Behavior Research Methods (2018) 50:1–25
https://doi.org/10.3758/s13428-017-1001-8

# On the predictive validity of various corpus-based frequency norms in L2 English lexical processing

Xiaocong Chen[1] · Yanping Dong[1] · Xiufen Yu[1]

Published online: 16 January 2018
© Psychonomic Society, Inc. 2018

**Abstract**
The predictive validity of various corpus-based frequency norms in first-language lexical processing has been intensively investigated in previous research, but less attention has been paid to this issue in second-language (L2) processing. To bridge the gap, in the present study we took English as a case in point and compared the predictive power of a large set of corpus-based frequency norms for the performance of an L2 English visual lexical decision task (LDT). Our results showed that, in general, the frequency norms from SUBTLEX-US and WorldLex–Blog tended to predict L2 performance better in reaction times, whereas the frequency norms from corpora with a mixture of written and spoken genres (CELEX, WorldLex–Blog, BNC, ANC, and COCA) tended to predict L2 accuracy better. Although replicated in both low- and high-proficiency L2 English learners, these patterns were not exactly the same as those found in LDT data from native English speakers. In addition, we only observed some limited advantages of the lemma frequency and contextual diversity measures over the wordform frequency measure in predicting L2 lexical processing. The results of the present study, especially the detailed comparisons among the different corpora, provide methodological implications for future L2 lexical research.

**Keywords** Corpus-based frequency norms · L2 lexical processing · Lemma frequency · Contextual diversity · Predictive validity

Word frequency has been shown to be one of the most important predictors in lexical processing (e.g., Brysbaert, Buchmeier, et al., 2011; Brysbaert, Mandera, & Keuleers, 2017; Ellis, 2002). High-frequency words are processed more quickly and accurately than low-frequency words, a well-attested finding in visual word recognition (e.g., Adorni, Manfredi, & Mado Proverbio, 2013; Baayen, Feldman, & Schreuder, 2006; Balota & Chumbley, 1984, 1990; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Howes & Solomon, 1951; Monsell, Doyle, & Haggard, 1989; Yap & Balota, 2009), spoken word recognition (e.g., Connine, Mullennix, Shernoff, & Yelen, 1990; Dahan, Magnuson, & Tanenhaus, 2001; Dufour, Brunellière, & Frauenfelder, 2013; Garlock, Walley, & Metsala, 2001; Luce & Pisoni, 1998; Moulin & Richard, 2015; Savin, 1963), reading (e.g., Rayner & Duffy, 1986; Schilling, Rayner, & Chumbley, 1998), and word production (e.g., Balota & Chumbley, 1985; Forster & Chambers, 1973; Jescheniak & Levelt, 1994; Shatzman & Schiller, 2004). This robust frequency effect is observed not only in the first language (L1) among native speakers of different languages, but also in the second language (L2) among different types of bilinguals (e.g., Akbari, 2015; Bradlow & Pisoni, 1999; Cop, Keuleers, Drieghe, & Duyck, 2015; De Groot, Borgwaldt, Bos, & van den Eijnden, 2002; Diependaele, Lemhöfer, & Brysbaert, 2013; Gollan, Montoya, Cera, & Sandoval, 2008; Gollan et al., 2011; Imai, Walley, & Flege, 2005; Lemhöfer et al., 2008; Schmidtke, 2014; Shi, 2014, 2015; van Wijnendaele & Brysbaert, 2002; Whitford & Titone, 2012). Given its importance, word frequency has played a key role in many computational models of lexical processing (see the reviews in Adams, 1979; Baayen, 2010; Colombo, Pasini, & Balota, 2006; Kuperman & van Dyke, 2013; Monaghan, Chang, Welbourne, & Brysbaert, 2017).

In practice, word frequencies have been widely estimated by corpus-based frequency norms, which are derived from word counts in a corpus. Given the availability of various corpus-based frequency norms, the *predictive validity* of these frequency norms—that is, how well they can predict lexical processing—has received increasing interest in recent research, with a large number of studies making comparisons between different

✉ Yanping Dong
ypdong@gdufs.edu.cn

[1] Bilingual Cognition and Development Lab, Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China

corpus-based frequency norms (e.g., see the reviews in Brysbaert, Buchmeier, et al., 2011; Brysbaert & New, 2009). However, almost all these studies have focused on L1 lexical processing, and less attention has been paid to this issue in L2. Thus, taking English as a case in point, in the present study we aimed to compare the predictive power of a large set of frequency norms from different corpora on L2 lexical processing, hoping to provide methodological implications for future L2 lexical research.

