Behav Res (2012) 44:287–304
DOI 10.3758/s13428-011-0118-4

The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words

Emmanuel Keuleers · Paula Lacey · Kathleen Rastle · Marc Brysbaert

Published online: 1 July 2011

Abstract We present a new database of lexical decision times for English words and nonwords, for which two groups of British participants each responded to 14,365 monosyllabic and disyllabic words and the same number of nonwords for a total duration of 16 h (divided over multiple sessions). This database, called the British Lexicon Project (BLP), fills an important gap between the Dutch Lexicon Project (DLP; Keuleers, Diependaele, & Brysbaert, *Frontiers in Language Sciences. Psychology, 1,* 174, 2010) and the English Lexicon Project (ELP; Balota et al., 2007), because it applies the repeated measures design of the DLP to the English language. The high correlation between the BLP and ELP data indicates that a high percentage of variance in lexical decision data sets is systematic variance, rather than noise, and that the results of megastudies are rather robust with respect to the selection and presentation of the stimuli. Because of its design, the BLP makes the same analyses possible as the DLP, offering researchers with a new interesting data set of word-processing times for mixed effects analyses and mathematical modeling. The BLP data are available at http://crr.ugent.be/blp and as Electronic Supplementary Materials.

Keywords Visual word recognition · Lexical decision · Megastudy · British English · Reaction times · Trial level analysis · Virtual experiments

Electronic supplementary material The online version of this article (doi:10.3758/s13428-011-0118-4) contains supplementary material, which is available to authorized users.

E. Keuleers (✉) · M. Brysbaert
Department of Experimental Psychology, Ghent University,
Henri Dunantlaan 2,
B-9000 Gent, Belgium
e-mail: emmanuel.keuleers@ugent.be

P. Lacey · K. Rastle
Department of Psychology, Royal Holloway, University of London,
London, UK

The last decade has seen an increasing effort in creating and analyzing large data sets of behavioral word-processing data. So far, a considerable amount of word-processing data has been made available to the research community (Table 1).

These large-scale data sets have many applications, including the following:

- Continuous variables can be treated as such, allowing examination of effects over an entire range (e.g., investigating the word frequency effect is not limited to a comparison of high- vs. low-frequency words).
- The relative importance of various word and task characteristics can be determined by looking at explained variance (e.g., word frequency explains most of the variance in lexical decision times, whereas the first phoneme explains most of the variance in naming times of monosyllabic words).
- Researchers can run *virtual experiments* on the data set, to evaluate new hypotheses, to check the reliability and generality of their own findings, or to better control their stimulus sets.
- The word-processing data can be used to evaluate computational models of word recognition and to assess the impact and the quality of word indices (e.g., different measures of word frequency).
- Mathematical psychologists can use the data sets to test and develop models of binary decisions (e.g., by analysis of the reaction time [RT] distributions of lexical decisions).

As is true for all aspects of scientific research, the availability of multiple word recognition data sets is a strength rather than a weakness, since the data sets differ in task, language, and experimental design, decreasing the risk of idiosyncratic findings from a single database.

At the same time, researchers should be aware of the ways in which the data sets differ from each other. For instance, the lexical decision study performed by Balota et

Table 1 List of visual word recognition megastudies

Source	Task	Material	Participants
Seidenberg and Waters (1989)	Naming	2,897 monosyllabic English words	30 students
Treiman, Mullenix, Bijeljac-Babic, and Richmond-Welty (1995)	Naming	1,329 English monosyllabic CVC words	27 students
Spieler and Balota (1997)	Naming	2,820 monosyllabic English words	31 students
Spieler and Balota (2000)	Naming	2,820 monosyllabic English words	29 older adults (mean age = 73 years)
Chateau and Jared (2003)	Naming	1,000 disyllabic six-letter words	29 undergraduate students
Balota et al. (2004)	Lexical decision	2,906 English monosyllabic words	30 students and 30 older adults (mean age = 74 years)
Balota et al. (2007)	Lexical decision and naming	40,481 English words	400 students (naming), 816 students (lexical decision)
Lemhöfer et al. (2008)	Progressive demasking identification	1,025 English words	20 native speakers, and three groups of bilinguals with English as L2
Ferrand et al. (2010)	Lexical decision	38,840 French words	975 students
Keuleers, Diependaele, & Brysbaert (2010)	Lexical decision	14,037 Dutch monosyllabic and disyllabic words	39 students and university personnel

al. (2007) differs from those by Ferrand et al. (2010) and Keuleers, Diependaele, & Brysbaert (2010) in language (English vs. French and Dutch), stimulus presentation (uppercase vs. lowercase), and type of nonwords used (manual vs. statistical construction). Keuleers, Diependaele, and Brysbaert's study additionally differs from the other studies in experiment design and duration. As long as data analysis of the different studies yields comparable results, one can confidently conclude that the differences are irrelevant for the topic being studied. However, when there are discrepancies between the results, it becomes nearly impossible to trace their origin.

To fill the gap between the various databases, we ran a new megastudy in English, using the design of Keuleers, Diependaele, & Brysbaert (2010). By comparing the new data with those in the existing studies in English, we can investigate to what extent the results depend on procedural choices, and by comparing the new data with those already collected in Dutch, we can examine to what extent they depend on language. For convenience, we refer to the new study as the *British Lexicon Project* (BLP), in analogy with the *English Lexicon Project* (ELP; Balota et al., 2007), the *French Lexicon Project* (FLP; Ferrand et al., 2010), and the *Dutch Lexicon Project* (DLP; Keuleers, Diependaele, & Brysbaert, 2010).

Since existing data sets often fall short in the number of low-frequency words they contain, we decided to use words with frequencies as low as 0.02 per million words (see below). As a result, over 28,000 words were included in the stimulus list (note that all words were mono- or disyllabic).

One of the goals in collecting the data for the BLP was to have an English data set that would allow for straightforward analysis on individual RTs, rather than on average RTs per item. While the ELP was designed primarily to allow analysis on item means, the BLP design follows the approach taken by Keuleers, Diependaele, & Brysbaert (2010) and is better suited for analysis on individual RTs, using linear mixed effects models with crossed random effects for participants and items (Baayen, Davidson, & Bates, 2008), eliminating the need for separate participant (F_1) and item (F_2) analyses. It is also possible to run this type of analysis on the ELP data, but, due to the lack of orthogonal variation, trial-level analyses are less powerful. We will come back to this point in the discussion.

Method

Participants

A total of 78 participants completed the experiment. They were students (both undergraduates and graduates) or employees of Royal Holloway, University of London. Participants were recruited via the university Web site and word of mouth. They were informed that successful completion of the experiment would take about 16 h, for which they would receive a payment of £200. They were also informed that they had to attain a consistent accuracy level of 80%,[1] that their average RT should stay below 1 s, and that all trials had to be completed. Participants were informed that if they failed to meet these targets, they would be excluded from further participation and would be paid £5 per hour completed.

In addition to the participants who successfully completed the experiment, 27 more participants did not continue to the end, either because they did not return after

[1] In Keuleers, Diependaele, & Brysbaert (2010), the accuracy level was set at 85%. However, given the high number of very low-frequency words in the present study, this accuracy level was not thought realistic here.

the first few sessions (13) or because their performance level was consistently below 80% correct (14).

Stimuli

To select the word stimuli, we used two sources. The first source, of which all words were included in the study, was a list of 8,010 monosyllabic words with a minimal length of two letters used in the DRC model of visual word recognition (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). The second source consisted of 22,725 disyllabic words with a total frequency of at least 2 in the British National Corpus (BNC, available at http://www.natcorp.ox.ac.uk).[2] Since the BNC corpus contains 100 million words, the lower frequency bound was 0.02 words per million. The list was cleared of typing errors, acronyms, low-frequency names of people and places, and non-British English spellings, leaving us with a total of 20,720 disyllabic words.

Because it takes over 32 h to make lexical decisions to 28,700 words and the same number of nonwords, a random permutation of the list was split in two. One half was given to the participants with odd ranks; the other half was given to the participants with even ranks. Because the dropout was not the same in both groups, we ended up with 40 participants in the list with odd ranks, against 38 participants in the list with even ranks.

