Behavior Research Methods
2009, 41 (4), 977-990
doi:10.3758/BRM.41.4.977

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

Marc Brysbaert
Ghent University, Ghent, Belgium
and Royal Holloway, University of London, London, England

and

Boris New
Université Paris Descartes, Paris, France

Word frequency is the most important variable in research on word processing and memory. Yet, the main criterion for selecting word frequency norms has been the availability of the measure, rather than its quality. As a result, much research is still based on the old Kučera and Francis frequency norms. By using the lexical decision times of recently published megastudies, we show how bad this measure is and what must be done to improve it. In particular, we investigated the size of the corpus, the language register on which the corpus is based, and the definition of the frequency measure. We observed that corpus size is of practical importance for small sizes (depending on the frequency of the word), but not for sizes above 16–30 million words. As for the language register, we found that frequencies based on television and film subtitles are better than frequencies based on written sources, certainly for the monosyllabic and bisyllabic words used in psycholinguistic research. Finally, we found that lemma frequencies are not superior to word form frequencies in English and that a measure of contextual diversity is better than a measure based on raw frequency of occurrence. Part of the superiority of the latter is due to the words that are frequently used as names. Assembling a new frequency norm on the basis of these considerations turned out to predict word processing times much better than did the existing norms (including Kučera & Francis and Celex). The new SUBTL frequency norms from the $SUBTLEX_{US}$ corpus are freely available for research purposes from http://brm.psychonomic-journals.org/content/supplemental, as well as from the University of Ghent and Lexique Web sites.

Since the seminal work of Howes and Solomon (1951), it has been well established that word frequency is a very important variable in cognitive processing. High-frequency words are perceived and produced more quickly and more efficiently than low-frequency words (e.g., Balota & Chumbley, 1984; Jescheniak & Levelt, 1994; Monsell, Doyle, & Haggard, 1989; Rayner & Duffy, 1986). At the same time, high-frequency words are easier to recall but more difficult to recognize in episodic memory tasks (e.g., Glanzer & Bowles, 1976; Yonelinas, 2002).

To investigate the word frequency effect, psychologists need estimates of how often words occur in a language. Howes and Solomon (1951), for instance, made use of Thorndike and Lorge's (1944; hereafter, TL) list of words as counted in books. Subsequently, Kučera and Francis's (1967; hereafter, KF) frequency norms became the measure of preference and formed the basis of over 40 years of psycholinguistic and memory research in the U.S. The latter may be surprising, because the KF list was based on a corpus of 1.014 million words only, whereas TL was based on a corpus of 18 million words. The reasons why KF became more popular may have been that the texts were more recent (from 1961 vs. the 1920s and 1930s) and were entirely based on adult reading material, whereas TL also contained children's books. Differences in availability may have played a role as well, in addition to a snowball effect (once KF was used in a number of key articles, it became the measure of choice for the group of researchers working on that topic).

The Continuing Popularity of the Kučera and Francis Norms

The central role of the KF frequencies in current psychological research can be gauged by counting the number of articles citing the 1967 database. In 2007, this was 183; in 2008, it was 215 (retrieved on January 31, 2009, from

M. Brysbaert, marc.brysbaert@ugent.be

Table 1
Frequency Norms Used in Research on Memory and Language Processing in the November 2008 Issue of the *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Source	Topic	Frequency Norms
Huber, Clark, Curran, & Winkielman (2008)	recognition memory	KF
Szpunar, McDermott, & Roediger (2008)	memory for word lists	KF
O'Malley & Besner (2008)	reading aloud	HAL
Hockley (2008)	recognition memory	KF
McDonough & Gallo (2008)	autobiographical memory	KF
McKay, Davis, Savage, & Castles (2008)	reading aloud	KF
Klepousniotou, Titone, & Romero (2008)	understanding ambiguous words	KF
Drieghe, Pollatsek, Staub, & Rayner (2008)	eye movements in reading	KF

Note—KF, Kučera and Francis (1967).

http://apps.isiknowledge.com). The prevalence of KF can further be seen in individual journal issues devoted to language and memory research. Table 1, for instance, lists the frequency measures used in the articles of the November 2008 issue of the *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Apart from one article, they all made use of KF.

The continued use of the KF norms is surprising, given that Burgess and Livesay, in 1998, already raised problems with them and that, currently, frequency norms are available on the basis of much larger corpora. Burgess and Livesay selected two samples of 240 words of varying frequencies and asked participants to name them. In particular, for words with low and medium frequencies, the correlation between naming latencies and KF frequencies was low. Furthermore, Burgess and Livesay showed that the correlations were significantly higher for a new frequency measure calculated on the basis of a corpus of approximately 131 million words gathered from Usenet groups on the Internet. Burgess and Livesay attributed the meager performance of the KF norms to the small size of the corpus (1 million vs. 131 million). They called their new corpus the HAL corpus (from Hyperspace Analogue to Language). Unfortunately, they did not make the new frequency estimates available.

The validation of frequency norms was continued by Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004). They collected naming latencies and lexical decision latencies of young and healthy older adults for over 2,400 monosyllabic English words and correlated them with different frequency norms. They tested five frequency measures: KF, based on written texts (1.014 million words; Kučera & Francis, 1967); Celex, based on American and British written texts (16.6 million words) and transcripts of spoken interactions (1.3 million words; Baayen, Piepenbrock, & van Rijn, 1993); HAL, based on Internet news groups (130+ million words,[1] obtained from Burgess & Livesay, 1998); Zeno, taken from various American textbooks geared toward primary and secondary school children (Grades 1–12; 17 million words; Zeno, Ivens, Millard, & Duvvuri, 1995); and MetaMetrics, based on computer text files (350 million words; see Balota et al., 2004).

Across the various analyses (naming, lexical decision, young adults, old adults), the KF frequency norm consistently underperformed in predicting the reaction times (RTs). The best performance was for Zeno and MetaMetrics; HAL and Celex followed closely behind.

Zevin and Seidenberg (2002) also compared KF and Celex with Zeno. They made use of the data from three different word recognition megastudies (including one by Seidenberg & Waters, 1989). In each case, they observed that KF was the worst one, followed by Celex, and then by Zeno. Zevin and Seidenberg, in particular, warned researchers that when a bad frequency measure is used, stimuli "matched" on frequency are unlikely to be completely confound free if the variable of interest is correlated to word frequency. In such cases, researchers may easily report an "effect" of their variable of interest that, in reality, is a frequency effect in disguise.

Given the findings of Burgess and Livesay (1998), Balota et al. (2004), and Zevin and Seidenberg (2002), it is surprising to see how little this has affected the choice of frequency norms in research. One reason may be the limited availability of the alternatives to KF. The HAL norms have been released only recently by Balota et al. (2007) as part of their Elexicon project.[2] Celex, Zeno, and MetaMetrics are copyright protected, meaning that researchers have no free access to them. Another reason, however, may be that authors do not realize how bad the KF norms are and which alternatives are indicated most.

In the sections below, we will present a new frequency measure and examine how well it does relative to other norms. In this enterprise, we were helped enormously by the recently released Elexicon Project (Balota et al., 2007; http://elexicon.wustl.edu). This project consists of the collection of word processing times for over 40,000 English words (lexical decision and word naming) and provides us with a criterion against which to validate the various frequency measures.