## Predictive validity of corpus-based frequency norms in L1

Recent research has investigated intensively how well various corpus-based frequency norms predict L1 lexical processing, with the aim to discover the "optimal" frequency norm that can be used in lexical processing research. As is shown by this line of research, the predictive power of corpus-based frequency norms may be influenced by the characteristics of a corpus, including the language register of corpus materials, corpus size, and time period of corpus materials (see the reviews in Brysbaert, Buchmeier, et al., 2011).

The language register of corpus materials has been thought to strongly affect the predictive power of the derived frequency norms (Balota et al., 2004; Brysbaert, Buchmeier, et al., 2011; Brysbaert & New, 2009; Burgess & Livesay, 1998; Zevin & Seidenberg, 2002). In particular, the genre of corpus materials plays an important role. For instance, a series of recent studies demonstrated that frequency norms derived from subtitles of films and TV programs tended to outperform those from printed texts in accounting for the variance of lexical processing time (and sometimes also accuracy) among native speakers of different languages (Brysbaert, Keuleers, & New, 2011; Brysbaert, Buchmeier, et al., 2011; Brysbaert & New, 2009; Cai & Brysbaert, 2010; Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011; Dimitropoulou & Carreiras, 2010; Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013; Keuleers, Brysbaert, & New, 2010; Mandera, Keuleers, Wodniecka, & Brysbaert, 2015; New, Brysbaert, Veronis, & Pallier, 2007; Soares et al., 2015; van Heuven, Mandera, Keuleers, & Brysbaert, 2014; but see an exception in Pham, 2014, for Vietnamese). Internet-based frequency norms (e.g., based on Web newsgroup discussion), particularly those derived from recent social media sources (e.g., materials from blogs, Facebook, or Twitter), were also found to show comparable or better performance in predicting lexical processing, compared with other frequency norms (Balota et al., 2004; Burgess & Livesay, 1998; Gimenes & New, 2016; Herdağdelen & Marelli, 2017). The superiority of subtitle-based or Internet-based frequency norms could be attributed to the increasing dominance of TV and the Internet in people's daily lives, which makes subtitles and Internet materials more representative of language use (e.g., Brysbaert, Buchmeier, et al., 2011; Gimenes & New, 2016; but see the different view in Baayen, Milin, & Ramscar, 2016, and Heister & Kliegl, 2012). In contrast, the poor predictive power of traditional written frequency norms in lexical processing may be explained partly by the fact that frequencies of certain words are often underestimated or overestimated due to a tendency to avoid repetitions (and thus with exaggerated lexical diversity) in printed texts (Brysbaert & New, 2009). In addition to genre, the dialectal variety also matters. For example, it was discovered that frequency norms from UK sources were better correlated with RTs from UK participants, whereas frequency norms from US sources were better correlated with RTs from US participants (Herdağdelen & Marelli, 2017; van Heuven et al., 2014).

Two other important factors influencing the predictive power of corpus-based frequency norms are the corpus size (Balota et al., 2004; Burgess & Livesay, 1998; Herdağdelen & Marelli, 2017) and the time period of corpus materials (e.g., Brysbaert, Keuleers, & New, 2011). In terms of corpus size, it was argued that frequency norms obtained from a large corpus were better at predicting lexical processing than those obtained from a small corpus, because a large corpus provided more reliable estimates for words at the lower end of the frequency spectra (Brysbaert, Buchmeier, et al., 2011). However, as was shown by Brysbaert and New (2009), when the corpus size was large enough (e.g., >16 million), a larger corpus size did not further improve the predictive power. In terms of the time period of corpus materials, frequency norms obtained from more recent materials were found to be better at predicting lexical processing than those obtained from old materials (Brysbaert, Keuleers, & New, 2011; Brysbaert & New, 2009, see footnote 6; Gimenes & New, 2016).

