your first trip to Canada, particularly if you drove despite warnings of a squall on television that morning. Chances are higher that you will remember the word after having experienced the second situation than after having been part of the first (see Barsalou, Santos, Simmons, & Wilson, 2008, for the contribution of embodied experiences to the meaning of words). We think that differences in semantic depth and richness determine why knowledge of some words is highly prevalent despite these words being relatively rare. Another variable known to influence the encoding and consolidation of memory traces is the congruency with existing knowledge. Information that fits in existing schemas is remembered better than information not in line with prior knowledge. There is even evidence that different brain mechanisms are involved (van Kesteren, et al., 2013). So, new words naming elements from familiar schemas are likely to be learned faster than new words referring to unfamiliar entities or contradicting prior knowledge.

A caveat with respect to the above considerations is that we must keep in mind that the evidence so far is limited to correlational data. As is well known, correlation itself does not inform about causation. Word prevalence and lexical decision time may be correlated with each other, not because word prevalence captures a feature that influences word processing efficiency, but because word processing efficiency affects word prevalence, or because both are the outcome of a third, underlying variable. This concern is often raised in word recognition research with respect to subjective evaluations, such as age-of-acquisition ratings. A word may not be processed more easily because it has been acquired early in life, but the correlation between age-of-acquisition and word processing efficiency is observed because a person asked to assign an age of acquisition rating to a word gives a rating on the basis of how easy it is to process that word (this is an easy word; thus, I must have learned it early). The same concern has been raised with respect to objective measures such as word frequency. Hayes (1988), for instance, wondered whether the observation that spoken discourse contains fewer low-frequency words than written texts could be due to people avoiding the use of low-frequency words in spoken discourse to preserve the fluency of their speech. According to Hayes (1988) the difficulty of producing a word could therefore determine the frequency of occurrence (and not the other way around).

Word prevalence and word frequency share the advantage that they are objective measures rather than subjective impressions based in a small group of raters. The disadvantage of the

word prevalence measure is that it is derived from a word judgment task ('Is this a word that you know? Yes/No'), which is very similar to the lexical decision task. The main difference between the tasks is that in the vocabulary test with which we measured word prevalence participants were motivated to obtain an estimate of the size of their vocabulary without time pressure while the typical lexical decision task instructs participants to perform as rapidly and as accurately as possible without further motivation. Still, there is no failsafe guarantee that word prevalence is a *cause* of word difficulty. One could argue that processing speed and word prevalence are correlated because they are both outcome variables caused by processing difficulty (whatever 'difficulty' then may be). Our only defense at the moment is that the prevalence measure follows well-known differences in word usage, such as an increase in vocabulary size with age and level of education (Keuleers et al., 2015)[12]. These are indications of construct validity, but in the end they are correlations as well.

There are three ways to further corroborate the causal role of word prevalence in word recognition. The first is to use different test formats, which probe more into the meaning of the words. Two such formats are multiple choice ('Which alternative agrees most with the target word?') and dictionary definition ('Is this definition correct?'). A challenge with these formats will be to ensure that all questions are matched in difficulty, because the difficulty not only depends on knowledge of the target word, but also on the similarity and difficulty of the answer alternatives. A second way to tackle the causality issue is to use word-learning studies, in which researchers have more control about which words are introduced at which moment in time and under which training regime. [13] If semantic richness is the causal factor, it must be possible to manipulate this variable, while keeping frequency of exposure constant. Examples of studies using the learning approach are Bowers, Davis, and Hanley (2005), Merkx, Rastle, and Davis (2011), Catling, Dent, Preece, and Johnston (2013), and Hawkins, Astle, and Rastle (2014). Finally, we can further corroborate the causal status of word prevalence by correlating the measure with performance in other tasks (e.g., word naming or gaze durations in reading).

---

[12] There were also interesting differences between countries (Belgium vs. the Netherlands) and between genders. Some words were much better known to men than to women and other words were more prevalent among women. English equivalents of gender-specific words are *paladin, kevlar, dreadnought, and golem* (better known by men) and *tresses, taupe, peony, and bodice* (better known by women).

[13] The authors thank Jeff Bowers for this suggestion.

Independent of the ontological status of word prevalence, the correlation between prevalence and word processing speed means that the variable can be used for prediction and control. Both Forster (2000) and Kuperman (2015) warned against the potential of experimenter bias in word selection on the basis of the features currently used (e.g., when the words in two conditions must be matched). Our data show that taking word prevalence into account will considerably alleviate this problem.

Word frequency, word length, and part of speech

Next to word prevalence, six other variables were investigated. As before, word frequency and word length (number of letters) had strong impacts. However, an interesting aspect of the present study is that both effects seem to have become linear, whereas in previous studies non-linear effects were described. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004), for instance, reported that the frequency effect reached a floor at high frequency levels, a finding subsequently confirmed by Keuleers, Diependaele, and Brysbaert (2010), among others. One reason for the higher than expected RTs for high frequency words is that many of these words are function words and number words (one, two, …), two types of words that take longer to respond to in a lexical decision than content words. Two reasons (not necessarily mutually exclusive) for the longer response times may be (1) the fact that function and number words form a minority in lexical decision tasks, and (2) that function and number words usually are accompanied by other words and, therefore, not seen in isolation. Even though part-of-speech has but a small impact on lexical decision times overall, it does have a strong impact on the high end of the frequency range and the low end of the word length range, which may be the origin of the nonlinearity of these effects reported in previous studies (see New, Ferrand, Pallier, & Brysbaert, 2006, for nonlinearity in the word length effect). Of course, the introduction of word prevalence may have taken away some of the nonlinearity as well. The effect of number of syllables did not survive the introduction of the other variables.