The Potential of Film and Television Subtitles As an Estimate of Everyday Language Exposure

An important factor for the quality of frequency counts is the sources from which the corpus is made (i.e., the language register the corpus taps into). The main sources for a long time were books, newspapers, and magazines. In particular, for research on visual word recognition, these were thought to be the most important sources of input. Problems with these forms of input, however, are that they usually have been edited (to polish the language), that they tend to exaggerate lexical variation (in order not to repeat

the same word over and over again), and that they often deal with topics that are not at the forefront of people's lives. Therefore, researchers have seized the advent of the Internet to search for more spontaneous language. As was indicated above, Burgess and colleagues saw user groups as an interesting new supply. In these groups, Internet users participate in discussions on a variety of topics without much supervision or editing. The experiences with this source have been largely positive, because the corpus is easy to collect and to analyze (given that the texts already exist in digital form) and because the correlation with behavioral data is good (Balota et al., 2004; see also below). As a result, Balota et al. (2007) included the HAL frequencies in their Elexicon project (together with the KF frequencies).[3] Other researchers have proposed the outcome of Internet search engines as another interesting estimate of word frequency (Blair, Urland, & Ma, 2002; New, Pallier, Brysbaert, & Ferrand, 2004).

In order to better approximate everyday language exposure, New, Brysbaert, Veronis, and Pallier (2007) explored film and television subtitles as an alternative source of language use.[4] They did so because subtitles are easy to gather and usually involve people in social interactions. In addition, most participants in psychology studies watch television more than they read books, magazines, or newspapers. New et al. (2007) were able to compile a corpus of nearly 50 million French words coming from 9,474 different films and television series, including French films, English and American films and television series, and non-English films from Europe. To their own surprise, they found that the frequency measures derived from this corpus significantly outperformed those previously derived from books and Internet searches (on the basis of two lexical decision experiments involving 234 and 240 words, respectively). As a result, New and colleagues added the subtitle frequencies to the third version of their Lexique project (see www.lexique.org).

In order to further explore the use of subtitles for frequency norms, we decided to assemble a subtitle corpus for American English as well, which we call the $\text{SUBTLEX}_{\text{US}}$ corpus. We first will describe how we decided on the size of the corpus. Then we will describe how the corpus was made. As will become clear, this is a procedure that can easily be copied in other languages.

How Large Must a Corpus Be?

As was indicated above, Burgess and Livesay (1998) criticized the KF norms for the small corpus on which they were based. Particularly with respect to rare words, this provides unreliable estimates. For many decades, there was no alternative. However, nowadays, as a result of the widespread availability of texts in digital format, it is quite easy to compile a corpus of a few hundred million words. Burgess and Livesay made a corpus of approximately 131 million words coming from Internet newsgroups. Westbury tapped the same source for several years and, by November 2008, reported a corpus of over 16 billion words (Shaoul & Westbury, 2008). For the same reason, the corpora in other countries regularly exceed 10–20 million words (e.g., the British National Corpus [BNC] contains 88 million words from written sources and 12 million words from spoken sources; the Celex norms for the Dutch language are based on 42 million words; Lexique 3 for the French language is based on 15 million words from written sources and over 50 million words from film subtitles).

An interesting question with respect to the corpus size is how large a cost-effective corpus should be. When the corpora were based on counting words in books and rarely exceeded 1 million words, there was no question that the bigger the corpus, the better the frequency counts (Burgess & Livesay, 1998). However, now that we are entering the stage of comparisons between 17 million (Zeno et al.), 130+ million (HAL), and 16+ billion (Shaoul & Westbury) words, we must be entering the zone of diminishing returns. The main question, then, is where this zone begins. To find out, we used the 88 million written part of the BNC (Leech, Rayson, & Wilson, 2001) and calculated word frequencies on various sections of the corpus (500,000 words, 1 million words, . . . , the complete corpus). Then we correlated them with the lexical decision times from Elexicon. We will report only data on the lexical decision times, because the effect of word frequency is particularly strong in this task (Balota et al., 2004) and the results did not differ as a function of the task (see Table 6).

All regression analyses reported in this article included four predictors: log10(frequency+1), $\log^2 10$(frequency+1), number of letters in the word, and number of syllables in the word. $\text{Log}^2 10$(frequency+1) was included because Balota et al. (2004) observed that the relationship between word frequencies and word processing times is not fully captured by the logarithmic curve. In particular, for words with a frequency of more than 100 per million, there seems to be a floor effect, in that these words do not result in increasingly shorter RTs. This floor effect can be captured by using the square of the logarithm of the frequency as an extra predictor in the regression analysis (a polynomial of degree 2 is able to capture concave functions).

We included the number of letters in the word and the number of syllables as additional variables, because word length is a very important variable in the lexical decision times of the Elexicon project, explaining more than 30% of the variance (New, Ferrand, Pallier, & Brysbaert, 2006, Figure 1). Word length has also been found to be an important predictor in event-related potential data (Hauk & Pulvermüller, 2004), brain imaging data (Yarkoni, Speer, Balota, McAvoy, & Zacks, 2008), and eye movements in reading (Brysbaert, Drieghe, & Vitu, 2005; Rayner, 1998). The linear length effect of the number of letters in reality is a compound of word length itself (New et al., 2006, Figure 2) and the number of words resembling the stimulus word (as measured by *N*, the number of orthographic neighbors). We do not make this distinction here, since that would involve extra nonlinear variables. The influence of other variables, although theoretically important for understanding the mechanisms of word recognition,

Table 2
Percentage of Variance Accounted for in the Elexicon Lexical Decision Times by Various Portions of the British National Corpus (N = 31,201)

Size (Million Words)	R^2 (%)
0.5	48.7
1	51.3
2	53.3
4	55.1
8	55.9
16	56.4
32	56.1
88	56.1

will not be addressed in this article either, since, together, they account for, at most, 10% additional variance (see, e.g., Baayen, Feldman, & Schreuder, 2006; Cortese & Khanna, 2007) and require different, more in-depth analyses, which would distract us from the core issue.

Since the Elexicon includes all types of words, we made a selection similar to the one used by New et al. (2006). In particular, we excluded abbreviations, names and adjectives that, according to Elexicon, start with a capital (e.g., *American*), and words that had a lexical decision task (LDT) accuracy lower than 67% (i.e., words that were rejected by more than a third of the participants). This resulted in a total of 31,201 usable word stimuli. Table 2 shows the results of the analyses for the various corpus sizes (percentages of variance accounted for by the regression analysis). From this table, we can see that the gain to be made levels off at a corpus size of 16 million words. This is rather small.

To test Burgess and Livesay's (1998) hypothesis that the optimal corpus size depends on the frequency of the words one is interested in, we made a distinction between high-frequency words (frequency $>$ 20 per million; N = 3,754) and low-frequency words (frequency $<$ 10 per million; N = 27,572). Table 3 shows the results.

From Table 3, we can conclude that the optimal corpus size indeed depends on the frequency of the words one is interested in: Whereas frequency counts for high-frequency words reach a stable level at a corpus size of 1 million, low-frequency words seem to require a corpus size of at least 16 million words for reliable estimates. The lower the word frequency, the larger the corpus must be. At the same time, little gain seems to be made beyond 30 million words, which is in agreement with Balota et al.'s (2004) observation that the Zeno frequencies (based on 17 million words) were not inferior to the HAL frequencies (based on 130+ million words).[5] Similarly, we observed that the percentages of variance explained by the Westbury corpus when it had 7.8 billion words were slightly lower than the percentages explained by the HAL corpus of 130+ million words.

Table 3
Percentage of Variance Accounted for in High-Frequency (HF) and Low-Frequency (LF) Words of the Elexicon Lexical Decision Times by Various Portions of the British National Corpus

Size (Million Words)	HF R^2 (%)	LF R^2 (%)
0.5	50.3	38.2
1	51.3	40.8
2	51.1	43.1
4	51.3	45.4
8	51.6	46.7
16	51.3	47.6
32	51.1	47.7
88	51.2	48.0

Note—N = 3,754 and 27,572 for high- and low-frequency words, respectively. HF words had a frequency of $>$20 per million; LF words had a frequency of $<$10 per million.

The basic message from our analyses, therefore, is that for most practical purposes, a corpus of 16–30 million words suffices for reliable word frequency norms. In particular, there is no evidence that a corpus of 3 billion words is much better than a corpus of 30 million words. For these sizes, it becomes more important to know where the words of the corpus came from.