Nonwords were generated using Wuggy, a multilingual pseudoword generator (Keuleers & Brysbaert, 2010). For a given target word, the Wuggy algorithm generates the best corresponding nonword, given a number of criteria. For the present experiment, we used the following criteria: (1) The nonword matched the syllabic and subsyllabic structure of the target word; (2) it differed from the target word in exactly one subsyllabic segment (onset, nucleus, or coda) for monosyllabic target words and in two subsyllabic segments for disyllabic target words; (3) the transition frequencies of the subsyllabic segments of the target word were matched as closely as possible; and (4) the morphological structure of the word was retained (e.g., if the word was a plural form, we tried to make a matching pseudoplural). Nonwords were created independently for each word list, meaning that there was a small overlap. As a result, while each participant saw a particular nonword only once, some nonwords were presented to all 78 participants. Most nonwords, however, were used in only one stimulus list and, therefore, were responded to by either 40 or 38 participants.

Procedure

The experiment started with an intake session, in which participants received information about the experiment and completed a questionnaire about their reading behavior and knowledge of other languages. Participants then completed a practice session of 200 trials using trisyllabic words and matching nonwords, allowing us to demonstrate the main features of the experiment. Responses were collected using a response box connected to the USB port. Participants used their dominant hand for word responses and their nondominant hand for nonword responses.

The experiment consisted of 57 test blocks of 500 trials and a final block of 230 trials. A trial consisted of the following sequence of events. First, two vertical fixation lines appeared slightly above and below the center of the screen, with a gap between them wide enough to clearly present a horizontal string of letters. Then participants were asked to fixate on the gap as soon as the lines appeared. A stimulus was presented 500 ms later in the gap between the vertical lines, which stayed on-screen. Following the response, the screen was blanked for 500 ms, after which the next trial started. We did not impose a time limit for responses, thus allowing us to collect very long RTs and limiting the loss of data from trials due to inattention.

After every 100 trials (about 3 min), a pause-screen was presented that gave participants information about their progress in the block and gave them the opportunity to take a short break if needed. After each block of 500 trials, participants received feedback about the percentage of correct trials for that block.

After the intake session, participants were free to organize their subsequent sessions by signing up for available time slots. Five experimental computers were used in parallel, and participants could choose to sit at any available computer. After entering a registration code, they were presented with a screen displaying their name and the number of their last completed block. After the participant confirmed this information, the next block started. Every evening, data were copied to a central computer, which automatically generated a review sheet, allowing the experimenters to verify participants' progress and performance. Display and timing routines for the experiment were programmed in C using the Tscope library (Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006).

Results

The results section starts with a general overview of the data. Unless indicated otherwise, analyses are limited to word trials.

Descriptive statistics

For each word, three dependent variables were defined: (1) the percentage of correct responses, or accuracy, calculated

[2] This list was kindly provided to us by James Adelman.

on all 38 or 40 participants; (2) the mean RT of the correct word responses; and (3) the mean standardized RT (*z*RT) of the correct word responses, calculated on the *z*-scores of the word RTs per participant and per block. Before computing mean RTs and *z*RTs, 2.3% of outliers were removed. Outliers were defined per participant and block, using a method commonly applied for box plots: First the interquartile distance (the distance between quartile 1 and quartile 3) was computed; RTs were then defined as outliers when they were higher than 3 interquartile distances above quartile 3 or lower than 3 interquartile distances below quartile 1. Since there were no time limits for responses in our study, this method, which is robust to the influence of extreme outliers, is particulary suitable. Of course, other researchers are invited to use their own choice of trimming method on the raw data.

Table 2 lists the summary statistics of the items. For comparison purposes, the statistics of the mono-and disyllabic words of the ELP and the DLP are included as well. From this table, the similarities between the BLP and the DLP are clear. The main differences are the higher error rates to the words and the lower RTs to the nonwords in the BLP. The main reason for this, arguably, is that the BLP included more very low frequency words, which the participants experienced as nonwords. In both the BLP and the DLP, participants' responses to words were 80 ms faster, on average, than in the ELP (the RTs of which, in general, are longer than in published studies; see below).

Practice effects

Figure 1 displays the effect of practice by plotting the average accuracy and RT over blocks. For RT, the effect is on the order of 100 ms on the word trials, which is larger than the effect observed in the DLP (where it was 40 ms). In addition, participants' response pattern to words and nonwords seems to have shifted during the experiment. Whereas the beginning of the experiment showed the usual pattern of longer RTs to nonwords than to words, around block 16 responses to nonwords became faster than responses to words. In our opinion, this is because a reasonably large number of words (up to 25%) were perceived as nonwords, so that participants had the impression that the experiment contained more nonword trials than word trials and adapted their response bias accordingly. In this respect, a study by Wagenmakers, Ratcliff, Gomez, and McKoon (2008, Experiment 2) may be particularly informative. These authors showed that in a lexical decision task with 25% nonwords and 75% words, responses to words were faster than responses to nonwords, whereas in a task with 75% nonwords and 25% words, the difference was reversed, with a similar effect for error rates. Interestingly, the word frequency effect remained the same: 103 ms in the 25% nonword condition versus 109 ms in the 75% nonword condition. By means of analyses with the diffusion model, Wagenmakers et al. showed that while the percentage of nonword trials did not change word-processing latencies, it altered the participants' response criteria. In terms of the diffusion model, it affected the starting point of the diffusion process and the separation between the decision criteria.

While the practice effect in our study has an intrinsic interest, researchers interested purely in word processing may prefer to partial out the effect. An easy way to do this is to use the normalized RTs (*z*RTs), which have a mean of 0 per block. Alternatively, time-specific variables, such as block number, can be entered as covariates in the statistical analysis.

Reliability of the dependent variables

The simplest way to determine the reliability of a variable is to calculate the split-half correlation and attenuate it for length, a method that has been applied for the FLP (Ferrand et al., 2010) and DLP (Keuleers, Diependaele, & Brysbaert 2010). Using this method, the reliability of the BLP word responses is .72 for RT (there is a difference of less than .01 between the odd and the even groups, due to larger number of participants in the former), .81 for *z*RT, and .96 for accuracy. These values are very similar to those obtained in DLP. The most likely explanation for the increased reliability of *z*RT, as compared with RT, is that the *z*RTs are less sensitive to the effects of practice.

Since lexical decision experiments usually have a rather high number of missing data (due to the errors made), Courrieu, Brand-D'Abrescia, Peereman, Spieler, and Rey (2011) proposed evaluating reliability using intraclass correlation (ICC), a method that is less sensitive to missing data. With the method described by Courrieu et al., a reliability of .82 was obtained for RT. For *z*RT it was .87, and for accuracy .96 (reliabilities were nearly equivalent for the odd and even groups of participants).[3] In line with Courrieu et al.'s analysis, the ICC reliabilities are higher for RT and *z*RT, but not for accuracy (where there were no missing data). Courrieu et al. also proposed the Expected Correlation Validation Test (ECVT) to test whether the ICC method is valid for a given data set by comparing the expected and observed ICCs for different numbers of participants. When applied to our data, the expected and observed correlations were indistinguishable. Figure 2 shows the results of the ECVT for *z*RTs for the odd group of participants. Similar results were obtained for the even group and for RTs.

[3] The authors thank Pierre Courrieu for kindly agreeing to run the analyses and the Expected Correlation Validation tests.