Concreteness

Concreteness did not have a significant effect in the ELP data when the other variables were controlled for (Table 3). In contrast, the effect remains significant in the DLP2 dataset, but importantly is reversed from what is traditionally assumed: Reaction times to abstract words

(low concreteness values) are faster than RTs to concrete words (Figure 4). As we have seen in the introduction, this phenomenon has been observed before (Kousta et al., 2011) and related to the fact that abstract words tend to be more emotionally laden than concrete words: Words with a strong valence (in particular positive words) tend to be responded to more rapidly than neutral words. The reason why the reverse concreteness effect has not been picked up more widely before is that it is usually disguised by confounded variables, such as word frequency, length, or OLD20.

Interestingly, the reversed concreteness effect seems to be particularly strong for late acquired words (Figure 2) and low-frequency words (Figure 6). There is even some evidence that familiar, often encountered concrete words (arguably those acquired early) can be processed faster than matched abstract words.

- -

Insert Figure 6 about here

- -

OLD20

Words similar to many other words (i.e., with low OLD20 values) were responded to faster than more distinct words (Figure 4). As explained in the introduction, this is usually explained by assuming that words with a high similarity to other words activate many lexical representations, so that the yes-response can be based on the total activation in the lexicon rather than on the activation level in the representation of the target word. In line with this idea, the word frequency effect and the word length effect are absent for words with low OLD20 values (Figure 7). Also the prevalence effect is smaller for words with low OLD20 values. The same, however, is not true for AoA. It is not clear at the moment how to interpret these divergent findings.

- -

Insert Figure 7 about here

- -

AoA

The main effect of AoA survived the introduction of word prevalence (Table 9), in line with the finding that the order of acquisition has an effect on the ease of processing, independent of the cumulative frequency with which words or objects have been encountered (Catling, et al., 2013; Ellis & Lambon Ralph, 2000; Monaghan & Ellis, 2010). The AoA effect is particularly strong for long, concrete words with medium frequency and non-maximal prevalence. This pattern seems to be most in line with those theories that see a semantic origin of (part of) the AoA effect (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Steyvers & Tenenbaum, 2005).

Conclusion

In this paper we introduced the Dutch Lexicon Project 2 (DLP2) and we showed that the addition of word prevalence increases the percentage of variance explained in lexical decision times to known words by 6%. The reason for this considerable increase is the high correlation of word prevalence with lexical decision times (outperformed only by word frequency) and the reduced correlations with the other predictors currently used: Word frequency, word length, age of acquisition, similarity to other words, concreteness, and part of speech. We also showed why taking word prevalence into account in stimulus selection is likely to reduce the potential of experimenter bias. At the same time, the maximal percentage of variance explained (52%) still remains well below the percentage of systematic variance in the database. This suggests that there may be other variables to discover.

At present, word prevalence is only available for the Dutch language. Given its importance, it is worthwhile to collect similar information for other languages. We have started to collect the necessary data for English (http://vocabulary.ugent.be/) and Spanish (http://vocabulario.bcbl.eu/) and hope to publish these soon.

References

Adelman, J. S., Sabatos-DeVito, M. G., Marquis, S. J., & Estes, Z. (2014). Individual differences in reading aloud: A mega-study, item effects, and some models. *Cognitive Psychology, 68*, 113–160.

Anderson, R.C., & Nagy, W.E. (1993). *The vocabulary conundrum*. University of Illinois at Urbana-Champaign: Technical Report No 570 from the Center for the Study of Reading (available at https://www.ideals.illinois.edu/bitstream/handle/2142/18019/ctrstreadtechrepv01993i00570_opt.pdf, September 3, 2015).

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

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55(2),* 290-313.

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

Balota, D. A., Yap, M. J., Hutchison, K.A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In James S. Adelman (Ed). *Visual word recognition* (pp. 90-115). Hove: 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., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). Language and simulation in conceptual processing, in M. De Vega, A. M. Glenberg, and A. Graesser (Eds), *Symbols, Embodiment, and Meaning* (pp. 245–283). Oxford: Oxford University Press.

Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development, 8(3),* 278-302.

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Interfering neighbours: The impact of novel word learning on the identification of visually similar words. *Cognition, 97(3),* B45-B54.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, *Communication & Society, 15(5),* 662-679.

Breton, J., & Robertson, E. M. (2014). Flipping the switch: Mechanisms that regulate memory consolidation. *Trends in Cognitive Sciences, 18(12),* 629-634.

Brysbaert, M. (2013). LEXTALE_FR: A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica, 53,* 23-37.

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., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior research methods, 45(2),* 422-430.

Brysbaert, M., Keuleers, E., & Mandera, P. (2014). A plea for more interactions between psycholinguistics and natural language processing research. *Computational Linguistics in the Netherlands Journal, 4* (Available at: http://www.clinjournal.org/sites/default/files/14-Brysbaert-etal-CLIN2014.pdf).