Assembling the SUBTLEX$_{US}$ Corpus

Subtitles were downloaded from the Web site www.opensubtitles.org. This Web site allows users to select films and television series on the basis of various criteria, such as the year of production, the language of the movie, and the country of origin.

We started by downloading files from different sources:[6] U.S. films from 1900–1990 (2,046 files); U.S. films from 1990–2007 (3,218 files); and U.S. television series (4,575 files).

Once the files were downloaded, they needed to be cleaned for optical character recognition (OCR) errors, because most subtitle files were obtained by scanning them from DVDs with an OCR system. We rejected all files with more than 2.5% type errors according to the spelling checker Aspell. In the end, 8,388 films and television episodes were retained, with a total of 51.0 million words (16.1 million from television series, 14.3 million from films before 1990, and 20.6 million from films after 1990). For some programs, the subtitles were limited to a short fragment (the shortest file contained only 84 words). We included them, since we saw no good reason to use the length of the fragment as a selection criterion.

How Well Do the Different Frequency Norms Predict Lexical Decision Times?

As was indicated above, the best way to validate frequency counts is to see how well they predict human processing latencies. Burgess and Livesay (1998) and New et al. (2007) did this for samples of 240 words; Zevin and Seidenberg (2002) and Balota et al. (2004) did it for 2,400+ monosyllabic words. However, the collection of LDTs for over 40,000 words in the Elexicon project makes it possible to validate the frequency norms on a sample that encompasses most of the generally known words in English. For each word, we used six frequencies: KF (based on 1 million words; Kučera & Francis, 1967); Celex (based on 17.9 million words; Baayen et al., 1993); HAL (based on 130+ million words; Burgess & Livesay,

Table 4
Percentages of Variance Explained in the Elexicon Accuracy (Acc) Rates and Lexical Decision Times by the Different Frequency Norms (Polynomials of Degree 2)

Measure	$Acc_{all\ words}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{all\ words}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
KF	19.6	28.6	19.1	57.7	38.9	32.8
Celex	25.2	36.1	25.8	60.6	41.1	37.6
HAL	31.1	38.2	32.1	63.4	44.5	43.9
Zeno	31.6	41.9	31.6	62.9	43.9	42.2
BNC	25.6	36.3	25.0	60.3	41.2	37.5
SUBTL	30.1	40.7	33.6	62.3	45.2	43.5
HAL + SUBTL + Zeno	33.7	45.5	35.8	64.1	47.7	45.9

Note—Other variables included in the regression were word length in number of letters and in number of syllables, if applicable; LDTs were calculated on the z scores of the Elexicon project for words with an accuracy >.66. KF, Kučera and Francis (1967); BNC, British National Corpus.

1998); BNC written (based on 88 million words; Leech et al., 2001); Zeno (based on 17 million words; Zeno et al., 1995); and SUBTL (based on 51 million words from American subtitles; the present article).

We opted for BNC, rather than MetaMetrics, because it is freely available on the Internet and it allowed us to look at the importance of differences in spelling and word use between British and American English. If a word did not have an entry in one of the corpora, it was given a frequency of 0. As in the analysis above, abbreviations and words starting with a capital in Elexicon were excluded.

In addition to the full set of words, we also calculated the percentages of variance explained for the monosyllabic and the bisyllabic words. Psycholinguists have a special relationship with these words, because the vast majority of experiments and computational models of visual word recognition thus far have been based on monosyllabic words and are currently being extended to bisyllabic stimuli. Therefore, it is important to know how well the frequency measures are doing for these words.

We based the regression analyses of the RTs on the z scores of the participants. These were calculated by subtracting the mean of each participant from their raw RT and dividing the remainder by the participant's standard deviation. In this way, individual differences in overall speed and variation were partialed out. Regression on z scores resulted in extra variance accounted for, relative to regression on the raw RTs or on the log of the RTs (see Table 7).

Table 4 lists the results. For each corpus, the table shows the percentage of variance explained by log10(freq+1), $\log^2 10$(freq+1), number of letters in the word, and number of syllables in the word.

There are several noteworthy findings in Table 4. First, the bad performance of KF mentioned by Burgess and Livesay (1998), Zevin and Seidenberg (2002), and Balota et al. (2004) is replicated for the entire Elexicon project. This is a sad finding, given that so much research in American English still is based on this frequency norm (Table 1).[7]

Second, Celex and the BNC seem to be less good than the top three. A likely explanation for the performance of the BNC is the influence of differences in orthography and word use between American and British English. If authors want to use an English corpus for American experiments (or vice versa), they should be very careful about differences in spelling and the implications this may have for their stimulus selection. The most likely reason for the poor performance of Celex is the age and the representativeness of the texts on which it is based (this was the COBUILD corpus assembled by linguists before the advent of the Internet). The Celex norms particularly perform suboptimally at predicting whether letter strings will be perceived as existing words or not (i.e., the accuracy rates). Note that, for many words, Celex provides both British and American spellings; where provided, we used the American.

The third important observation is that the three big American corpora (HAL, Zeno, $SUBTLEX_{US}$) overall have quite similar performance but do not yet seem to have depleted the full potential, given that a combination of the three sources still accounts for some 2%–4% of extra variance (last line of Table 4; see Table 11 for one of the reasons). Further scrutiny of the data revealed that HAL is doing particularly well for long words (Table 5). The most likely reason for this superiority is that long words tend to be avoided in subtitles and children's books. $SUBTLEX_{US}$ does particularly well for short words.

To further examine how well the three corpora do with respect to words frequently used in psycholinguistic research, we looked at the percentages of variance accounted for in the studies of Balota et al. (2004; see also Cortese &

Table 5
Proportions of Variance Explained by HAL and SUBTLEX for Words of Different Lengths

Word Length	HAL	SUBTLEX
3	.38	.51
4	.47	.53
5	.47	.49
6	.47	.47
7	.45	.44
8	.44	.42
9	.42	.38
10	.41	.36
11	.40	.35
12	.39	.33
13	.39	.30

Table 6
Percentages of Variance Explained by the Various Frequency Counts in the Lexical Decision Task Accuracy (Acc) Data Reported by Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) and the Elexicon Project (Balota et al., 2007)

Measure	Acc_{young}	Acc_{old}	Acc_{Elex}
KF	18.0	7.0	22.5
Spoken	16.8	5.5	23.0
Celex	24.2	10.4	26.0
HAL	24.7	8.2	31.3
Zeno	25.5	10.7	29.8
BNC	22.8	9.0	25.4
SUBTL	27.7	12.4	38.3

Note—Multiple regression analysis involved log(freq+1), $\log^2$(freq+1), and word length in number of letters. All stimuli were monosyllabic (N = 2,406). KF, Kučera and Francis (1967); BNC, British National Corpus.

Khanna, 2007). These studies involved naming and lexical decision for 2,406 monosyllabic words with frequencies of more than 1 per million. There were separate groups of young and old adults doing these tasks.

Table 6 first gives the results of the accuracy data in the LDT. For comparison purposes, the data from Elexicon (obtained with young adults) on the same words are included as well. In addition to the frequency measures discussed above, Table 6 also includes recently published frequency norms of spoken American English based on the Michigan Spoken corpus (Pastizzo & Carbone, 2007). These norms are based on a corpus including 1.6 million words obtained from 152 transcriptions of lectures, meetings, advisement sessions, public addresses, and other educational conversations recorded at the University of Michigan.

Table 6 further illustrates the inferior performance of the KF measure. It also illustrates the rather low fit of Celex for data obtained with students (it does better with older participants). Because of its small corpus size, the spoken frequency measures also perform less well. Of the three remaining contenders, there is a clear advantage for the SUBTL frequencies, confirming New et al.'s (2007) initial observations in French (note, however, that the amount of data here is much larger than the small samples used by New et al., 2007).