In addition to the selection of corpora, the selection of different types of frequency measures has been discussed. The wordform frequency measure (i.e., the frequency of a single wordform in a corpus, also called "surface frequency") has been primarily adopted in relevant literature, but alternative measures have also been proposed. One is the lemma frequency measure, defined as the sum of frequencies of all inflected forms of a word (e.g., Brysbaert & New, 2009). Although most previous research failed to find an advantage of lemma frequencies over wordform frequencies in predicting lexical processing (e.g., Baayen, Wurm, & Aycock, 2007; Brysbaert, Buchmeier, et al., 2011; Brysbaert & New, 2009), some recent studies had observed stronger predictive power of lemma frequencies (Gimenes, Brysbaert, & New, 2016; Keuleers et al., 2010). Another alternative is the contextual diversity measure, referring to the number of documents or texts in a corpus containing a given word (McDonald & Shillcock, 2001). The contextual diversity measure has been found to offer a slightly better prediction for lexical processing time than the wordform frequency measure (Adelman, Brown, & Quesada, 2006; Brysbaert & New, 2009; Johns, Gruenenfelder, Pisoni, & Jones, 2012), but this was not replicated in Gimenes and New (2016).

## Predictive validity of corpus-based frequency norms in L2

Surprisingly, despite the importance of word frequency in L2 lexical processing, few studies have investigated the predictive validity of various corpus-based frequency norms in the L2. Since L2 learners may receive very different language input from native speakers, frequency norms based on L1 corpora may not very well reflect L2 learners' lexical knowledge. Although the validity of using existing L1 corpus-based frequency norms in L2 lexical processing was questioned (Dong & Yuan, 2008, in Chinese), only Shi (2015) has explored this issue. He compared the strength of correlations between four corpus-based frequency norms and the error rates of spoken English words among monolinguals and different groups of bilinguals. Overall, for both native and nonnative speakers, he found that frequency norms from the three recent corpora (the comprehensive corpus ANC [American National Corpus; Reppen & Ide, 2004]; Web-based corpus HAL [Hyperspace Analogue to Language; Lund & Burgess, 1996]; and the subtitle-based corpus SUBTLEX-US [Brysbaert & New, 2009; Brysbaert, New, & Keuleers, 2012]) outperformed the traditional written frequency norm from HML (the Hoosier Mental Lexicon database; Nusbaum, Pisoni, & Davis, 1984), with the former three frequency norms showing stronger correlations with the error rates of spoken words. Moreover, although the three recent frequency norms (ANC, HAL, and SUBTLEX-US) showed similar correlation strengths with the error rates, ANC appeared to show a slightly larger correlation coefficient than the other two. However, the general patterns described above varied among bilinguals with different ages of onset of English acquisition (AoA) and lengths of English learning experience. For late bilinguals, all of the four frequency norms significantly correlated with the error rates. For early bilinguals, significant correlations with error rates were found only for ANC and HAL. But for intermediate bilinguals (larger English AoAs than early bilinguals but more L2 English schooling experience than late bilinguals), none of the frequency norms showed significant correlations with the error rates.

However, there are limitations of Shi's (2015) study. First, he did not employ online measures (such as RTs), and nothing is known about the predictive power of those frequency norms on L2 lexical processing efficiency. Second, the bilinguals in Shi's (2015) study had been immersed in the L2-speaking country to varying degrees, and it is unknown whether the conclusions can be extended to L2 learners in an L1-dominant environment. Third, only a small set of frequency norms from American English corpora were compared. With the availability of other influential English corpora, including those from British English, it would be better to compare a large set of frequency norms in a single study so that their use in L2 lexical research is better justified.