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. & Rastle, K. (2013). *Historical and Conceptual Issues in Psychology* (2nd edition). Harlow: Pearson Education.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150,* 80-84.

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

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

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine, 9(2),* 48-57.

Catling, J., Dent, K., Preece, E., & Johnston, R. (2013). Age-of-acquisition effects in novel picture naming: A laboratory analogue. *The Quarterly Journal of Experimental Psychology, 66(9),* 1756-1763.

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

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

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108(1),* 204-256.

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

Cortese, M.J., McCarty, D.P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *Quarterly Journal of Experimental Psychology*.

Cortese, M. J., & Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *The Quarterly Journal of Experimental Psychology, 66(5)*, 946-972.

Coutanche, M. N., & Thompson-Schill, S. L. (2014). Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology: General, 143(6)*, 2296-2303.

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

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology*.

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

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

Fischer, S., & Born, J. (2009). Anticipated reward enhances offline learning during sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(6)*, 1586-1593.

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

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19(6)*, 975-991.

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.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35(3)*, 515-531.

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

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review, 111(3),* 662-720.

Hawkins, E., Astle, D. E., & Rastle, K. (2015). Semantic Advantage for Learning New Phonological Form Representations. *Journal of Cognitive Neuroscience, 27(4),* 775-786.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... & Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature, 455(7209),* 47-50.

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

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *The Quarterly Journal of Experimental Psychology, 64(9),* 1683-1691.

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

Keuleers, E. (2015). *Package 'vwr': Useful functions for visual word recognition research*. Retrieved from http://cran.r-project.org/web/packages/vwr/vwr.pdf (available at 03/13/2015).

Keuleers, E., & Balota, D.A. (2015). Megastudies, Crowdsourcing and Large Datasets in Psycholinguistics: An Overview Of Recent Developments. *Quarterly Journal of Experimental Psychology*.

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

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*, 643-650.

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*.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General, 140(1),* 14-34.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PloS one, 9(9),* e105825.

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

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.

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

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

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

Merkx, M., Rastle, K., & Davis, M. H. (2011). The acquisition of morphological knowledge investigated through artificial language learning. *The Quarterly Journal of Experimental Psychology, 64(6),* 1200-1220.

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

McGraw, K. O. & Wong, S. P. (1996). Forming Inferences about some Intraclass Correlation Coefficients. *Psychological Methods, 1,* 30-46.

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.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words.* Indiana University: Research on speech perception progress report no. 10.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349 (6251),* aac4716 [DOI:10.1126/science.aac4716].

Popper, K. R. (1934), *Logik der Forschung,* Vienna: Springer.

Popper, K. R. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge & Keegan Paul.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96(4),* 523-568.

Shrout, P. E. & Fleis, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86,* 420-428.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011) False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22,* 1359–1366.

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

Stevens, M., Lammertyn, J., Verbruggen, F., and Vandierendonck, A. (2006). Tscope: a C library for programming cognitive experiments on the MS Windows platform. *Behavior Research Methods 38,* 280-286.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29(1),* 41-78.

Taylor, J.S.H., Duff, F.J., Woollmans, A.M., Monaghan, P., & Ricketts, J. (2015). How word meaning influences word reading. *Current Directions in Psychological Science, 24(4),* 322-328.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second language research, 13(2),* 138-163.

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

van Kesteren, M. T., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. *Neuropsychologia, 51(12),* 2352-2359.

Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology, 67(5),* 1037-1040.

Ward, N. (1998). Artificial intelligence and other approaches to speech understanding: Reflection on methodology. *Journal of Experimental and Theoretical Artificial Intelligence, 10,* 487-493.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. Knowledge and Data Engineering, *IEEE Transactions, 26(1),* 97-107.

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

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

Table 1: Main differences between the traditional small-scale factorial designs and megastudies in psycholinguistics (based on Balota et al., 2012; Brysbaert et al., 2014; Keuleers & Balota, 2015).

Factorial designs	***Megastudies***
Research targets specific variables (e.g., word frequency and/or word concreteness)	Research targets the collection and analysis of large datasets (e.g., 2,500 English monosyllabic words)
Experimental approach with independent variables (frequency, concreteness) and control variables (word length, similarity to other words, age-of-acquisition, …)	Correlational approach with predictor variables (frequency, concreteness, length, similarity to other words, age-of-acquisition, …)
Selection of small samples of words (typically some 20 per condition) that make (orthogonal) testing of the independent variables possible and that meet the control requirements	No selection of words; all words with data are included in the analysis; preference for large datasets
Variables are tested in a categorical way (e.g., high frequency vs. low frequency)	Variables are tested in a continuous way along their entire range (e.g., frequencies going from 2 Zipf to 7 Zipf)
Typically assumes linear effects (e.g., the frequency effect is a linear effect from high to low frequency)	Uses nonlinear regression to establish the best fitting curve of the effect
Primary interest in the statistical significance of the effect ($p < .05$)	Primary interest in the size of the effect (e.g., percentage of variance accounted for)
Primary interest in the position of the stimuli relative to each other (higher, lower frequency)	Also interest for the absolute values of the measures (e.g., word frequencies expressed in Zipf units)
Tends to limit the study to words at the extremes of the variable (very low and very high on the continuum of interest)	Takes all words into account, to make sure that the extreme words are no outliers.