Table 7 presents the same information for the RT data. For illustration purposes, we here include the data for both the raw lexical decision times of the Elexicon project and the z scores. As can be seen, the percentage of variance accounted for is substantially higher for the z scores than for the LDTs, because individual differences in RTs have been partialed out. As in Table 6, there is a clear gradation in performance. The two small corpora (KF and Spoken) perform significantly worse; SUBTLEX does consistently better. As is shown in Table 4, Celex does better for RT data of known words than for accuracy data. The differences between the corpora are smaller for naming times than for lexical decision times, in line with the observation that the impact of word frequency is much smaller in word naming than in the LDT (Balota et al., 2004; Cortese & Khanna, 2007). However, the pattern of results remains the same.

Are Lemma Frequencies Better Than Word Form Frequencies?

In the analyses thus far, we used the simple word form (WF) frequencies. These are the frequencies of the words as they appear in the corpus. For instance, there are 18,081 occurrences of the word *play* in $SUBTLEX_{US}$, 1,521 of the word *plays*, 2,870 of the word *played*, and 7,515 of the word *playing*. However, these words all refer to the same base word, *play*, which can be a noun or a verb. Is it better to combine these frequencies or to leave them as they are?

The "American" solution has been to work with the WF frequencies as observed in the corpus (also called *surface frequency*). This is true for HAL, Zeno, and KF (although Francis & Kučera, in 1982, published a list with lemma frequencies). In Europe, more importance has been attached to lemma frequency. The lemma frequency is the sum of the frequencies of all the inflected forms of a particular noun, verb, or adjective. For instance, the lemma frequency of the noun *dog* is the sum of the frequencies of its inflected forms *dog* and *dogs*. The lemma frequency of the verb *to beg* is the sum of the frequencies of its inflected forms *beg*, *begs*, *begged*, and *begging*.

The idea behind the use of lemma frequencies is that processing times of inflected forms profit from each other, so that the total number of encounters with *to play* is the summed frequency of $play_{verb}$, $plays_{verb}$, $played_{verb}$, and $playing_{verb}$. Such a view is particularly appealing within theories that postulate a process of morphological

Table 7
Percentages of Variance Explained in the Reaction Time Data Reported by Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) and Balota et al. (2007)

	R^2 (%)							
Measure	LDT_{young}	LDT_{old}	LDT_{Elex}	$z_{LDTElex}$	NMG_{young}	NMG_{old}	NMG_{Elex}	$z_{NMGElex}$
KF	31.8	23.8	32.1	38.0	20.0	21.5	22.7	23.9
Spoken	31.1	19.9	31.9	38.5	19.7	20.6	23.0	24.4
Celex	37.0	28.4	33.8	40.8	20.0	21.3	22.3	23.9
HAL	36.7	24.2	37.7	45.6	20.5	21.9	23.9	25.3
Zeno	38.8	30.1	35.6	43.3	20.5	22.2	23.5	24.7
BNC	34.8	26.8	34.4	41.2	19.7	21.5	22.8	24.0
SUBTL	42.2	29.3	40.1	48.6	21.0	22.9	24.1	25.2

Note—Multiple regression analysis involving log(freq + 1), $\log^2$(freq + 1), and word length in number of letters. All stimuli were monosyllabic (N = 2,406). LDT, lexical decision task; NMG, word naming; KF, Kučera and Francis (1967); BNC, British National Corpus.

decomposition for the recognition of inflected words (e.g., Clahsen, 1999; Rastle, Davis, & New, 2004; Taft, 2004). According to these theories, the word *played* is recognized by decomposing it into its morphemes *play* + *-ed*. Lemma frequencies are much less important in theories that postulate separate lexical entries for all morphologically complex words, except for the ones with a very low frequency (e.g., Caramazza, Laudanna, & Romani, 1988). In between are the many dual-route theories that see the processing of morphologically complex words as the result of an interaction between decomposition and whole-word lookup (e.g., Baayen, Dijkstra, & Schreuder, 1997; New, Brysbaert, Segui, Ferrand, & Rastle, 2004).

The first corpus to really promote lemma frequencies was Celex (Baayen et al., 1993). To calculate these frequencies, the corpus had to be parsed so that words were tagged for their syntactic role within the text. This was done semiautomatically with a manual cross-check on selected samples. Nowadays, automatic taggers provide a more acceptable outcome than do the semiautomatic parsers of the late 1980s in a fraction of the time. As a result, the BNC project also gives lemma frequencies in addition to surface frequencies (Leech et al., 2001).

To find out whether lemma frequencies explain lexical decision times better than do WF frequencies (and hence, whether we could further improve the SUBTL norms by tagging the corpus), we compared both types of frequencies for the Celex and the BNC. A problem we rapidly encountered, however, was that lemma frequency can be defined in several ways, depending on the theory of word recognition one adheres to. The lemma frequency can be the sum of the lemma frequencies of the different syntactic roles (e.g., *play* as a verb and as a noun, in line with what is done for the WF frequencies). However, the lemma frequency could also be the highest lemma frequency (e.g., that of the verb in the case of *play*). Furthermore, the lemma frequency can be given to all possible inflections of the base word (*play*, *plays*, *played*, *playing*) or to the base word only (*play*). Finally, there is the question of what to do with irregular forms (*men*, *ate*, *worse*): Should they be included in the lemma frequency or kept separately?

To make sure that we gave lemma frequencies each and every opportunity, we ran tens of analyses with different definitions.[8] The overall finding, however, was always the same: Whereas it is possible under some circumstances to account for 1%–2% more extra variance with lemma frequencies than with WF frequencies when the analysis is limited to $\log(\text{freq}+1)$ and $\log^2(\text{freq}+1)$, as soon as word length (number of letters and number of syllables) is added, most of the advantage is lost, because lemma frequencies are correlated more with word length than with WF frequencies (this is particularly the case when only the uninflected forms are given the lemma frequency; e.g., *play*, but not *plays*, *played*, or *playing*).

Table 8 shows some illustrative examples for Celex. As in the previous analyses, they are based on the Elexicon and the Balota et al. (2004) data. Lemma frequency was defined as the summed lemma frequencies of all the syntactic roles a word can have (e.g., *play* both as a verb and as a noun) and used for all the inflections (i.e., *play*, *plays*, and *played* got the same lemma frequency).[9] The correlation between log WF frequency and log lemma frequency thus defined was .92 for the words of Elexicon project and .95 for the words of Balota et al. (2004). The data are shown for regressions that included frequency only and regressions that included both frequency and word length.

As can be seen in Table 8, the use of lemma frequencies does not seem to result in a marked improvement of the fit for American English, certainly not when word length is included in the regression analysis. This is surprising,

Table 8
A Comparison of the Variance Explained by CELEX Word Form (WF) Frequencies and Lemma Frequencies for Performance in the Experiments Reported in the Elexicon Project (*z* Scores) and Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004), When Word Length (Number of Letters and Number of Syllables If Applicable) Was Included and When It Was Not Included

	Elexicon					
	$Acc_{\text{all words}}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{\text{all words}}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
Frequency						
Celex WF	21.3	33.9	21.4	36.2	39.4	34.6
Celex lemma	21.9	36.6	21.3	37.9	37.1	32.4
Frequency + word length						
Celex WF	25.2	36.1	25.8	60.7	41.1	37.6
Celex lemma	25.8	37.9	25.4	60.2	40.0	35.9
	Balota et al. (N = 2,406)					
	Acc_{young}	Acc_{old}	LDT_{young}	LDT_{old}	NMG_{young}	NMG_{old}
Frequency						
Celex WF	23.9	10.3	36.9	27.9	6.4	9.8
Celex lemma	25.3	10.1	36.5	27.3	6.2	9.2
Frequency + word length						
Celex WF	24.2	10.4	37.0	28.4	20.0	21.3
Celex lemma	25.5	10.3	36.7	28.0	20.0	21.2

Note—Acc, accuracy; RT, reaction time; LDT, lexical decision task; NMG, word naming.