## The present study

Thus, the goal of the present study is to compare the predictive power of various corpus-based frequency norms on L2 lexical processing in an L1-dominant environment (China, in this study). We selected frequency norms from different English corpora with varying corpus sizes, language varieties and times, in order to present a more comprehensive picture. L2 lexical processing was gauged by a visual lexical decision task (LDT hereafter), as it is the most common task used in the relevant literature and sensitive to the influence of word frequency (e.g., Soares et al., 2015). Both RT (online measure) and accuracy (offline measure) were analyzed. Besides, using the same stimuli, we compared our results of L2 lexical processing with those of L1 lexical processing from native English speakers, with the L1 baseline data obtained from the English Lexicon Project (ELP hereafter; Balota et al., 2007, http://elexicon.wustl.edu/). We also ran similar comparisons with the L1 baseline data obtained from the British Lexicon Project (BLP hereafter; Keuleers, Lacey, Rastle, & Brysbaert, 2012, http://crr.ugent.be/blp/) but the results were reported in Appendix D.[1] In addition, we further compared the predictive power of different frequency norms on the LDT data from learners with different levels of L2 English proficiency. Moreover, we primarily focused on the wordform frequency measure as with most previous studies (e.g., Brysbaert & New, 2009), but lemma frequency and context diversity measures were also examined in the present study.

## Method

### Participants

A total of 77 Chinese university students (68 females and nine males) from Guangdong University of Foreign Studies, aged between 17 and 24 ($M$ = 19.9 years, $SD$ = 1.35), majoring in English language and culture, were paid to take part in the LDT. All the participants were native Chinese speakers and mostly learned English as a foreign language at school (years of English learning: $M$ = 12.8 years, $SD$ = 2.23, range = 6 to 17; age of onset of English Acquisition: $M$ = 8.1 years old, $SD$ = 2.40, range = 3 to 13). Their English proficiency, assessed

[1] In the present study, we mainly compared our L2 data with the L1 data from the ELP (Balota et al., 2007) rather than from the BLP (Keuleers et al., 2012), because the BLP data only included mono- and disyllabic words, whereas 25% of our stimuli were multisyllabic (over two syllables). Nevertheless, we still ran parallel regression analyses on our L2 data and the BLP data by excluding the multi-syllabic items in the datasets, and the results were reported in Appendix D.

by the Quick Placement Test (QPT hereafter; University of Cambridge Local Examination Syndicate, 2001),[2] covered the B1 (lower-intermediate, 15 participants), B2 (upper-intermediate, 38 participants), and C1 (advanced, 24 participants) levels defined in the Common European Framework.

## Selection of frequency norms

The present study selected frequency norms from the following 12 representative English corpora:

1) KF frequency norms (Kučera & Francis, 1967);
2) CELEX English database (Baayen, Piepenbrock, & Gulikers, 1995);
3) BNC (British National Corpus, based on the wordlist from Kilgarriff, 2006);
4) COCA (Corpus of Contemporary American English; Davies, 2009);
5) ANC (The American National Corpus; Reppen & Ide, 2004);
6) SUBTLEX-UK (van Heuven et al., 2014);
7) SUBTLEX-US (Brysbaert & New, 2009; Brysbaert et al., 2012);
8) HAL (Hyperspace Analogue to Language; Lund & Burgess, 1996);
9) WestburyLab USENET (Shaoul & Westbury, 2006);
10) WorldLex (Gimenes & New, 2016);
11) Google Books British English corpus (GB_BrE) (Davies, 2011b; Michel et al., 2011);
12) Google Books American English corpus (GB_AmE) (Davies, 2011a; Michel et al., 2011).

The first two corpora (KF and CELEX) are two traditional corpora commonly used in experimental psychology (see, e.g., the detailed discussion in Brysbaert & New, 2009; van Heuven et al., 2014), with KF derived from the Brown corpus containing written American English texts, and CELEX derived from the COBUILD corpus containing written texts and a small proportion of spoken materials. The next three corpora (BNC, COCA, and ANC) are influential large-scale corpora often referred to in corpus linguistics (see Davies, 2009), all of which incorporate a wide range of genres, including both printed texts and a certain proportion of spoken materials. The next two corpora (SUBTLEX-UK and SUBTLEX-US) are recently developed subtitle-based corpora, respectively derived from BBC broadcasts in the UK, and films and TV programs in the US. The next three corpora (HAL, WestburyLab USENET, and WorldLex) are Internet-based corpora collected from different periods of time. HAL and USENET are both based on the USENET newsgroup postings, whereas WorldLex consists of three subcorpora (Blog, Twitter, News) derived, respectively from blogs, tweets, and newspapers. The last two corpora (GB_BrE and GB_AmE) in the present study are based on Google's databases of millions of digitalized books published, respectively in the UK and US between 1810 and 2009. Detailed information and availability of these corpora are presented in Appendix A.