Table 2: The use of falsification and verification in scientific research (inspired by Ward, 1998).

Falsification	***Verification***
The ultimate method to find out whether a scientific theory is an adequate explanation or not. Corrects for the confirmation bias people have.	The most efficient method to find out whether an idea has practical value (i.e., accounts for variance in empirical data).
Requires a theory to test plus a good understanding of the assumptions one is making and the techniques one is using for the test.	Is especially useful when one does not (yet) have a full-fledged theory, is building a new artificial system (e.g., a computational model), or testing a new research method.
Focuses on radically new perspectives and theoretical breakthroughs: Scientific progress consists of replacing bad old ideas by good new ones.	Focuses on the improvement of existing systems; scientific progress is seen as incremental.
Is motivated by long-term, theoretical goals (full understanding of a phenomenon)	Is motivated by short-term, pragmatic goals (increased efficiency).
May lead to protracted yes-no discussions between opponent views.	May get bogged down in non-optimal, local minima (solutions that seem to slightly improve the fit to the empirical data, but that turn out to be fruitless).
Has a high risk of proposing false trails (bold new explanations that show some initial empirical support, but in the end fail the relentless empirical testing)	Has a high risk of overlooking theoretically important phenomena that account for a small amount of variance.

Table 3: The contribution of word concreteness to the lexical decision times in the ELP dataset. Data limited to the words for which we had measures for all variables and that were recognized by more than 67% of the participants (N = 19,405).

Correlations between variables

	N_{lett}	N_{syll}	AoA	OLD20	Conc	RT
Frequency SUBTLEX	-0.49	-0.42	-0.61	-0.48	0.14	-0.64
Length in letters		0.84	0.47	0.87	-0.31	0.63
Length in syllables			0.51	0.74	-0.37	0.59
AoA				0.47	-0.39	0.59
OLD20					-0.20	0.64
Concreteness						-0.24

Percent variance in RTs accounted for in regression analyses. We used restricted cubic splines with 3 knots used for frequency and N_{lett} to take into account the nonlinear nature of these effects. The word length variables were entered before AoA, in order not to favor the latter. All variables had a significant contribution, except for concreteness.

Frequency	42.4%
Frequency + N_{lett}	57.6%
Frequency + N_{lett} + N_{syll}	58.4%
Frequency + N_{lett} + N_{syll} + AoA	60.6%
Frequency + N_{lett} + N_{syll} + AoA + OLD20	61.5%
Frequ ency + N_{lett} + N_{syll} + AoA + OLD20 + Conc	61.5%

Table 4: Words with very low frequencies according to SUBTLEX-US and HAL, but with high (left) or low (right) word prevalence measures.

– toolbar	- scourage
– screenshot	- kestrel
– soulmate	- thunk
– uppercase	- whicker
– hoodie	- caudle

Table 5: Reliability of the RT measures in DLP and the two sublists of DLP2, based on the intraclass correlation. The data for the imputed values were the same.

		DLP	DLP2a	DLP2b	DLP2 a + b
Words					
	Rt	.84	.83	.84	.84
	Rate	.87	.88	.87	.88
	Zrt	.87	88	.88	.88
	Zrate	.89	.90	.90	.90
	Accuracy	.96	.95	.95	.95
Nonwords					
	Rt	.83	.80	.80	.80
	Rate	.86	.85	.83	.84
	Zrt	.86	.86	.86	.86
	Zrate	.88	.87	.87	.87
	Accuracy	.89	.91	.91	.91

Table 6: Intraclass correlations for the 7,530 words shared between DLP and DLP2, and correlations between the dependent variables of both lexicon projects (last column).

Overlapping		DLP	DLP2a	DLP2b	DLP2 a + b	Correlation DLP-DLP2
Words						
	Rt	.83	.76	.76	.76	.69
	Rate	.87	.83	.81	.82	.75
	Zrt	.86	.84	.82	.83	.76
	Zrate	.89	.86	.85	.86	.79
	Accuracy	.94	.92	.92	.92	.89

Table 7: Correlations between the performance measures of DLP2 and those of BALDEY for the 1,160 words in common.

	Baldey	Dlp2a	Dlp2b	Dlp2 a + b	Correlation Baldey-DLP2
Rt	.60	.81	.83	.82	.26
Accuracy	.86	.93	.94	.93	.57

Table 8: The contribution of the various word predictors to the lexical decision times in the DLP2 dataset. Data were limited to the words for which we had measures for all variables and that were recognized by more than 67% of the participants (N = 24,560).

Correlations between variables

	N_{lett}	N_{syll}	AoA	OLD20	Conc	Prevalence	RT
Frequency SUBTLEX	-0.38	-0.32	-0.50	-0.36	0.02	0.51	-0.57
Length in letters		0.80	0.33	0.74	-0.23	-0.04	0.34
Length in syllables			0.41	0.59	-0.29	-0.12	0.32
AoA				0.34	-0.43	-0.45	0.48
OLD20					-0.09	-0.12	0.39
Concreteness						0.16	-0.06
Prevalence							-0.53

Percent variance in RTs accounted for in regression analyses. We used restricted cubic splines with 3 knots used for frequency and N_{lett} to take into account the nonlinear nature of these effects. The word length variables were entered before AoA, in order not to favor the latter. The intraclass correlation for this dataset was .83, which gives the upper limit of the variance that can be explained.