Table 9
Percentages of Variance Accounted for by the Word Frequency SUBTL Index and the Contextual Diversity SUBTL Index for the Elexicon Project and the Monosyllabic Words Investigated by Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004; see Tables 5 and 6)

	Elexicon					
	$Acc_{all\ words}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{all\ words}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
Frequency						
$SUBTL_{WF}$	22.0	32.9	26.4	49.2	45.2	42.5
$SUBTL_{CD}$	23.4	36.8	28.0	49.5	46.8	43.6
Frequency + word length						
$SUBTL_{WF}$	30.1	40.7	33.6	62.3	45.2	43.5
$SUBTL_{CD}$	31.3	44.0	34.9	62.9	46.8	44.6
	Balota et al. (N = 2,406)					
	Acc_{young}	Acc_{old}	LDT_{young}	LDT_{old}	NMG_{young}	NMG_{old}
Frequency						
$SUBTL_{WF}$	26.4	12.1	42.1	29.5	9.7	13.6
$SUBTL_{CD}$	29.3	13.9	44.2	31.0	9.4	13.3
Frequency + word length						
$SUBTL_{WF}$	27.7	12.5	42.3	29.6	21.1	22.8
$SUBTL_{CD}$	30.6	14.3	44.3	31.1	21.2	23.0

Note—Acc, accuracy; RT, reaction time; LDT, lexical decision task; NMG, word naming; WF, word form frequency; CD, contextual diversity.

because there is empirical evidence that the frequencies of regular plural nouns affect the lexical decision times to the singular form: Participants are faster to accept singular nouns with high-frequency plurals (e.g., *acre*, *boot*, *critic*) than matched nouns with low-frequency plurals (e.g., *aunt*, *earl*, *flint*; Baayen et al., 1997; New, Brysbaert, et al., 2004). However, our analyses with the entire Elexicon suggest that, for most practical purposes, lemma frequencies in English are not more informative than WF frequencies. This also seems to be the conclusion reached by Baayen in his most recent articles (e.g., Baayen et al., 2006; Baayen, Wurm, & Aycock, 2007). In these articles, Baayen makes use of the Celex WF frequencies and adds extra variables to capture the morphological richness of a word, such as the word family size or information-theoretical measures of morphological complexity.

Is Contextual Diversity Better Than WF Frequency?

Another variable that has been proposed as an alternative to WF frequency is the contextual diversity (CD) of a word (Adelman, Brown, & Quesada, 2006). This variable refers to the number of passages (documents) in a corpus containing the word. So, rather than calculating how often a word appeared in the BNC, Adelman et al. measured how many of the 3,144 text samples in the corpus contained the

Table 10
Percentages of Variance Accounted for When the Word Form (WF) Frequency and Contextual Diversity (CD) Measures Are Based on All the Occurrences of the Words or Only on the Occurrences of the Words Starting With a Lowercase Letter, Separately for the Elexicon Project and the Monosyllabic Words Investigated by Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004; see Tables 5 and 6)

	Elexicon					
	$Acc_{all\ words}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{all\ words}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
Frequency + word length						
$SUBTL_{WF}$	30.1	40.7	33.6	62.3	45.2	43.5
$SUBTL_{CD}$	31.3	44.0	34.9	62.9	46.8	44.6
$Frequency_{lowercase}$ + word length						
$SUBTL_{WF}$	31.1	44.2	34.5	62.7	47.5	44.0
$SUBTL_{CD}$	31.8	46.1	35.2	63.0	47.8	44.4
	Balota et al. (N = 2,406)					
	Acc_{young}	Acc_{old}	LDT_{young}	LDT_{old}	NMG_{young}	NMG_{old}
Frequency + word length						
$SUBTL_{WF}$	27.7	12.5	42.3	29.6	21.1	22.8
$SUBTL_{CD}$	30.6	14.3	44.3	31.1	21.2	23.0
$Frequency_{lowercase}$ + word length						
$SUBTL_{WF}$	31.0	14.3	45.3	32.1	20.9	22.8
$SUBTL_{CD}$	32.0	15.3	45.5	32.1	21.0	22.9

Note—Acc, accuracy; RT, reaction time; LDT, lexical decision task; NMG, word naming.

Table 11
Percentages of Variance Accounted for by the Different Frequency Measures for the Elexicon Project When the Analyses Are Limited to the Words That More Often Start With a Lowercase Letter Than With an Uppercase Letter (RT Analyses Limited to Words With an Accuracy Level >.66)

Measure	Acc_{all} (N = 31,246)	$Acc_{Nsyl=1}$ (N = 5,281)	$Acc_{Nsyl=2}$ (N = 12,439)	RT_{all} (N = 27,350)	$RT_{Nsyl=1}$ (N = 4,721)	$RT_{Nsyl=2}$ (N = 10,840)
HAL	29.5	38.3	30.0	59.1	46.4	42.1
Zeno	30.2	40.6	29.9	58.5	44.4	40.7
$SUBTL_{WF}$	28.6	40.7	31.5	58.1	47.4	42.2
$SUBTL_{WFlow}$	28.9	41.3	31.8	58.3	47.6	42.4
$SUBTL_{CD}$	29.7	43.2	32.7	58.7	47.8	42.9
$SUBTL_{CDlow}$	30.0	43.6	32.8	58.7	47.8	42.9
Hal + Zeno + SUBTL	31.1	40.4	31.8	60.0	47.8	43.6

Note—Acc, accuracy; RT, reaction time; WF, word form frequency; CD, contextual diversity.

word. They found that the CD measure explained 1%–3% more of the variance in the Elexicon data.[10]

To further assess the relative merits of the CD index and the WF index, we counted how many of the 8,388 films in $SUBTLEX_{US}$ contained the various words. Then we entered log10(CD+1) and $\log^2 10(CD+1)$ in the regression analyses. Table 9 lists the results, again when the analysis is limited to word frequency and when word length is included. From this table, it is clear that (possibly with the exception of naming latencies), the CD index outperforms the WF index. There is a difference of 1%–3% in the variance accounted for, which is in line with the figure reported by Adelman et al. (2006). In addition, this extra variance accounted for is not a confound of word length. The difference remains pretty much the same whether or not the numbers of letters and syllables in the words are included.

A closer look at the data suggested one reason why CD is better than WF. Many words with a high WF/CD ratio turned out to be words that were used as names in a few films (such as *drake* and *prince*). To address this issue, we calculated how often each word in $SUBTLEX_{US}$ started with a lowercase letter. To our surprise, it turned out that this lowercase frequency measure was substantially better than the total word frequency. In addition, for this measure, the difference between WF and CD was much attenuated (Table 10).

Further scrutinizing of the results suggested that the better performance of the lowercase frequency measure was partially due to the fact that not all the words in Elexicon have been given the capital they deserve on the basis of their usage in the language (examples are *beltway*, *bock*, *gable*, *sloan*).[11] If we omitted all the words that, in $SUBTLEX_{US}$, occurred more often with a capital than with a lowercase first letter, the advantage of the lowercase frequency largely disappeared, and the difference between CD and WF diminished (Table 11). So, although there is no big gain to be made by using lemma frequencies rather than WF frequencies in English, researchers should avoid using words that frequently occur as names.

In What Respect Do HAL and $SUBTL_{CD}$ Differ?