Table 2 Comparison of the British Lexicon Project (BLP) with the Dutch Lexicon Project (DLP) and the monosyllabic and disyllabic words of the English Lexicon Project (ELP)

	BLP	DLP	ELP (mono + di)
Number of words	28,730	14,034	22,143
Length (characters)	6.5 (2–13)	6.3 (2–12)	6.5 (1–13)
Length (syllables)	1.7 (1–2)	1.8 (1–2)	1.7 (1–2)
SUBTLEX frequency[a]	31.5 (.02–41,857)	59.7 (0.02–39,883)	42.6 (0.02–41,857)
Accuracy words	77% (0–100)	84% (0–100)	85% (0–100)
RT words	654 (300–1,617)	654 (312–1382)	730 (415–1,589)
Accuracy nonwords	94% (0–100)	94% (2–100)	88%[b]
RT nonwords	639 (444–1,159)	674 (508–1,135)	856[b]

[a] SUBTLEX frequencies refer to word form frequencies calculated on a corpus of 40–50 million words from film and television subtitles. Frequencies are expressed as frequency per million words. English frequencies are from Brysbaert and New (2009); Dutch frequencies are from Keuleers, Brysbaert, and New (2010). For the BLP words, there were SUBTLEX frequencies for only 25,316 words, partly because of spelling differences between British and American English. Therefore, unless indicated otherwise, for the analyses reported in this article, we used the BNC frequencies, which had an average of 26.9 per million and ranged from 0.01 to 61,879 per million.

[b] Based on the full ELP

Correlations between the BLP and ELP

The most straightforward way to compare the BLP with the ELP is to correlate the various dependent variables. There were 18,969 words in common between the BLP and ELP.[4] In order to interpret the intercorrelations shown in Table 3, it is useful to know that the reliability as calculated above gives an estimate of how much a variable would correlate with itself if the study were repeated. The reliability of .81 for *z*RT means that one can expect a correlation of .81 between *z*RT and the *z*RT calculated on a new, similar study. If we look at the obtained correlations, we see that the correlations for standardized RTs (r = .77) and for accuracy scores (r = .79) approach these ceiling levels. The correlation for *z*RTs remains high even when the data are limited to those words that are known to two thirds of the participants in both the British and the American studies (r = .73, N = 15,241).

Table 4 gives the 40 words with the highest residuals when the ELP *z*RTs are regressed on the BLP *z*RTs, for words that were known by at least two thirds of both groups of participants. Apart from some typical British and American words, we see that the ELP participants were faster on names, possibly because, in the ELP, words were presented in capitals (HOMER), whereas in the BLP, they were presented in lowercase letters (homer).

Correlation with the Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) data

Balota et al. (2004) collected lexical decision times for 2,906 monosyllabic words from a group of young adults very similar to the participants for the BLP. Table 5 lists the correlations between the BLP and the Balota et al. (2004) data. The two studies had 2,328 words in common. Despite the smaller range of some variables (e.g., fewer very low frequency items in this database, all words monosyllabic), the correlations range between .6 and .7.

In addition to examining correlations between similar variables, it is interesting to correlate RTs and accuracies with some of the major predictors of lexical decision performance: word length, word frequency (SUBTLEX; Brysbaert & New, 2009), age of acquisition (AoA), word familiarity, and imageability. These correlations, based on measurements collected by Cortese and Khanna (2007, 2008) are listed in Table 6 and clearly show the overwhelming similarity between the Balota et al. (2004) lexical decision data set and the BLP data.

Figure 3 shows the frequency effect in the BLP and ELP for the words present in both databases. Stimuli were binned in groups of 1,000, and the means are given. The figure also shows the standard deviation of the RTs in each bin.

In Fig. 3, we see that the frequency effect is very similar for the BLP and ELP but that RTs were up to 100 ms shorter in the BLP than in the ELP. Also, the standard deviations of the RTs were larger in the ELP than in the BLP. In Fig. 4, the same data are plotted using standardized RTs, showing that for this variable, the frequency effect is more pronounced in the BLP than in the ELP. When interpreting these data, it is important to keep in mind that the ELP *z*RTs were calculated on all 40 K words present in that database, including the long words of more than two syllables. As a result the *z*RTs of the monosyllabic and disyllabic words in the ELP are lower than those in the BLP.

Finally, Fig. 5 shows the effect of word frequency on accuracy. Here, too, the effect is more pronounced for the

[4] Note that this analysis excludes words written differently in the two databases (e.g., *labor* vs. *labour*).

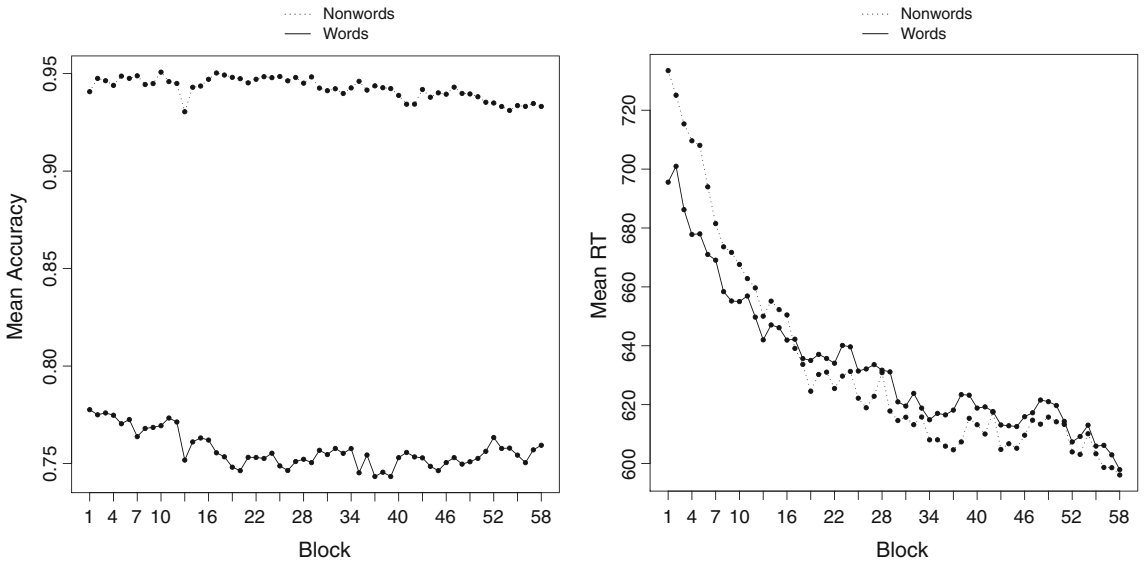

Fig. 1 The effects of practice on accuracy (*left panel*) and response latency (*right panel*)

BLP—in particular, at the very low frequency end (below 0.5 per million).

Virtual experiments

The BLP approach, in which two groups of participants each see half of the stimuli, allows for straightforward generalization across stimuli and across participants. Therefore, we can run virtual experiments by extracting the relevant information from the full database (containing the data of the individual participants) and running a linear mixed effects analysis (see also Keuleers, Diependaele, & Brysbaert 2010).

Below, we compare some classical findings in English word recognition research with those for virtual experiments, using the BLP data.

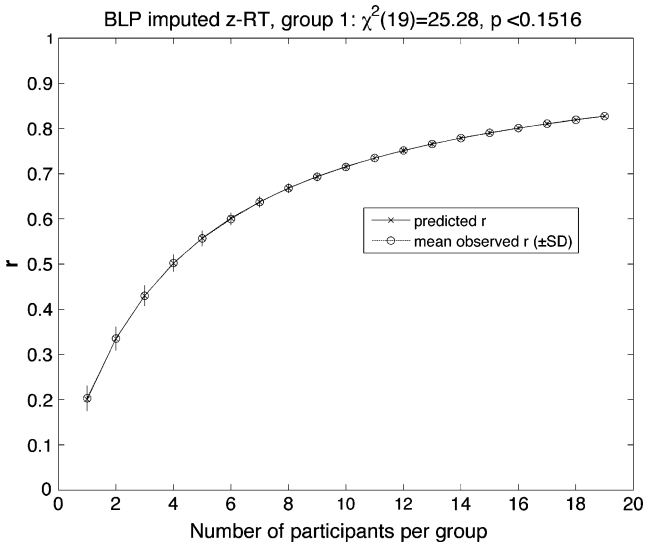

Fig. 2 ECVT test (Courrieu et al., 2011) on *z*RTs data of Group 1 ($n = 38$), after imputation of 26% missing data by the CRARI algorithm (Courrieu & Rey, 2011). The "predicted" and "observed" curves are indistinguishable, and the χ^2 test does not detect a significant difference between them. Therefore, the ICC method can be considered valid for these data

Word frequency As was shown above, there is a healthy frequency effect in the full BLP data. In Table 7, we examine how the RT effects of frequency from some classical studies compare with effects found using virtual experiments involving the same items from the BLP.