Frequency	32.96%
Frequency + N_{lett}	36.98%
Frequency + N_{lett} + N_{syll}	37.53%
Frequency + N_{lett} + N_{syll} + AoA	41.38%
Frequency + N_{lett} + N_{syll} + AoA + OLD20	42.49%
Frequency + N_{lett} + N_{syll} + AoA + OLD20 + Conc	42.86%
Frequency + N_{lett} + N_{syll} + AoA + OLD20 + Conc + Prev	49.05%
Frequency + N_{lett} + N_{syll} + AoA + OLD20 + Conc + Prev + PoS	49.37%

Table 9: The ANOVA table of the final model in Table 8, together with the unique variance each predictor accounted for once the influence of the other was partialed out. This table shows that the number of syllables no longer affects RTs once the other variables are taken into account. It also shows that word prevalence accounts for the most unique variance in RTs because it has a high correlation with RT and is less correlated with the other predictors than word frequency.

	F	P	R square change
Frequency	F(2, 24546) = 904.8	p <<0.001	3.73%
N_{lett}	F(2, 24546) = 501.23	p <<0.001	2.06%
N_{syll}	F(1, 24546) = 0.427	p = 0.427	0.00%
AoA	F(1, 24546) = 943.02	p <<0.001	1.94%
OLD20	F(1, 24546) = 309.83	p <<0.001	0.63%
Concretenesss	F(1, 24546) = 370.32	p <<0.001	0.76%
Prevalence	F(1, 24546) = 2695.4	p <<0.001	5.55%
PoS	F(4, 24546) = 39.02	p <<0.001	0.32%

Table 10: ANOVA table of the ordinary regression with two-way interactions (type III tests on centered predictors[14]). The R squared of this model was 52.96%.

	F	p	R square change
Frequency	F(2,24521) = 1222.3	p << 0.001	4.68 %
N_{lett}	F(2,24521) = 110.16	p << 0.001	0.42 %
AoA	F(1,24521) = 874.51	p << 0.001	1.67 %
OLD20	F(1,24521) = 42.21	p << 0.001	0.08 %
Concretenesss	F(1,24521) = 342.10	p << 0.001	0.65 %
Prevalence	F(1,24521) = 2735.1	p <<0.001	5.24 %
PoS	F(4,24521) = 12.37	p <<0.001	0.09 %
Frequency:N_{lett}	F(4,24521) = 0.34	p = 0.847	0.00 %
Frequency:AoA	F(2,24521) = 18.01	p <<0.001	0.07 %
Frequency:OLD20	F(2,24521) = 26.19	p <<0.001	0.10 %
Frequency:Conc	F(2,24521) = 26.80	p <<0.001	0.10 %
Frequency:Prev	F(2,24521) = 106.65	p <<0.001	0.41 %
N_{lett}:AoA	F(2,24521) = 26.90	p <<0.001	0.10 %
N_{lett}:OLD20	F(2,24521) = 76.19	p <<0.001	0.29 %
N_{lett}:Conc	F(2,24521) = 0.78	p = 0.416	0.00 %
N_{lett}:Prev	F(2,24521) = 13.23	p <<0.001	0.05%
AoA:OLD20	F(1,24521) = 1.73	p = 0.187	0.00 %
AoA:Conc	F(1,24521) = 39.83	p <<0.001	0.08 %
AoA:Prev	F(1,24521) = 54.68	p <<0.001	0.10 %
OLD20:Conc	F(1,24521) = 2.76	p = 0.096	0.01 %
OLD20:Prev	F(1,24521) = 32.87	p <<0.001	0.06 %
Conc:Prev	F(1,24521) = 0.01	p = 0.913	0.00 %

[14] A complicating factor here is that the rcs() function does not center variables well. This requires some calculations by hand. An alternative is to work with polygons.

Figure 1

Practice effects in the Dutch Lexicon Project 2 for accuracy (left panel) and Reaction times (right panel)