The fact that $SUBTL_{CD}$ does better than HAL in accounting for the Balota et al. (2004) young adult lexical decision times (31% of the variance accounted for in accuracy vs. 25%; 44% of the variance accounted for in RT vs. 37%) raises the question of what types of words get a different estimate. To find out, we ran a regression on the 2,407 monosyllabic stimulus words in which we predicted $SUBTL_{CD}$ on the basis of HAL and then examined the standardized residuals. Table 12 lists the 25 words for which the $SUBTL_{CD}$ frequencies were most overestimated on the basis of HAL and the 25 words for which they were most underestimated. These words, more than any statistical analysis, illustrate the difference in register tapped by the two corpora. Whereas the HAL corpus focuses on a written Internet world (*pub*, *text*, *mime*, *web*, *sphere*, *mode*), the SUBTLEX corpus is centered on everyday interactions (*swear*, *dad*, *bye*, *sir*, *beg*, *slept*, *sit*, *kid*, *hush*, *shut* [*up*], *cheer* [*up*], *wake* [*up*], *hang* [*on*]).

Table 12
Frequencies of Words Used in Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) Overestimated and Underestimated on the Basis of HAL (Relative to $SUBTLEX_{US}$)

Overestimated on the Basis of HAL	Underestimated on the Basis of HAL
thru	swear
null	dad
lisp	bye
pub	staunch
warp	calm
death	breath
node	cinch
text	sir
stilt	slept
mime	beg
spool	shut
Web	knock
sphere	toast
mint	till
vale	sit
and	kid
width	hush
volt	cheer
prompt	drink
strand	wow
strait	drunk
dole	sweet
ram	booze
hind	wake
mode	hang

Table 13
Frequencies of Words Used in Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) Overestimated and Underestimated on the Basis of Kučera and Francis (KF) (Relative to $SUBTLEX_{US}$)

Overestimated on the Basis of KF	Underestimated on the Basis of KF
chive	hook
whig	sneak
shear	stuff
strode	freak
oust	lab
fig	bet
spire	trash
daunt	fry
and	heck
gaunt	swear
loath	weird
flux	thank
clad	scare
strait	steal
strove	ouch
null	dad
thru	jerk
wry	yeah
scribe	cute
quill	bike
clung	pal
scant	hey
sprawl	ass
sparse	bye
blithe	wow

In What Respect Do KF and $SUBTL_{CD}$ Differ?

The same analysis can be done to examine the differences in register tapped by KF and $SUBTL_{CD}$ (Table 13). This analysis clearly illustrates why KF no longer is a good instrument, because it overestimates the frequency of outdated words (*chive*, *whig*, *shear*, *strode*, *oust*, *fig*, *spire*, *daunt*) and underestimates the frequency of words used in everyday informal social interactions (*steal*, *dad*, *jerk*, *cute*, *bike*, *pal*).

Discussion

The Kučera and Francis (1967) word frequency norms are still widely used in American research, despite the fact that serious concerns have been raised about them (Balota et al., 2004; Burgess & Livesay, 1998; Zevin & Seidenberg, 2002). This illustrates the observation made by Peirce (1877) that the scientific method based on empirical evaluation is only one of the four ways in which people acquire knowledge. The other methods are the method of tenacity (people hold to assumptions and beliefs because they have been around for a long time), the method of authority (people form opinions by consulting experts), and the a priori method (people use information because it looks sound on the basis of their own reasoning, logic, and intuition). Arguably, psychology researchers keep on using the KF norms because the norms have been around for several decades (tenacity) and have been used by experts on the topic (authority).

The analyses presented in this article confirm the bad quality of the KF norms. If we use them to predict how well words are known, we are able to explain but 19% of the variance (including word length in the analysis). This compares with the 31% accounted for by the best frequency measure. The comparison is slightly better when it comes to predicting RTs to known words (58% vs. 63%), probably because most of these words are of a higher frequency. A similar picture emerges when we limit the analysis to words typically used in word processing tasks. If we take the Balota et al. (2004) lexical decision study with young adults as a representative study, we see that the KF norms account for 18% of the variance in the accuracy data and 32% in the latency data. In contrast, the most successful predictor, $SUBTL_{CD}$, accounts for 30% and 44%, respectively, of the variance.

The analyses presented in this article further show that by using the scientific method, it is not difficult to improve on the existing word frequency norms. Three variables are important for evaluating the quality of a frequency measure: the size of the corpus, the register the corpus taps into, and the frequency measure used.

As for the corpus size, it is clear that any corpus smaller than 16 million words lacks the level of detail to reliably estimate word frequencies below 10 per million (see the results of KF and the spoken corpus of Pastizzo & Carbone, 2007). At the same time, gains become marginal above sizes of 16–30 million words. It is not the case that a corpus of 1 billion words will be better than a corpus of 30 million words; it can easily be worse if the language register on which it is based is not equally representative for everyday language use.

The second variable affecting the quality of the frequency norms is the representativeness of the materials on which the norms are based: The more natural the language use is, the better the frequency norms account for lexical decision times. The Zeno frequency counts do as well as HAL for monosyllabic words, despite the fact that the underlying corpus is eight times smaller (17 million words vs. 130 or 160 million words). This suggests that books for children are a more interesting source of language use than Internet discussion groups. Similarly, $SUBTLEX_{US}$ outperforms HAL for short words, despite the fact that the corpus on which it is based is three times smaller. Apparently, the types of words used in films and television series are a better approximation of real-life exposure than are the types of words used in Internet user groups.

Finally, the third factor affecting the quality of word frequency norms is the way in which the norms have been defined. Interestingly, for English, we found no advantage for lemma frequencies, relative to WF frequencies, as was hypothesized by the linguists behind Celex and BNC and as has been implemented, for instance, by Duyck, Desmet, Verbeke, and Brysbaert (2004) in their WordGen program. The few percentages of extra variance that sometimes can be explained by the lemma frequency seem to be obtained at the expense of the word length effect. This is a rather intriguing finding, because it suggests that many inflected forms in English are stored as separate entities in the mental lexicon (i.e., there are separate representations for *play*, *plays*, *played*, and *playing*), in line with full-storage models of morphological processing (Caramazza et al., 1988) and parallel dual-route models (Baayen et al., 1997; New, Brysbaert, et al., 2004). It will be interesting to see

whether WF frequencies keep on being as good as lemma frequencies for languages with a much richer inflectional system. After all, inflections in English are rather limited, and there is the fact that many noun forms are verb forms as well, cutting through the syntactic boundaries that underlie the definition of a lemma.

Although frequency of occurrence intuitively seems to be the most appealing definition of word frequency, Adelman et al. (2006) have suggested that CD may be a better measure, a finding upheld in our analyses. Rather than counting the number of times a word is present in corpus, it is better to count the number of documents that contain the word. Several factors are likely to contribute to the superior performance of word frequency operationalized as CD.

First, WF frequencies can be overestimated because of multiple occurrences in a single source. For instance, the word *creasy* (defined in dictionaries as *full of creases*) occurs 63 times in $SUBTLEX_{US}$, giving it a frequency of 1.24 per million. However, all these occurrences come from a single source (a film about John Creasy). Compare this with the word *measly*, which occurs 63 times as well but is found in 59 different film excerpts. In particular, words that can be used as names are prone to this type of distortion. An additional problem with such words is that names are likely to involve a different type of processing than do content words such as nouns, adjectives, and verbs. How many people make a link with bows and arrows when they are watching a film about the Archer family? Still, the family name gives the word *archer* a respectable frequency of 5.5 per million in $SUBTLEX_{US}$.

Because family names occur in rapid succession (and in the same context), the CD measure is more robust against this type of distortion than is the WF measure, as can be seen in Table 10: The correlation between word processing times and CD is less influenced by words starting with a capital than is the correlation between word processing times and WF. It can be expected that the same will be true for other distortions due to the repeated use of a word in one particular film (e.g., in a movie about a dam).

A second reason why CD is a better measure than WF may be due to the CD itself, as hypothesized by Adelman et al. (2006). In their view, the more contexts a word has occurred in, the more likely it is to be needed in a new context. CD may also play a role as a semantic variable, since it has been observed that the more easily participants can generate a context in which the word appears, the faster they are to recognize the word (van Hell & de Groot, 1998).