Moreover, instead of using corpus raw frequencies, the present study adopted the standardized Zipf frequency values proposed by van Heuven et al. (2014), a logarithmic scale resembling a 7-point Likert scale. The Zipf values were calculated by the formula: $zipf\ value = \log10\left(\frac{f+1}{N}\right) + 3$, in which $f$ stands for raw frequency of a word in a given corpus, and $N$ stands for the corpus size in millions that equals the sum of the number of word tokens and word types in a corpus.[3] The use of the Zipf scale allowed us to compare frequency estimates from different-sized corpora, and handle certain words that were missing and thus had zero frequencies in some corpora (see the relevant disussions in Brysbaert & Diependaele, 2013).

## Stimuli

A total of 370 English words (including 246 stem forms and 124 inflected forms) were selected as the word stimuli in the LDT. These words were originally sampled from word frequency lists taken from ten corpora (KF, CELEX, BNC, ANC, COCA, SUBTLEX-UK, SUBLEX-US, HAL, USENET, and WorldLex). To ensure that the word stimuli span a wide frequency scale in each corpus, we first selected a set of words from each of the three predefined frequency bands (low: 1–10 per million; medium: 10–100 per million;

[2] The Quick Placement Test (QPT) used in this study was a paper-and-pencil test, which consisted of 60 multiple choice items, with each correctly answered item scored 1. According to the QPT manual (Geranpayeh, 2003), English proficiency can be mapped onto the Council-of-Europe levels or the ALTE levels based on their total QPT scores (11–17: A1–Breakthrough; 18–29: A2–Elementary; 30–39: B1–Lower intermediate; 40–47: B2–Upper intermediate; 48–54: C1–advanced; 55–60: C2–very advanced).

[3] The exact corpus size of HAL is not clearly documented in the literature. It has been claimed to be 131 million words (Balota et al., 2007; Burgess & Livesay, 1998), or 160 million words (Lund & Burgess, 1996), or even more than 400 million (Herdağdelen & Marelli, 2017). In the present study, following Marc Brysbaert's note (see the blog at http://crr.ugent.be/archives/1352, retrieved on June 20, 2016), we estimated the corpus size of HAL to be roughly 437 million by summing up all the HAL occurrences of the words in English Lexicon Project (Balota et al., 2007; http://elexicon.wustl.edu/). Moreover, WorldLex only listed the frequency of a word in each of the three respective subcorpora (Blog, Twitter, and News). The raw frequency of a word in the whole WorldLex corpus in the present study was calculated by adding up the raw frequencies of that word in the three subcorpora and then transformed into the corresponding Zipf values.

high: 100–1,000 per million) in each corpus[4] and then assembled the selected words from each list as the final set of word stimuli (see Appx. Table 10 for the detailed frequency distribution of the stimuli). Moreover, we tried to select a wide range of regular and irregular inflected forms including plural nouns, different types of inflected verb forms (third person singular forms, *-ing* forms, past tense forms and past participles), and comparative and superlative forms of adjectives. We also made sure that our selected words were included in ELP (Balota et al., 2007) so that comparison with the data from native English speakers was possible. Pearson correlation analyses showed that different corpus-based wordform frequency norms for the final word stimuli were highly correlated (all *r*s > .75, *p*s < .001; see Appx. Table 11 for more details). The 370 words varied in the number of letters ($M$ = 6.59, $SD$ = 2.13, range = 3 to 13) and number of syllables ($M$ = 2.03, $SD$ = 0.94, range = 1 to 5).

Another 370 pronounceable nonword stimuli were created by the pseudoword generator Wuggy (Keuleers & Brysbaert, 2010), which generated nonwords perfectly matched with the word stimuli in terms of the number of letters and syllables, by substituting subsyllabic elements of words with other subsyllabic elements from other words. Among a set of nonword candidates generated by the software for each word, we picked up the nonwords that had small deviation from the matched words in terms of transition probability and number of neighbors. Furthermore, for those inflected word stimuli, we selected the nonword stimuli that retained the same inflectional affixes. Appendix C lists all the stimuli.