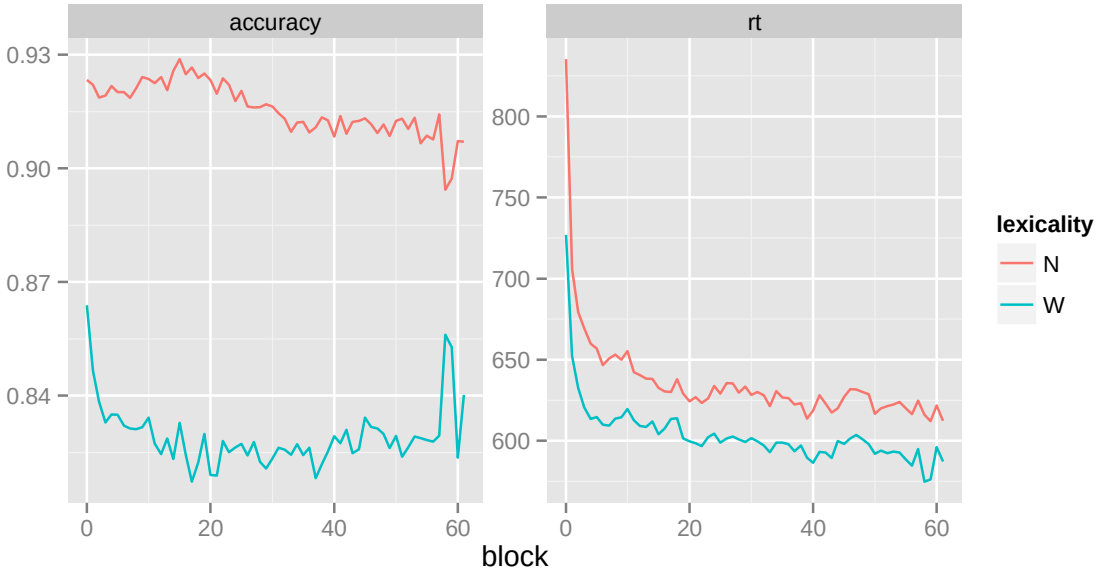

Figure 2: The two-way interactions between AoA and the other predictors. The plots are filtered so that we do not compute predictions for the whole grid of combinations between two variables, but only for combinations that are likely to occur. For instance, no words were available in the stimulus list that were very frequent and late acquired; these parts of the space are left out of the plots.