Table 7 illustrates that the frequency effect in the BLP is similar to that of published experiments. As will become clear below, this is true only for experiments with short RTs. In experiments with longer RTs, the frequency effect tends to be more pronounced than in the BLP. An interesting study in this context was published by Yap, Balota, Tse, and Besner (2008, Experiments 2–4). They examined the size of the frequency effect as a function of vocabulary size, by comparing the results of three universities with different student populations.[5] As can be seen in Table 8, larger vocabulary sizes corresponded with shorter RTs and smaller frequency effects. Therefore, the most likely explanation for the short RTs and the relatively small size of the frequency effect is that the participants we tested, on average, had a large vocabulary size, which seems plausible given the nature of our experiment and the fact that we required them to have an accuracy level above 80%.

In a similar vein to Yap et al. (2008), Chateau and Jared (2000) provided evidence that the frequency effect is

[5] Because the nonwords used in these studies were pseudohomophones (e.g., *brane,* which sounds like the existing word *brain*), they were not included in Table 7.

Table 3 Correlations between the word data of the British Lexicon Project (BLP) and English Lexicon Project (ELP; lexical decision)

	BLPzrt	BLPacc	ELPrt	ELPzrt	ELPacc
BLPrt	.954	-.685	.679	.730	-.588
BLPzrt		-.767	.710	.770	-.656
BLPacc			-.580	-.653	.788
ELPrt				.937	-.595
ELPzrt					-.690

All correlations are significant at the .0001 level (N = 18,969)

Table 5 Correlations between the word data in the BLP and Balota et al. (2004), young adults;[B04])

	BLP zRT	BLP Accuracy	B04 RT	B04 Accuracy
BLPrt	.968	-.682	.693	-.589
BLPzrt		-.728	.716	-.616
BLPacc			-.589	.612
BAL04rt				-.589

All correlations are significant at the .0001 level (N = 2,328).

mediated by print exposure, as measured with a reading test and an author recognition test. In their experiment, the high print exposure group showed a frequency effect of 128 ms (537 vs. 665 ms), whereas the low-exposure group had a frequency effect of 278 ms (618 vs. 896 ms). Interested in seeing whether the effect found by Chateau and Jared's (2000) experiment was caused by their use of pseudohomophones as nonwords, Sears, Siakaluk, Chow, and Buchanan (2008) compared a condition using pseudohomophonic nonwords with a condition using typical pseudowords (e.g., *brint*). Whereas the difference in the frequency effect between print exposure groups was found using pseudohomophonic nonwords, the interaction between print exposure and the frequency effect was not present in the experiment with legal nonwords. However, the participants in Sears et al. (2008) were fast readers with small frequency effects, overall, in line with those for the BLP participants. We will return to the Chateau and Jared (2000) and Sears et al. (2008) studies below when we discuss the effect of orthographic neighborhood size (Table 11).

Table 7 further shows that the interaction between frequency and animacy found by Monsell, Doyle, and Haggard (1989) was not present in the BLP data. This is interesting, because Monsell et al. did not expect to find this interaction effect. As they wrote, "For unknown reasons, possibly to do with our selection of items, the frequency effect was significantly weaker for the thing nouns than for the person nouns" (p. 55). The BLP data show that the unexpected interaction indeed cannot be replicated using megastudy data, in which participants respond to a large numbers of words and nonwords. On the other hand, we did obtain a significant interaction between frequency and the stressed syllable position (note the same trend in the data in Monsell et al., 1989). Why the frequency effect is smaller in the words with final stress, which as a group are less common than words with initial stress, is at present not clear.

Table 4 Words with the largest residual difference in *z*RT between the British Lexicon Project (BLP) and English Lexicon Project (ELP)

BLP Much Faster	BLP Much Slower
nightie	homer
greaseproof	lincoln
offence	boston
postcode	johnny
catchphrase	mom
sulphate	speedway
oxtail	sears
wholemeal	roger
levelled	softball
signposts	plato
gasworks	lawless
ferries	farthest
heartstrings	ralph
instructs	butts
transience	wick
tongs	dean
strengthens	heather
defence	peter
yachtsman	singed
drainpipe	babes
moulding	tooling

Table 6 Correlations between word characteristics and average RTs and accuracies for Balota et al. (2004), young adults; [B04]). and the British Lexicon Project (BLP; words for which all data are available; N=2,328)

	RT		Accuracy	
	B04	BLP	B04	BLP
Length	.092**	.147**	.022	.039
Frequency	-.598**	-.617**	.414**	.456**
AoA	.649**	.645**	-.501**	-.500**
Familiarity	-.605**	-.608**	.445**	.454**
Imageability	-.274**	-.273**	.303**	.266**

AoA, age of acquisition.

** $p < .01$

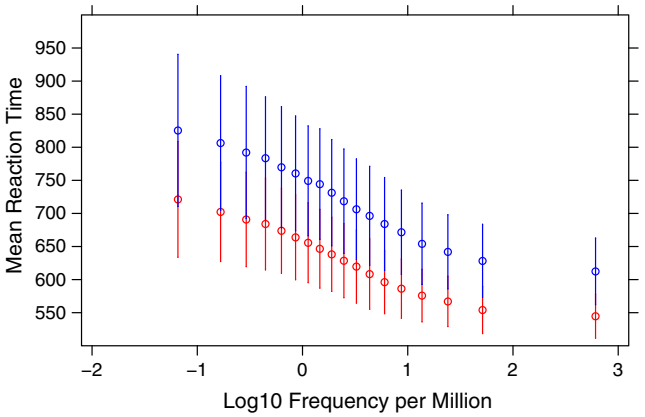

Fig. 3 The frequency effect in the English Lexicon Project (*blue lines*) and the British Lexicon Project (*red lines*). Stimuli were binned by frequency in groups of 1,000. Each line shows the mean RT and the standard deviation for a bin. Frequencies were based on British National Corpus. Log10 frequency is the log10 of the frequency per million (i.e., for 0.1 per million, its value is -1.0)

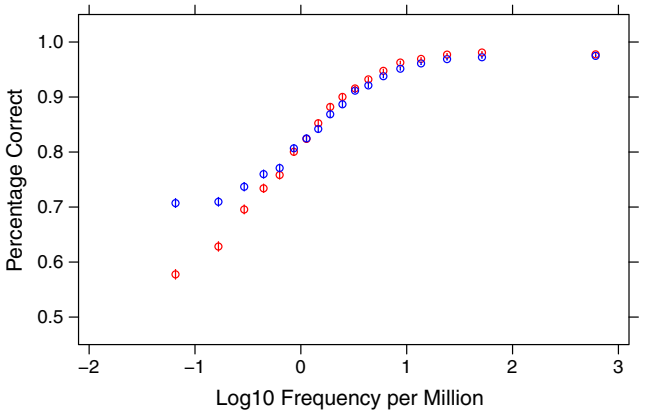

Fig. 5 The word frequency effect on accuracy in the English Lexicon Project (*blue*) and British Lexicon Project (*red*)

Age of acquisition It has repeatedly been claimed that early-acquired words are processed faster than late acquired words, independently of word frequency. As Table 9 illustrates, the BLP data show the same results, and the size of the age-of-acquisition effect is the same as that reported in published experiments, even when the latter had longer RTs.