## Procedure

In the LDT, the presentation of stimuli and the recording of responses were controlled by E-prime 2.0 (Schneider, Eschman, & Zuccolotto, 2001). Participants were asked to decide as quickly and accurately as possible whether the string of letters presented at the center of the computer screen was an English word or not. Participants were instructed to press the "F" key (labeled "YES") on the keyboard if the string was an English word they knew, and the "J" key (labeled "NO") if the string was not.

The 740 items were arranged into two lists, with each list further divided into five blocks, respectively. Each block thus contained 74 trials (half words and half nonwords). Each trial began with a fixation point (+) presented at the center of the computer screen for 500 ms. Then the fixation point was immediately replaced by an experimental stimulus (word or nonword). The stimulus was displayed in 18-point lowercase Courier New font at the center of the screen, and remained on the screen until participants responded or until 2,500 ms elapsed. The next trial then started after a blank screen lasting 500 ms. The order of the two lists was counterbalanced across the participants, and the order of the blocks in a list and the trials in a block was randomized per participant. There were short breaks between blocks and lists. Before the experimental trials, participants received 24 practice trials (12 words and 12 nonwords, not used in the experimental trials) to familiarize themselves with the procedure. Participants were allowed to start the experiment trials only if they achieved an accuracy of 70% in practice trials; otherwise they were asked to have another practice. The decision task lasted approximately 30–40 min.

After the decision task, participants were asked to complete the QPT within 30 min and a language background questionnaire on their language learning experience. In all, it took about 75 min for each participant to complete the entire procedure.

## Statistical analysis

The LDT data from five participants were excluded, due to their extremely low accuracy rates of nonwords (less than 60%) that indicated excessive use of guessing strategies. Data from the remaining 72 participants were retained for further statistical analysis implemented in the software R (R Development Core Team, 2017). The RT data of erroneous responses were discarded. To reduce the disproportionate influence of extreme RT data on subsequent item analysis, we also excluded RTs shorter than 200 ms, and RTs beyond three *SD*s from the mean raw RT of each participant, which accounted for 1.96% of the RT data of all the correct responses. Participants' mean accuracy rates were 84.7% for words ($SD$ = 7.7%, range = 65.1% to 96.4%) and 88.5% for nonwords ($SD$ = 9.0%, range = 66.2% to 98.9%). Participants' mean RTs of correct responses were 783.2 ms for words ($SD$ = 112.22, range = 570.0 to 1,097.6 ms) and 890.1 ms for nonwords ($SD$ = 150.46, range = 601.0 to 1,278.8 ms).

Besides, we retrieved the RT and accuracy data of the LDT data from ELP (Balota et al., 2007) for further comparisons. The mean accuracy rate of all our word stimuli for English native speakers in ELP was 95.3% ($SD$ = 8.1%, range = 32% to 100%), and the mean RT was 672.8 ms ($SD$ = 89.55, range = 511.7 to 1,112.0 ms). As compared to native English speakers, our L2 learners of English responded less accurately (as indicated by Wilcoxon signed-rank test, $z = -10.016$, $p < .001$) and more slowly [as indicated by paired $t$ test, $t(369) = 27.487$, $p < .001$].

To examine the predictive power of various frequency measures on lexical processing, we conducted regression analyses on the RT and accuracy data, with each corpus-based frequency norm included as a predictor in separate regression

[4] When selecting words from each corpus, we did not consider words with a frequency of less than 1 per million because these words are unfamiliar to most L2 learners. Words more than 1,000 per million were also not considered because most of these words are function words that L2 learners are extremely familiar with. However, the same words can occur in different frequency bands in different corpora, so when we assembled the words selected from all the corpora, a few words could fall below 1 per million or above 1,000 per million in certain corpora, but such words were still retained in the present study.