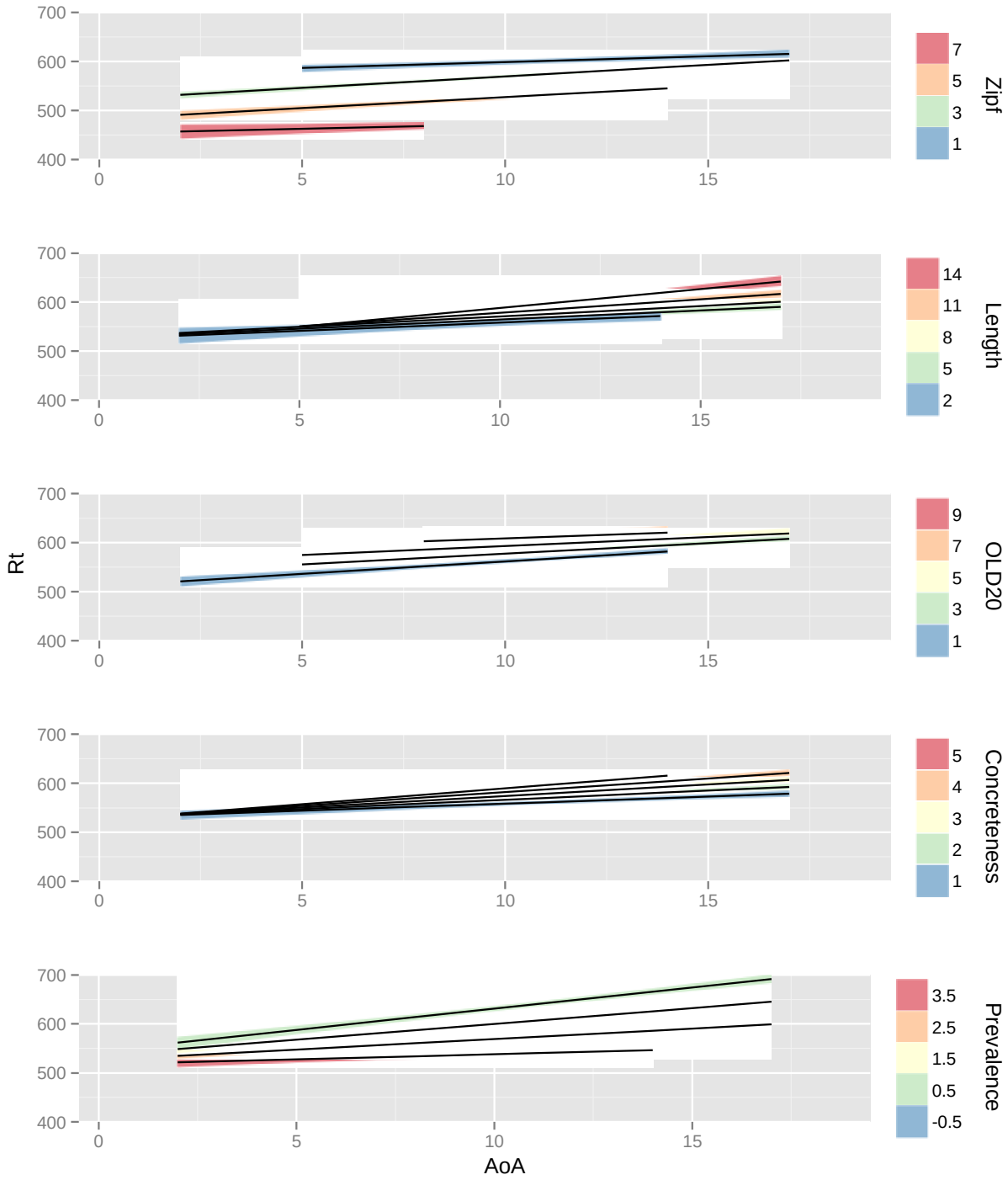

Figure 3: Two-way interactions between Prevalence and the other predictors

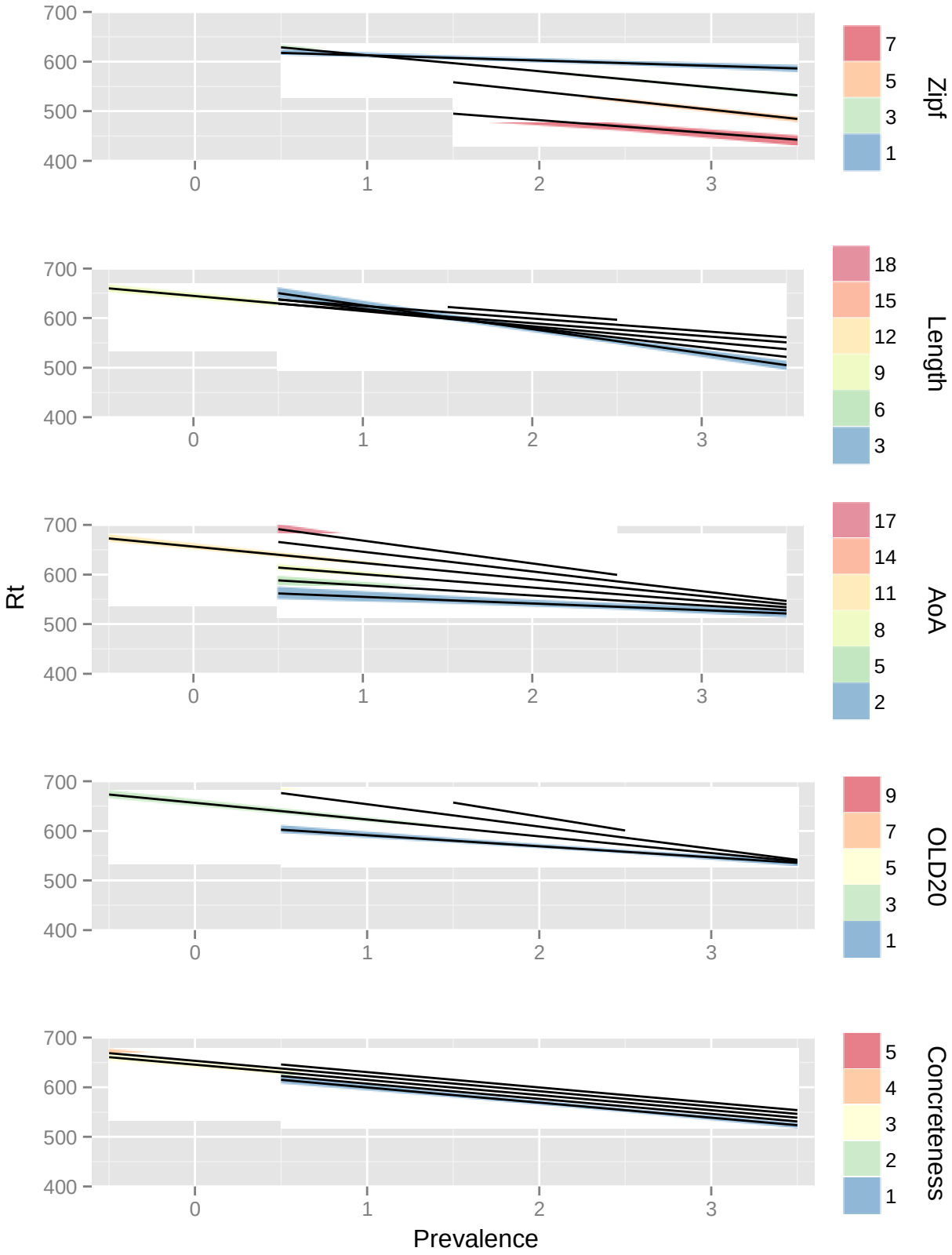

Figure 4: The main effects of word frequency, length, AoA, OLD20, concreteness, prevalence, and PoS.