Finally, it could be that rapid successions of words do not induce the same neurological changes as do slow successions. A representation that has been activated shortly before may not undergo the same physiological changes upon a new activation (cf. the phenomenon of repetition priming). In addition, there is a long-standing literature showing that distributed practice results in more enduring learning than does massed practice (Underwood, 1961), and there is growing evidence that exposure does not have an immediate effect on word representations in the mental lexicon but requires at least one night's sleep to be implemented (Gaskell & Dumay, 2003). All these factors may mitigate the effect of repeated exposures on a single day.

The superiority of the CD measure also has practical implications, because it indicates that a corpus consisting of a large number of small excerpts is better than a corpus consisting of a small number of large excerpts. It is better to have a 10-million-word corpus coming from 5,000 different sources than one coming from 100 different sources (e.g., books). On the basis of the work by Adelman et al. (2006) and our own, it would seem that the corpus should comprise at least 3,000 different samples, with presumably not much gain to be expected above 10,000 samples. In addition, if the idea of distributed practice is correct, it may not be good to use samples that succeed each other rapidly in time (e.g., articles from the same newspaper). As for the size of the samples, better results can be expected with samples of moderate size (a few hundred words to a few thousand words) than with samples of a large size (e.g., 100,000 words or more), because, in the latter case, many words will appear in all samples, so that the variability in the frequency norms will be compromised.

Availability

Knowing which frequency measure is the best is one thing; having access to it is another. In this respect, American researchers have been at a disadvantage with respect to their colleagues in the U.K., France, and Spain. Most of the indices are not freely available, either because they are subject to copyright restrictions or because they have never been published in their entirety. The latter is the case, for instance, for the HAL norms. They have become available only recently as part of the Elexicon project and, even then, only for the 40,000 words that are included in this project.

We are in the lucky situation that the work presented here was covered by educational, noncommercial grants. Therefore, we can give full access to the SUBTL word frequencies. They are available in different formats. First, there is a raw text file that contains the information for the 282K different letter strings in $SUBTLEX_{US}$. This is the file that will be of interest to people working specifically on word frequency measures. For the others, we have made two more user-friendly files, limited to the words that are likely to be of interest to users of word frequency norms (this was achieved by excluding the entries that did not appear in spelling checkers and, hence, are likely to be typos, infrequent names, or renditions of speech hesitations; we also combined words across letter cases).

The first user file is an Excel file containing the 74K entries of $SUBTLEX_{US}$ that had a spelling match in the spelling checkers. The second file comprises the same information but is limited to the 60.4K entries with a frequency of more than 1 in the corpus. Because this file contains fewer than 65K lines, it can be read by all spreadsheets. These files are available at the Web site of the Psychonomic Society (http://brm.psychonomic-journals.org/context/supplemental.org), at the Web site of the Department of Experimental Psychology at Ghent University (http://expsy.ugent.be/subtlexus), and at the Web site of Lexique (http://subtlexus.lexique.org). In addition, the latter Web site provides the possibility of online searches on the Internet and the opportunity to have a look at the contexts in which the words appear.

The Excel files contain the following information:

1. *The word*. To help users, this is written with a capital whenever the frequency of the lowercase word is smaller than half of the total frequency. This is the case for nearly one word in six. Words with capitals include all the words that are used more often as a name than as a content word (e.g., *Mark*, *Bay*), words with a frequency of 1 or 2 that happened to be the first word of a sentence (e.g., *Workwise*, *Unverifiable*), and words that in general tend to occur at the beginning of a sentence (e.g., *According*, *Thanks*).

2. *FREQcount*. This is the number of times the word appears in the corpus (i.e., on the total of 51 million words).

3. *CDcount*. This is the number of films in which the word appears (i.e., it has a maximum value of 8,388).

4. *FREQlow*. This is the number of times the word appears in the corpus starting with a lowercase letter. This allows users to further match their stimuli.

5. *CDlow*. This is the number of films in which the word appears starting with a lowercase letter.

6. $SUBTL_{WF}$ is the word frequency per million words. It is the measure one would preferably use in one's manuscripts, because it is a standard measure of word frequency independent of the corpus size. It is given with two-digit precision, in order not to lose precision of the frequency counts.

7. *Lg10WF*. This value is based on log10(FREQcount + 1) and has four-digit precision. Calculating the log frequency on the raw frequencies is the most straightforward transformation, because it allows one to give words that are not in the corpus a value of 0. One can easily lose 5% of the variance explained by taking log(frequency per million + 1), because, in this case, one is not making much of a distinction between words with low frequencies. Similarly, adding values lower than 1 (e.g., +1E−10) is dangerous, because one may end up with a big gap between the words in one's corpus and words for which one does not have a frequency measure (which will get a log value of −10). In addition, if one uses log(frequency per million), one will have negative values for words with a frequency lower than 1 per million, and one will have to enter negative values for missing words. The downside of using log10(FREQcount + 1) is that there is no intuitive relationship between Lg10WF and $SUBTL_{WF}$ (frequency per million). As a rule of thumb, the following conversions apply for $SUBTLEX_{US}$:

Lg10WF	$SUBTL_{WF}$
1.00	0.2
2.00	2
3.00	20
4.00	200
5.00	2,000

8. $SUBTL_{CD}$ indicates in what percentage of the films the word appears. For instance, the word *the* has a $SUBTL_{CD}$ of 100.00, because it occurs in each film. In contrast, the word *abbreviation* has a $SUBTL_{CD}$ of 0.10, because it appears only in eight films. This value has two-digit precision in order not to lose information.

9. *Lg10CD*. This value is based on log10(CDcount + 1) and has four-digit precision. As the present article has shown, this is the best value to use if one wants to match words on word frequency. The following rough conversions apply:

Lg10CD	$SUBTL_{CD}$
0.95	0.1
1.93	1
2.92	10
3.92	100

Conclusion

For all too long, psychologists simply have used the frequency measures that were easily available and that were used by their colleagues, without empirically testing their usefulness. With the development of the Elexicon project, such validation studies have become possible. Several interesting findings came out of our analyses, which generalize to other languages.

1. A corpus of 1–3 million words allows researchers to get reliable estimates only for high-frequency words. For words with a frequency smaller than 10 per million, a corpus of at least 16 million words is required.

2. Above 30 million words, the language register tapped into by the corpus is more important than the size.

3. The two most interesting language registers currently available are Internet discussion groups and subtitles. Both can easily be gathered, and they have the highest correlations with word processing variables. On the basis of the English findings, frequencies based on discussion groups seem to be indicated for words longer than seven letters, whereas for short words subtitle frequencies are better (Table 5).

4. There is an issue with words also used as names. If the name frequency is simply added to the word frequency, this results in an overestimation of the word frequency.

5. A frequency measure based on CD outperforms a frequency measure based on simple counts. This has implications for the ways in which corpora must be collected.

AUTHOR NOTE

The authors thank Dave Balota, Barbara Juhasz, Patrick Bonin, and an anonymous reviewer for their comments on an earlier draft. The authors also thank Mike Cortese for kindly providing them with the data of Balota et al. (2004). Correspondence concerning this article should be addressed to M. Brysbaert, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Gent, Belgium (e-mail: marc.brysbaert@ugent.be).

REFERENCES

ADELMAN, J. S., BROWN, G. D. A., & QUESADA, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, **17**, 814-823.

BAAYEN, R. H., DIJKSTRA, T., & SCHREUDER, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory & Language*, **37**, 94-117.

BAAYEN, R. H., FELDMAN, L. B., & SCHREUDER, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory & Language*, **55**, 290-313.

BAAYEN, R. H., PIEPENBROCK, R., & VAN RIJN, H. (1993). The CELEX Lexical Database [CD-ROM]. Philadelphia: Linguistic Data Consortium.

BAAYEN, R. H., WURM, L. H., & AYCOCK, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *Mental Lexicon*, **2**, 419-436.