models. As to the analysis of RT, the RT data were log-transformed to reduce the skewing in the distributions of the original RT data (Baayen, 2008, p. 31; Baayen et al., 2006). When modeling the data, for the data of native speakers from ELP, we employed polynomials (including a linear and a quadratic term) to model the nonlinear effect of frequency on RTs (i.e., a leveling-off effect for high-frequency words) found in previous studies (e.g., Brysbaert & New, 2009; Gimenes & New, 2016; Soares et al., 2015; van Heuven et al., 2014). In contrast, for the data of our L2 participants, only the linear term of the frequency norms was included in the final models because preliminary visual inspection of the RT data did not reveal a nonlinear relationship between our RT data and frequency norms, and application of the non-linear functions did not improve the model fit as indicated by the likelihood ratio tests. With respect to accuracy data, logistic regression analysis was applied[5] (Baayen, 2008, p. 197). Polynomials (including a linear, a quadratic, and a cubic term) were used to capture the nonlinear relationship between the word frequency and the accuracy data, since preliminary visual inspection showed that there was a ceiling effect for the accuracy of high-frequency words in both our L2 data and the ELP data. Moreover, to control for the word length effect on the RT and accuracy data, the number of letters[6] was entered into the regression models as a covariate. The nonlinear functions of the word length were also added into the models only when the inclusion of the nonlinear terms of the word length significantly improved the model fit, as indicated by the likelihood ratio tests.

Apart from the 12 corpora, we additionally analyzed the frequency norms from the three sub-corpora in WorldLex (Blog, Twitter, and News), as Gimenes and New (2016) demonstrated that frequency norms from the subcorpora Twitter and Blog performed well in predicting lexical processing. Thus, 15 linear regression models were constructed, respectively for both the accuracy and RT data. Besides, when we compared different types of frequency measures (wordform vs. lemma, wordform vs. contextual diversity), we built separate regression models for each type of frequency measure. The differences of the predictive power between various frequency measures on the accuracy and RT data were evaluated through the comparisons of different non-nested regression models. The models were assessed by the following criteria: (1) $R^2$ values (or McFadden pseudo-$R^2$ values for logistic regression), with a larger value indicating a larger amount of variance of the data explained by the regression model; (2) Akaike information criterion (AIC) values, with a smaller value indicating a better fit of the models for the data (Vrieze, 2012). Vuong tests (Vuong, 1989), implemented by the *nonnest2* package (Merkle, You, & Preacher, 2016) in R, were used as supplementary diagnostics to examine whether the difference between two given models was significant or not in terms of their goodness of fit.

[5] Unlike most previous research (e.g., Brysbaert & New, 2009), we did not conduct linear regression analyses on the proportional data (accuracy rates). Instead, we followed Baayen's (2008, p. 197) practice and applied the logistic regression by using the glm function in R to model the accuracy data, because proportion data did not fully satisfy the assumptions of linear regression modeling (see more details in Baayen, 2008, p. 196). We transformed our accuracy data into a format in which number of successes (correct responses trials) and failures (wrong responses trials) for each item were listed separately, as required by the glm function. For the accuracy data of native speakers from ELP (and BLP reported in Appx. D), we obtained the number of successes (correct responses trials) and failures (incorrect responses trials) for each item from the raw trial-level data.

[6] Another word-length variable (number of syllables) was not included as a predictor in the regression analyses, because the two word-length variables (number of syllables and number of letters) were highly correlated ($r = .844$, $p < .001$) for the word stimuli in the LDT. To avoid the multicollinearity problem, we only included the number of letters in the regression models to control for the word-length effect.

## Results

### Question 1: How do various corpus-based frequency norms differ in predicting L2 English lexical processing? Are these patterns found in L2 lexical processing the same as in the L1?

**Analysis of RT** Only words with accuracy rates of 66.7% or more (correct responses by more than two thirds of the participants in either group) in both our data and the ELP data (303 items in total) were entered into the linear regression analyses of RT. Table 1 summarizes the predictive power of the frequency norms on both L2 and L1 RT data, with the frequency norms ranked according to their explained variance ($R^2$) and model goodness of fit (according to the AIC). The results of our regression analysis showed that after controlling word length, all 15 frequency norms exerted statistically significant effects on the RTs of both our L2 and native English speakers. Combined with word length, each of the frequency norms accounted for a large proportion of RT variance for both our L2 participants (over 65%) and native English speakers (over 50%), as is shown by the $R^2$ values in Table 1. In addition, the amount of variance explained by the frequency norms seemed to be larger for our L2 RT data, suggesting a larger word frequency effect in L2 than in L1 lexical processing.