Regular versus irregular words An important topic in visual word recognition is the interaction between word frequency and the consistency of the letter–sound mappings. It is well established that in word naming, the cost of an irregular or inconsistent mapping is higher for low-frequency words than for high-frequency words (Seidenberg, Waters, Barnes, & Tanenhaus, 1984). There

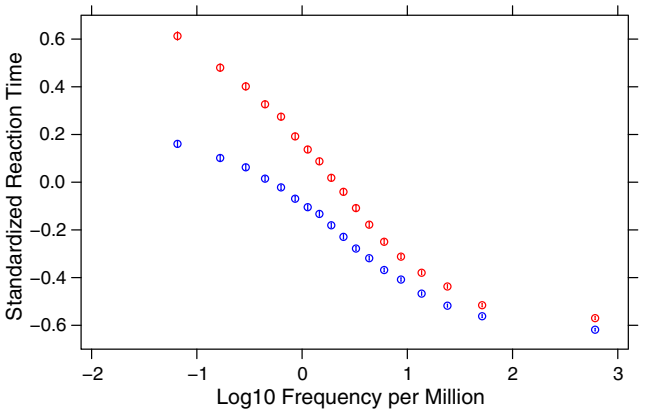

Fig. 4 The word frequency effect in the English Lexicon Project (*blue*) and the British Lexicon Peojct (*red*) when expressed in standardized reaction times. The whiskers represent standard errors

is more discussion about whether the same interaction exists for lexical decision times, with the "classical" studies giving null effects. Table 10 summarizes the findings. As can be seen, the expected interaction does not show up in the BLP data. For the stimuli of Hino and Lupker (1996), the regularity effect even seems to be the reverse of the expected one. The only effect that was replicable is that of feedback inconsistency reported by Stone, Vanhoy, and Van Orden (1997, Experiment 2): Participants needed more time to accept a word that is feedback inconsistent, meaning that its pronunciation body can be spelled in more than one way (for instance, the rime of a syllable with the sound /ip/ can be spelled *–eep*, as in *deep* or *–eap,* as in *heap*). Interestingly, this was the case only for spelling patterns that are feedforward consistent, meaning that they are unambiguous in their pronunciation (e.g.,–eap is always pronounced /ip/, while *–eat* is pronounced differently in *sweat* and in *heat*). Surprisingly, the feedback consistency effect reported by Stone et al. (1997, Experiment 1) was not found with the BLP virtual experiment.

Neighborhood size Another important topic in word recognition research is the way in which a word's recognition is affected by its orthographic neighbors—that is, the words that differ from the target word in one letter at one position. When a person sees the input word *lost*, its orthographic neighbors *most*, *list*, *loft*, and *lose* are thought to become activated as well and to compete with the target word in the word recognition process. Andrews (1992) found that words with a large neighborhood size (*N*) were responded to more quickly in lexical decision than were words with a small *N*—in particular, if they were low-frequency words. This finding was unexpected given that neighbors were assumed to compete with each other and, thus, to have inhibitory effects (Segui & Grainger, 1990). A possible

Table 7 Published lexical decision experiments involving frequency effects and virtual experiments with the same stimuli from the British Lexicon Project (studies are chronologically ordered)

		Original Experiment RTs	Virtual Experiment RTs
Monsell, Doyle, and Haggard (1989, Experiment 1)	High frequency, person	538	535
	High frequency, thing	541	539
	Medium frequency, person	553	570
	Medium frequency, thing	570	565
	Low frequency, person	639	640
	Low frequency, thing	617	618
	Effect of frequency	88**	92**
	Effect of animacy	1	8
	Frequency × animacy interaction	$p < .01$	n.s.
Monsell et al. (1989, Experiment 3)	High frequency, initial stress	538	541
	High frequency, final stress	543	551
	Low frequency, initial stress	642	646
	Low frequency, final stress	616	598
	Effect of frequency	89**	77**
	Effect of stress	10	19^{+}
	Frequency × stress Interaction	15	29**
Morrison and Ellis (1995, Experiment 6)	High frequency	548	542
	Low frequency	602	576
	Effect of frequency	54**	34**
Yap et al. (2008, Experiment 1)	High frequency	557	531
	Low frequency	605	574
	Effect of frequency	48**	43**

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$ or significant only in F_1 or F_2

explanation of this unexpected finding was that words with many neighbors look more wordlike than do words with few neighbors and elicit a word response on the basis of the general activation in the lexicon, rather than on the basis of recognition of the precise target word (Grainger & Jacobs, 1996).

Table 11 lists the most important studies on the topic, together with the results of the virtual experiments. From this table, we can conclude that, overall, the BLP data do not show significant effects of neighborhood size. At the same time, the original studies on the neighborhood size effect showed considerably stronger frequency effects than did the virtual experiments, not all of which can be explained by differences in overall RTs. This may indicate that lexical decision experiments including only words with extreme values of a particular dimension (high vs. low frequency, many vs. few neighbors) tend to exaggerate the importance of this dimension, relative to studies including words from the entire continuum.

Yates and colleagues (Yates, 2005, 2009; Yates, Locker, & Simpson, 2004) suggested that a word's phonological neighbors are more important than its orthographic neighbors. Table 12 shows the results of the various experiments run by Yates and colleagues, together with the simulations. Contrary to the orthographic neighborhood data, the phonological neighborhood data were observed mostly in the virtual experiments, although, again, the effects tended to be smaller than in the original studies. Only the most

Table 8 Data from Yap et al. (2008), illustrating the frequency effect as a function of the vocabulary size

University	Age	Years of Education	Vocabulary Age	RT_{HF}	RT_{LF}	Effect
Washington University	20.9	13.8	18.7	612	678	66
University of Waterloo	20.9	NA	17.7	658	753	95
University at Albany (SUNY)	19.4	12.2	16.9	732	844	112

Table 9 Age of acquisition (AoA) effects in published lexical decision experiments and in virtual experiments with the same stimuli as those from the British Lexicon Project

		Original Experiment RTs	Virtual Experiment RTs
Morrison and Ellis (1995, Experiment 5)	Early acquired	582	552
	Late acquired	648	604
	Effect of AoA	66**	52**
Gerhand and Barry (1998, Experiment 1)	Early acquired, high frequency	593	540
	Early acquired, low frequency	621	538
	Late acquired, high frequency	603	584
	Late acquired, low frequency	730	623
	Effect of AoA	59**	65**
	Effect of frequency	77**	19
	Frequency × AoA Interaction	50**	20^{+}

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$ or significant only in F_1 or F_2

recent study (Yates, 2009), comparing the lexical decision times to monosyllabic words with neighbors for all three subsyllabic segments ($P = 3$) with those to words with neighbors for just two subsyllabic segments ($P = 2$), did not reach significance in the virtual experiment.

Bigram frequency The frequency of a word's letter bigrams is often considered to affect word recognition and is, therefore, often controlled for in experiments. Surprisingly, in the only study we could find on this variable (Andrews, 1992, Experiment 3), words with familiar bigrams were not responded to faster than words with rare bigrams. The same pattern is observed in the BLP (Table 13).

Polysemy The final topic we will address is the effect of polysemy—that is, the existence of different meanings for the same word. As can be seen from Table 14, word recognition research has suggested a small facilitation effect of number of meanings. Although there is a similar tendency in the BLP, none of the effects reached significance in the virtual experiments.

Rodd, Gaskell, and Marslen-Wilson (2000) criticized the existing research on polysemy in word recognition because it failed to make a distinction between words with multiple related *senses* (e.g., the adjective *uniform* [similar in form] and the noun *uniform* [clothing worn by a particular group]) and words with multiple unrelated *meanings* (e.g., *bank* [financial institution] and bank [land alongside a river]). Rodd et al.'s results suggested that the effect of polysemy was limited to words with true polysemy (multiple related *senses*). As Table 15 shows, the same tendency is found in the BLP data, but the effects are again smaller than in the original experiment. In addition, the BLP data suggest that both number of meanings and number of senses impact word recognition.

Discussion

We started our analysis of the BLP data by documenting the high correlation between the BLP and the ELP data. As a matter of fact, the correlation between the two data sets is close to the maximum that can be expected given the reliability of the scores in the individual databases, meaning that the methodological differences do not entail divergent results. This is useful for researchers who want to run new megastudies (e.g., in other languages).