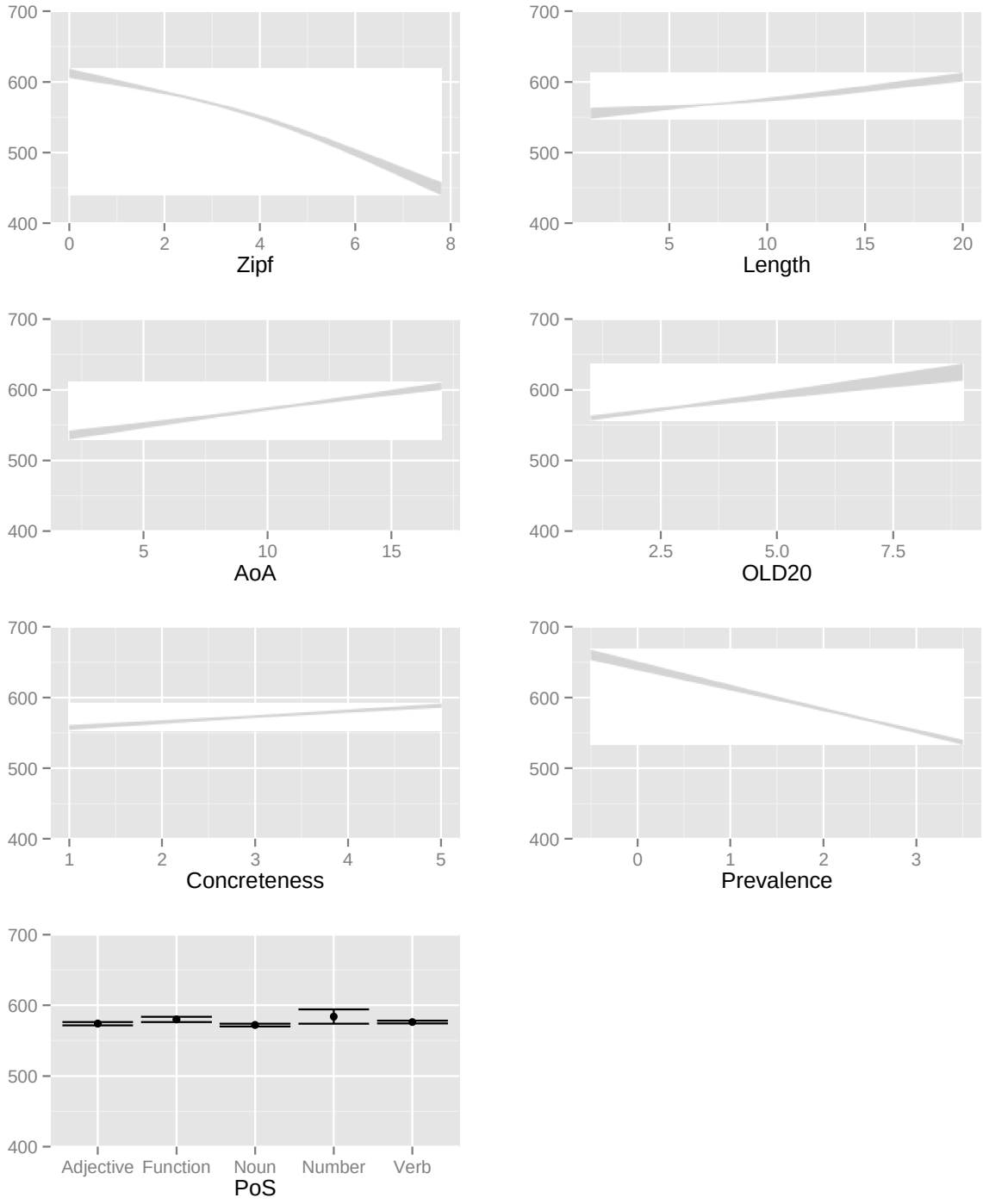

Figure 5: The distribution of word prevalence measures in the analyses run. Remark that most words were known to 95% and more of the people (prevalence values above 1.65)

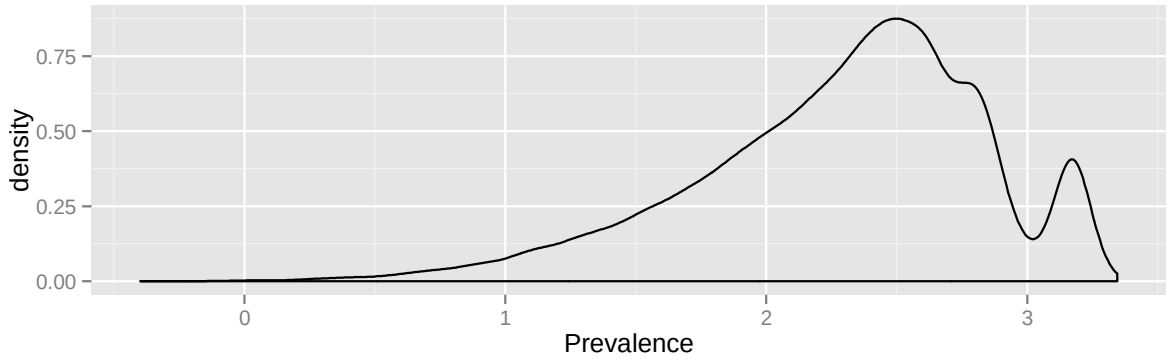

Figure 6: The interaction between word concreteness and word frequency.

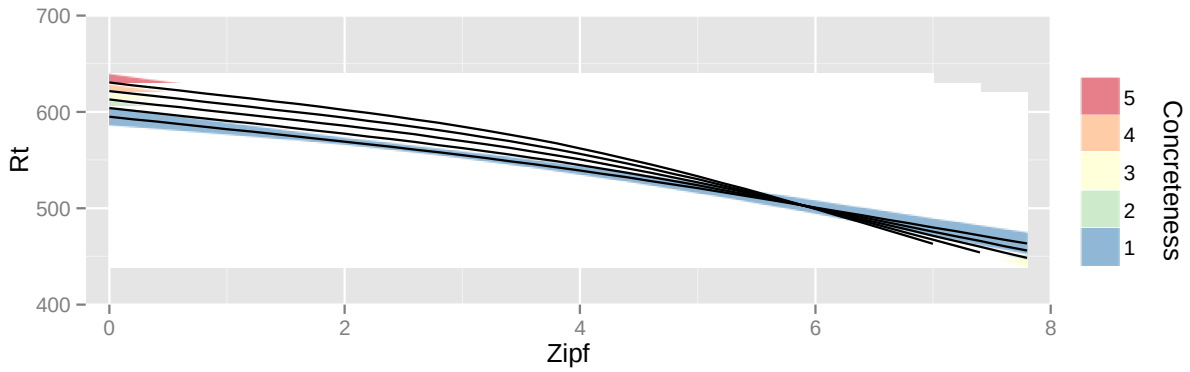

Figure 7: Two-way interactions between OLD20 and the other predictors.