BALOTA, D. A., & CHUMBLEY, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 340-357.

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.

BLAIR, I. V., URLAND, G. R., & MA, J. E. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers*, **34**, 286-290.

BRYSBAERT, M., DRIEGHE, D., & VITU, F. (2005). Word skipping: Implications for theories of eye movement control in reading. In G. Underwood (Ed.), *Cognitive processes in eye guidance* (pp. 53-77). Oxford: Oxford University Press.

BURGESS, C., & LIVESAY, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, **30**, 272-277.

CARAMAZZA, A., LAUDANNA, A., & ROMANI, C. (1988). Lexical access and inflectional morphology. *Cognition*, **28**, 297-332.

CLAHSEN, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral & Brain Sciences*, **22**, 991-1060.

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.

DRIEGHE, D., POLLATSEK, A., STAUB, A., & RAYNER, K. (2008). The word grouping hypothesis and eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1552-1560.

DUYCK, W., DESMET, T., VERBEKE, L. P. C., & BRYSBAERT, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, **36**, 488-499.

FRANCIS, W., & KUČERA, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

GASKELL, M. G., & DUMAY, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, **89**, 105-132.

GHYSELINCK, M., LEWIS, M. B., & BRYSBAERT, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, **115**, 43-67.

GLANZER, M., & ADAMS, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, **13**, 8-20.

GLANZER, M., & BOWLES, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 21-31.

HAUK, O., & PULVERMÜLLER, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, **115**, 1090-1103.

HOCKLEY, W. E. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1412-1429.

HOWES, D. H., & SOLOMON, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, **41**, 401-410.

HUBER, D. E., CLARK, T. F., CURRAN, T., & WINKIELMAN, P. (2008). Effects of repetition priming on recognition memory: Testing a perceptual fluency–disfluency model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1305-1324.

JESCHENIAK, J. D., & LEVELT, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 824-843.

JOHNSTON, R. A., & BARRY, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, **13**, 789-845.

JUHASZ, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, **131**, 684-712.

KLEPOUSNIOTOU, E., TITONE, D., & ROMERO, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1534-1543.

KUČERA, H., & FRANCIS, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

LEECH, G., RAYSON, P., & WILSON, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

LUND, K., & BURGESS, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, **28**, 203-208.

MCDONOUGH, I. M., & GALLO, D. A. (2008). Autobiographical elaboration reduces memory distortion: Cognitive operations and the distinctiveness heuristic. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1430-1445.

MCKAY, A., DAVIS, C., SAVAGE, G., & CASTLES, A. (2008). Semantic involvement in reading aloud: Evidence from a nonword training study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1495-1517.

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.

NEW, B., BRYSBAERT, M., SEGUI, J., FERRAND, L., & RASTLE, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory & Language*, **51**, 568-585.

NEW, B., BRYSBAERT, M., VERONIS, J., & PALLIER, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, **28**, 661-677.

NEW, B., FERRAND, L., PALLIER, C., & BRYSBAERT, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, **13**, 45-52.

NEW, B., PALLIER, C., BRYSBAERT, M., & FERRAND, L. (2004). *Lexique 2*: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**, 516-524.

O'MALLEY, S., & BESNER, D. (2008). Reading aloud: Qualitative differences in the relation between stimulus quality and word frequency as a function of context. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1400-1411.

PASTIZZO, M. J., & CARBONE, R. F., JR. (2007). Spoken word frequency counts based on 1.6 million words in American English. *Behavior Research Methods*, **39**, 1025-1028.

PEIRCE, C. S. (1877). The fixation of belief. *Popular Science Monthly*, **12**, 1-15.

RASTLE, K., DAVIS, M. H., & NEW, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, **11**, 1090-1098.

RAYNER, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**, 372-422.

RAYNER, K., & DUFFY, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, **14**, 191-201.

SEIDENBERG, M. S., & WATERS, G. S. (1989). Reading words aloud: A mega study [Abstract]. *Bulletin of the Psychonomic Society*, **27**, 489.

SHAOUL, C., & WESTBURY, C. (2008). *A USENET corpus (2005–2008)*. Edmonton: University of Alberta. Retrieved on 10/9/2008 from www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html.

SZPUNAR, K. K., MCDERMOTT, K. B., & ROEDIGER, H. L., III (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1392-1399.

TAFT, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology*, **57A**, 745-765.

THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words*. New York: Columbia University, Teachers College.

UNDERWOOD, B. J. (1961). Ten years of massed practice on distributed practice. *Psychological Review*, **68**, 229-247.

VAN HELL, J. G., & DE GROOT, A. M. B. (1998). Disentangling context availability and concreteness in lexical decision and word translation. *Quarterly Journal of Experimental Psychology*, **51A**, 41-63.

YARKONI, T., SPEER, N. K., BALOTA, D. A., MCAVOY, M. P., & ZACKS, J. M. (2008). Pictures of a thousand words: Investigating the neural mechanisms of reading with extremely rapid event-related fMRI. *NeuroImage*, **42**, 973-987.

YONELINAS, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory & Language*, **46**, 441-517.

ZENO, S. M., IVENS, S. H., MILLARD, R. T., & DUVVURI, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.

ZEVIN, J. D., & SEIDENBERG, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory & Language*, **47**, 1-29.

SUPPLEMENTAL MATERIALS

The norms discussed in this report may be downloaded from http://brm.psychonomic-journals.org/content/supplemental. For further details, see the Availability section above.

NOTES

1. A less interesting aspect of the HAL frequency norms is that the size of the corpus is not particularly clear. Whereas Balota et al. (2007; see also Burgess & Livesay, 1998) wrote that the corpus "consists of approximately 131 million words gathered across 3,000 Usenet newsgroups during February 1995" (p. 450), Lund and Burgess (1996), in their initial report, referred to a corpus of "160 million words of text from Usenet news groups" (p. 205). This makes it hard to interpret the absolute values of the norms.

2. Although the official announcement of the Elexicon project was published in 2007, researchers could already have consulted the database for a few years.

3. The other frequency norms tested in Balota et al. (2004) are copyright protected and, hence, cannot be made freely available.

4. The idea was originally proposed to them by Agnès Bontemps.

5. The observation that frequencies based on children's books do as well as frequencies based on language use in adults is in line with the finding that the age at which words are acquired also plays a role in visual word recognition (see, e.g., Ghyselinck, Lewis, & Brysbaert, 2004; Johnston & Barry, 2006; Juhasz, 2005).

6. We also downloaded a corpus of 10.2 million words coming from 1,867 British films, but these frequencies decreased the fit of the SUBTL norms and, hence, were excluded. Future work with British LDTs will have to indicate whether the lower fit was due to the sample we were able to download or to differences between British and American English. Interestingly, the estimates coming from the pre-1990 American films accounted for some 3% less of the variance in the young readers of Balota et al. (2004), whereas they accounted for 1.5% more in the older participants. This is a nice example of how word frequencies evolve over time and, hence, need to be recalibrated from time to time. We kept the pre-1990 frequencies included in our norms because the variance accounted for by the combined corpus was not lower than that accounted for by the post-1990 films alone. It also provided frequencies for some "old" words that may be of interest.

7. Statistical tests are not needed to check whether these differences are reliable. Due to the large number of observations, differences in explained variance of 1% go far beyond the conventional significance levels in psychology.

8. The authors thank Emmanuel Keuleers for providing them with a cross-check of some findings with Celex.

9. Making different choices did not make a practical difference.

10. The Zeno database also has a measure in which the raw frequencies are corrected for CD, the so-called U-index. However, when we used this index, we explained 0.5% less of variance of the Elexicon LDTs than with raw word frequency.

11. Participants did not notice this because, in the experiments of the Elexicon project, words were presented entirely in uppercase (Balota et al., 2007).

(Manuscript received December 23, 2008;
revision accepted for publication April 2, 2009.)