The high correlation between the BLP and ELP is of further interest because it addresses an issue raised by Sibley, Kello, and Seidenberg (2009). These authors noted that word naming megastudies did not replicate the frequency × irregularity interaction found in small-scale factorial experiments and ventured that this was because megastudies contained too much noise to allow fine-grained analyses. As long as there is only one major megastudy available in English, this criticism cannot be examined. However, now that we have the BLP and ELP, we can take a deeper look. If Sibley et al. are right and megastudies are characterized by a larger degree of noise (rather than other strategies used by participants when confronted with a random sample of words than with a sample of target words differing on one or two dimensions), the data of the two megastudies will not agree more with each other than with the findings of small-scale factorial experiments.

To illustrate the approach, let us have a look at the data of Stone et al. (1997) on the effects of feedforward and feedback spelling–sound consistency in lexical decision (Table 10). For these data, we observed that BLP failed to replicate the feedforward consistency effect but presented evidence in favor of the feedback consistency effect. If the null effect of feedforward consistency is due to poor power

Table 10 The interaction between word frequency and spelling–sound consistency in published lexical decision experiments and in virtual experiments with the same stimuli as those from the British Lexicon Project

		Original Experiment RTs	Virtual Experiment RTs
Seidenberg et al. (1984, Experiment 2)	High frequency, regular inconsistent	584	534
	High frequency, strange	570	526
	High frequency, regular	601	534
	Low frequency, regular inconsistent	626	603
	Low frequency, strange	673	613
	Low frequency, regular	633	598
	Effect of frequency	59**	73**
	Effect of regularity	n.s.	n.s.
	Frequency × regularity interaction	$p < .05$	n.s.
Seidenberg et al. (1984, Experiment 3)	High frequency, regular	533	534
	High frequency, exception	530	564
	Low frequency, regular	601	600
	Low frequency, exception	604	593
	Effect of frequency	71**	47**
	Effect of regularity	0	-11
	Frequency × regularity interaction	3	18
Hino and Lupker (1996, Experiment 5b)	High frequency, regular	500	543
	High frequency, exception	492	521
	Low frequency, regular	573	597
	Low frequency, exception	579	571
	Effect of frequency	80**	52**
	Effect of regularity	-1	-24*
	Frequency × regularity interaction	7	2
Stone et al. (1997), Experiment 1)	Feedback consistent	774	597
	Feedback inconsistent	807	595
	Effect of feedback consistency	33*	-2
Stone et al. (1997, Experiment 2)	Feedforward consistent, feedback consistent	732	574
	Feedforward consistent, feedback inconsistent	778	620
	Feedforward inconsistent, feedback consistent	780	593
	Feedforward inconsistent, feedback inconsistent	770	604
	Effect of feedforward consistency	20^{+}	2
	Effect of feedback consistency	18^{+}	29**
	Feedforward × feedback interaction	28^{+}	18^{+}

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$ or significant only in F_1 or F_2

or to a peculiarity of the BLP, the ELP and BLP should not be more in agreement with each other than with the data of the original experiment. Table 16 shows the outcome of the analysis, and also when the analysis is based on the standardized scores (recall that these are less noisy). From this table, it is clear that the two megastudies are much more in agreement with each other than with the original study, thus adding extra weight to the megastudy findings.

The high correlation between the ELP and BLP indicates that a high percentage of variance in megastudies is systematic rather than noise (see also Rey & Courrieu, 2010). Both databases can be used in combination to verify hypotheses, to cross-check a finding with one database on the other, or to train a mathematical model on one data set and to test it on the other, thereby avoiding the issue of overfitting. In this respect, the BLP design has an extra bonus because its design allows for more precise statistical analyses of the trial data than does the random sampling design used in the ELP (Keuleers, Diependaele, & Brysbaert 2010; Rey & Courrieu, 2010). To illustrate this, we performed a Monte Carlo simulation in which we took two random samples of words from the words occurring in both the BLP and the ELP at various sample sizes (between 10 and 160 words) and added a virtual effect (between 0

Table 11 Studies addressing the effect of neighborhood size (*N*) in published lexical decision experiments and in virtual experiments with the same stimuli from the British Lexicon Project

		Original Experiment RTs	Virtual Experiment RTs
Andrews (1992, Experiment 1)	High frequency, small *N*	570	539
	High frequency, large *N*	586	535
	Low frequency, small *N*	757	642
	Low frequency, large *N*	714	625
	Effect of frequency	157**	96**
	Effect of *N*	-13	-7
	Frequency × *N* interaction	29**	6
Sears, Hino, and Lupker (1995, Experiment 1)	High frequency, small *N*	528	532
	High frequency, large *N*	509	538
	Low frequency, small *N*	587	564
	Low frequency, large *N*	577	581
	Effect of frequency	63**	28**
	Effect of *N*	15^{+}	-12^{+}
	Frequency × *N* interaction	5	6
Sears et al. (1995, Experiment 3)	High frequency, small *N*	520	535
	High frequency, large *N*	518	546
	Low frequency, small *N*	669	595
	Low frequency, large *N*	617	587
	Effect of frequency	124**	50**
	Effect of *N*	27^{+}	1
	Frequency × *N* interaction	25^{+}	10
Sears et al. (1995, Experiment 4a)	Small *N*, no higher neighbors	625	584
	Small *N*, 1 higher neighbors	585	559
	Small *N*, many higher neighbors	591	563
	Large *N*, no higher neighbors	585	554
	Large *N*, 1 higher neighbors	570	574
	Large *N*, many higher neighbors	570	557
	Effect of *N*	25*	7
	Effect of higher neighbors	27^{+}	9
	N × neighbor level interaction	n.s.	$p = .11$
Chateau and Jared (2000)[a]	High frequency, small *N*	542	539
	High frequency, large *N*	533	535
	Low frequency, small *N*	694	642
	Low frequency, large *N*	636	625
	Effect of frequency	127**	96**
	Effect of *N*	33**	-7
	Frequency x *N* interaction	24**	6
Sears et al. (2008, Experiment 1)	High frequency, small *N*	520	532
	High frequency, large *N*	517	530
	Low frequency, small *N*	567	552
	Low frequency, large *N*	548	559
	Effect of frequency	39**	24**
	Effect of *N*	11^{+}	-3
	Frequency × *N* interaction	8^{+}	4

[a] Same stimuli as in Andrews (1992) but pseudohomophones as nonwords, only the high print exposure group. $^{**}p < .01$, $^{*}p < .05$, $^{+}p < .10$ or significant only in F_1 or F_2

Table 12 The effect of phonological neighborhood size in published lexical decision experiments and in virtual experiments with the same stimuli from BLP

		Original Experiment RTs	Virtual Experiment RTs
Yates et al. (2004, Experiment 1)	Small phonological neighborhood	681	633
	Large phonological neighborhood	620	578
	Effect of neighborhood	61**	55**
Yates et al. (2004, Experiment 2)	Small phonological neighborhood	638	602
	Large phonological neighborhood	601	580
	Effect of neighborhood	37^{+}	22
Yates (2005)	Small phonological neighborhood	729	610
	Large phonological neighborhood	656	567
	Effect of neighborhood	73**	43**
Yates (2009)	$P = 2$	647	575
	$P = 3$	620	566
	Effect of P	27**	9

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$ or only significant in F_1 or F_2

and 40 ms) to the RTs of one group of items. For each combination of sample size and effect size, we ran 1,000 tests and noted how often the data showed a significant effect in an analysis on item means (a traditional F_2 ANOVA) and in a linear mixed effects model with crossed random effects for participants and items. The results are shown in Fig. 6 and confirm that larger sample sizes are required to find an effect in the ELP than in the BLP. For instance, in the item analysis, we can expect to find an effect of 40 ms with 40 items in the BLP, whereas a sample of about 70 items is required for the ELP. For the BLP, the results of the Monte Carlo simulations, using item analysis and mixed effects modeling, are nearly equivalent. For the ELP, however, the mixed effects model is less powerful than the item analysis, showing that, due to its design, the BLP is more suited to trial-level analysis than is the ELP.

Megastudies further allow researchers to reassess entire research traditions against a common framework. That is, each and every factorial experiment of which the original stimuli are known can be projected against the same set of megastudy data. This makes it possible to take away the peculiarities of the original studies and to directly compare results. The frequency effect provides a nice illustration. Although the effect is clearly present in the megastudy data, it tends to be smaller in the BLP than in many of the original studies. Two factors seem to contribute to this pattern. First, participants with a large vocabulary and/or high reading exposure tend to be faster and to show a smaller frequency effect (Chateau & Jared, 2000; Yap et al., 2008). Second, the frequency effect seems to be larger in small-scale factorial lexical decision experiments that include only low- and high-frequency words than in megastudies comprising all types of words. This finding raises the question as to what extent the use of two extreme categories exaggerates the effect under investigation, which we think is an important topic for future research, because our analyses suggest that a similar phenomenon may be responsible for the effects of orthographic neighborhood and spelling–sound consistency reported in some small-scale factorial studies.

Although virtual experiments are interesting, it must not be forgotten that the main strength of megastudy data is that they allow researchers to examine a continuous effect across the entire continuum. The frequency effect in Figs. 3 and 4 is a case in point. A cursory review of studies investigating the frequency effect suggests that, as a rule of

Table 13 Effects of frequency and bigram frequency in Andrews (1992, Experiment 3) and in a virtual experiment with the same stimuli as those in the British Lexicon Project

	Original Experiment RTs	Virtual Experiment RTs
High word frequency, high bigram frequency	592	532
High word frequency, low bigram frequency	594	531
Low word frequency, high bigram frequency	690	577
Low word frequency, low bigram frequency	686	591
Effect of word frequency	95**	52**
Effect of bigram frequency	-1	6
Word frequency × bigram frequency interaction	3	7

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$ or significant only in F_1 or F_2

Table 14 Effects of polysemy on visual lexical decision times in published lexical decision experiments and in virtual experiments with the same stimuli as those in the British Lexicon Project

		Original Experiment RTs	Virtual Experiment RTs
Borowsky and Masson (1996, Experiment 3)	Polysemous	637	555
	Monosemous	647	562
	Effect of polysemy	10^{+}	7
Hino and Lupker (1996, Experiment 1)	High frequency, polysemous	548	524
	High frequency, monosemous	561	534
	Low frequency, polysemous	613	574
	Low frequency, monosemous	626	572
	Effect of frequency	65**	44**
	Effect of polysemy	13^{+}	4
	Frequency × polysemy interaction	0	6
Pexman, Hino, and Lupker (2004, Experiment 1)	High frequency, polysemous	513	529
	High frequency, monosemous	511	531
	Low frequency, polysemous	567	564
	Low frequency, monosemous	609	570
	Effect of frequency	76**	37**
	Effect of polysemy	20*	4
	Frequency × polysemy interaction	22*	2

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$ or significant only in F_1 or F_2

thumb, researchers choose their high-frequency words from frequencies above 50–100 per million, while low-frequency words have frequencies below 5–10 per million. Looking at the frequency curve in megastudies (see also Keuleers, Diependaele, & Brysbaert 2010), it becomes clear that there is virtually no frequency effect above 50 per million in the lexical decision task and that nearly half of the frequency effect is situated below 1–2 per million. The main reason why researchers have overlooked the importance of frequency differences below 1 per million may well be that nearly all research in English has been based on the Kučera and Francis measures, which are based on a corpus of only 1 million words (Brysbaert & New, 2009).

It is our conviction that analyses across the entire continuum based on large numbers of data will be interesting for other variables as well. Each of the research topics we have raised above can be addressed by looking at the effect across the entire data set, rather than across the few handpicked words used in the original publications (see, e.g., Baayen & Milin, in press, for an example). This considerably increases the power of the design (Keuleers, Diependaele, & Brysbaert 2010). Furthermore, it allows researchers to go beyond merely determining the statistical significance of a variable. They can now see the curve of the entire effect and examine how strong the effect is in terms of explained variance.

Availability

The BLP data are available at the Web site http://crr.ugent.be/blp formatted as text, as Excel files, and as R data objects. In addition, we are making available a file of stimulus characteristics, which can be merged with the data.

Table 15 Effect of number of senses and number of meanings in Rodd et al. (2000, Experiment 2) and in virtual experiments with the same stimuli as those in the British Lexicon Project

	Original Experiment RTs	Virtual Experiment RTs
Many meanings, few senses	587	571
Many meanings, many senses	578	559
One meaning, few senses	586	560
One meaning, many senses	567	550
Effect of number of meanings	6	10
Effect of number of senses	14*	11^{+}
Interaction	5	1

** $p < .01$, * $p < .05$, $^{+}$ $p < .10$ or significant only in F_1 or F_2

Table 16 The feedforward and feedback consistency effects reported by Stone et al. (1997, Experiment 2) in the British Lexicon Project (BLP), in the English Lexicon Project (ELP), and combined. The first three columns show mean RTs, the final three columns show standardized RTs. The three final rows display the *p*-values of the effects

	Original Experiment RTs	BLP RTs	ELP RTs	*z*BLP	*z*ELP	*z*BLP + *z*ELP
Feedforward consistent, feedback consistent	732	574	634	-.406	-.505	-.455
Feedforward consistent, feedback inconsistent	778	624	718	-.057	-.248	-.152
Feedforward inconsistent, feedback consistent	780	598	669	-.236	-.386	-.311
Feedforward inconsistent, feedback inconsistent	770	617	690	-.087	-.292	-.189
p Feedforward consistency	.059	.491	.814	.353	.537	.376
p Feedback consistency	.086	.005	.003	.001	.005	.001
p Feedback × feedforward interaction	.129	.189	.066	.182	.186	.137

Item-level data

At the item level, there are 55,867 rows of data. For each stimulus (word or nonword), the following information is given.

- Spelling: the spelling of the stimulus as it was presented.
- Lexicality: whether the stimulus was a word (W) or a nonword (N).
- RT: the average RT to the stimulus (correct trials only).
- Zscore: the average standardized RT. Standardized RTs were calculated separately for all levels of participant, block, and lexicality (e.g., all RTs to correct word trials in block 1 by participant 1).
- Accuracy: average accuracy for the stimulus.
- RT SD: standard deviation for the average RT.
- Zscore SD: standard deviation for the average *z*-score.
- Accuracy SD: standard deviation for the average accuracy.

Trial-level data

At the trial level, there are 2,240,940 rows of data. These are the raw data allowing everyone to run analyses as if they had collected the data themselves. For each trial, the following information is given.

- Environment: indicates which of the five computers the participant was using when the trial was recorded.
- Participant: identification number of the participant.
- Block: the number of the block in which the trial was presented.
- Order: the presentation order of the trial for the participant.
- Trial: the trial identification number.
- Spelling: the spelling of the stimulus.
- Lexicality: whether the stimulus was a word (W) or a nonword (N).
- Response: the response to the stimulus. Word (W), nonword (N), or time-out(T).
- Accuracy: 1 if the response matched the lexicality; otherwise, 0.
- Previous accuracy: accuracy on the previous trial.
- RT: RT on the trial, with outliers and incorrect responses set to NA.
- RT raw: RT on the trial without cleaning.
- Previous RT: RT on the previous trial.
- Microsec error: the timing error given by the tscope software (in microseconds).
- Unix seconds: date and time in Unix seconds format (seconds elapsed since 1970).
- Unix microseconds: decimal part of unix seconds (in microseconds).
- Trial day: indicates how many trials the participant responded to since the day began (including the current trial).
- Trial session: indicates how many trials the participant responded to since the session began (including the current trial). A session expired after no response was given for 10 min.
- Order in block: the presentation order of the trial in a block of 500 items.
- Order in subblock: the presentation order of the trial in a subblock of 100 items.

Stimulus characteristics

- Coltheart *N*: the number of words of same length differing in one letter, computed over all word forms in the English CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995).
- OLD20: The average orthographic Levenshtein distance of the 20 most similar words, computed over all word forms in the English CELEX lexical database.

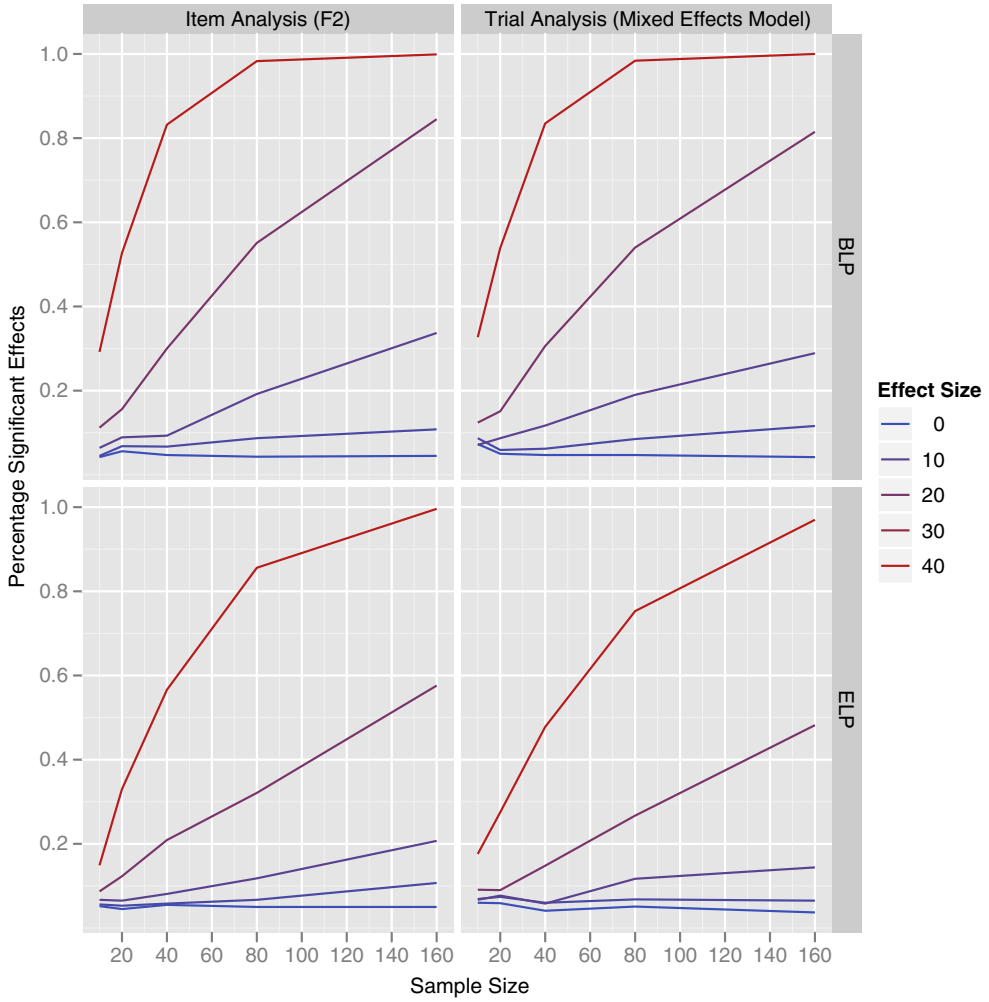

Fig. 6 Sample size required for finding an effect of a particular size (in milliseconds), derived from Monte Carlo simulation. For each combination of sample size (n = 10, 20, 40, 80, 160) and effect size (0, 5, 10, 20, 40 ms), we ran 1,000 simulations, each time taking two random samples of n words from the database. The y-axis indicates the proportion of simulations in which the null hypothesis (no effect) was rejected (alpha = .05). Sample sizes at which sufficient power (.8) is reached for the British Lexicon Project are about n = 40 for an effect of 40 ms and about n = 160 for an effect of 20 ms in both types of analyses. For the English Lexicon Project, sufficient power is reached at about n = 70 for an effect of 40 ms in the item analysis and about n = 100 for an effect of 40 ms in the trial analysis

- CELEX frequency: Raw frequency of the stimulus as given by CELEX.
- CELEX CD: Contextual diversity (dispersion) of the stimulus in CELEX.
- CELEX frequency lemma: sum of the raw frequencies of all possible lemmas for the stimulus in CELEX.
- SUBTLEX frequency: raw frequency of the stimulus in the SUBTLEX-US database.
- SUBTLEX CD: contextual diversity of the stimulus in SUBTLEX-US.
- SUBTLEX frequency million: frequency per million of the stimulus in SUBTLEX-US.
- SUBTLEX CD pct: contextual diversity as a percentage of contexts of the stimulus in SUBTLEX-US.
- BNC frequency: raw frequency of the stimulus in BNC.
- BNC frequency million: frequency per million of the stimulus in BNC.
- Summed monogram: sum of nonpositional letter frequencies, computed over all word forms in CELEX.
- Summed bigram: sum of nonpositional bigram frequencies.
- Summed trigram: sum of nonpositional trigram frequencies.
- Nletters: length of the stimulus in characters.
- Nsyl: length of the stimulus in syllables.
- Morphology: morphological status (e.g., monomorphemic, complex) of the form in CELEX. Different options are separated by a dot.
- Flection: flection (e.g., singular, plural) of the form in CELEX. Different options are separated by a dot.
- Synclass: syntactic class (e.g., verb, noun) of the form in CELEX. Different options are separated by a dot.

Acknowledgements This research was made possible by an Odysseus grant awarded by the Government of Flanders (Belgium) to M.B. The authors thank Kevin Diependaele for making Fig. 1 and running the split-half reliability analysis and Pierre Courrieu for providing us with the ICC reliability analysis.

References

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Baayen, R. H., & Milin, P. (in press). Analyzing reaction times. *International Journal of Psychological Research*.

Baayen, R. H., Piepenbrock, R., & Gulikers, H. (1995). *The CELEX Lexical Database [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium.

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 63–85.

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

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition, 28,* 143–153.

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

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

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology, 60,* 1072–1082.

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods, 40,* 791–794.

Courrieu, P., Brand-D'Abrescia, M., Peereman, R., Spieler, D., & Rey, A. (2011). Validated intraclass correlation statistics to test item performance models. *Behavior Research Methods, 43,* 37–55. doi:10.3758/s13428-010-0020-5.

Courrieu, P., & Rey, A. (2011). Missing data imputation and corrected statistics for large-scale behavioral databases. *Behavior Research Methods, 43,* 310–330. doi:10.3758/s13428-011-0071-2.

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

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

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

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1331–1356.

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

Keuleers, E., Brysbaert, M., & New, B. (2010a). 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. (2010b). 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 Language Sciences*. *Psychology, 1,* 174. doi:10.3389/fpsyg.2010.00174.

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

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

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

Pexman, P. M., Hino, Y., & Lupker, S. J. (2004). Semantic ambiguity and the process of generating meaning from print. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1252–1270.

Rey, A., & Courrieu, P. (2010). Accounting for item variance in large-scale databases. *Frontiers in Psychology, 1,* 200. doi:10.3389/fpsyg.2010.00200.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2000). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46,* 245–266.

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

Sears, C. R., Siakaluk, P. D., Chow, V. C., & Buchanan, L. (2008). Is there an effect of print exposure on the word frequency effect and the neighborhood size effect? *Journal of Psycholinguistic Research, 37,* 269–291.

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

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

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition. *Journal of Verbal Learning and Verbal Behavior, 23,* 383–404.

Sibley, D. E., Kello, C. T., & Seidenberg, M. S. (2009). Error, error everywhere: A look at megastudies of word reading. In N.

Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1036–1041). Amsterdam: Cognitive Science Society.

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

Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging, 15,* 225–231.

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

Stone, G. O., Vanhoy, M., & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory and Language, 36,* 337–359.

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

Wagenmakers, E. J., Ratcliff, R., Gomez, P. K., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58,* 140–159.

Yap, M. J., Balota, D. A., Tse, C. S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 495–513.

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

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

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