For both L2 and native English speakers, the performance of the frequency norms formed a continuum in terms of their predictive power. We observed a gradual change of goodness of fit along the continuum, with certain frequency norms showing an advantage over others, as evidenced by subsequent Vuong tests (see the details in Appx. Tables 12 and 13). For our L2 English participants, SUTBLEX-US seemed to show a superior performance, as it was (marginally) significantly better than most of the other frequency norms, ($ps < .1$; see Appx. Table 12). The second-best frequency norm was WorldLex–Blog, as it showed a (marginally) significant advantage over most of the remaining frequency norms ($ps < .1$; see Appx. Table 12). For the native English speakers,

**Table 1** Predictive power of Corpus-based frequency norms on L2 and L1 ELP RT data ($N$ = 303 items)

| L2 English Speakers | | | US Native English Speakers | | |
|---|---|---|---|---|---|
| **Corpus** | $R^2$ (%) | AIC | **Corpus** | $R^2$ (%) | AIC |
| **SUBTLEX-US** | 75.3 | – 717.916 | **COCA** | 57.4 | – 746.793 |
| **WL-blog** | 74.2 | – 704.791 | **SUBTLEX-US** | 57.1 | – 744.671 |
| **WL** | 73.2 | – 693.538 | **CELEX** | 56.3 | – 738.90 |
| **CELEX** | 72.9 | – 689.541 | **WL** | 56.0 | – 736.678 |
| **COCA** | 72.3 | – 682.989 | **WL-blog** | 55.8 | – 735.692 |
| **WL-twitter** | 72.2 | – 682.693 | **WL-news** | 55.7 | – 734.799 |
| **SUBTLEX-UK** | 71.7 | – 677.267 | **BNC** | 55.3 | – 732.408 |
| **BNC** | 70.9 | – 668.747 | **GB_AmE** | 55.2 | – 731.862 |
| **USENET** | 70.9 | – 668.589 | **ANC** | 54.9 | – 729.564 |
| **HAL** | 70.8 | – 667.025 | **HAL** | 54.9 | – 729.461 |
| **ANC** | 69.4 | – 653.353 | **USENET** | 54.9 | – 729.394 |
| **WL-news** | 68.8 | – 647.593 | **WL-twitter** | 54.2 | – 725.035 |
| **KF** | 68.3 | – 642.223 | **SUBTLEX-UK** | 54.1 | – 724.431 |
| **GB_AmE** | 67.7 | – 636.343 | **GB_BrE** | 53.3 | – 718.748 |
| **GB_BrE** | 65.6 | – 617.931 | **KF** | 53.0 | – 716.722 |

(1) The frequency norm models are ranked in the order of the goodness of fit, with a larger $R^2$ value indicating a larger proportion of variance explained by the model, and a smaller AIC value indicating a better fit of the model for the data. (2) WL: WorldLex; GB: Google books

however, COCA had the largest predictive power, showing a (marginally) significantly better performance than most of the other frequency norms ($p$s < .1; see Appx. Table 13). But SUTBLEX-US, as was found in previous research (Brysbaert & New, 2009), also offered good predictions, which had a (marginally) significantly better fit than the last five frequency norms (USENET, WorldLex–Twitter, SUBTLEX-UK, GB_BrE, and KF, $p$s < .1; see Appx. Table 13).

**Analysis of accuracy** All 370 words were entered into the regression analysis of the accuracy data. Table 2 gives a summary of predictive power for both the L2 and L1 accuracy data, with the frequency norms ranked according to the McFadden pseudo-$R^2$ and AIC values. Regression analysis again showed that all the frequency norms had some predictive power on both the L2 and L1 accuracy data, since the effects of the linear, quadratic, and cubic terms of each norm were found to be statistically significant. A close look at the data seemed to suggest that the word frequency effect was more prominent in L2 lexical processing than in L1. This may be due to the fact that there was a larger ceiling effect for the high-frequency words among the native English speakers than our L2 English speakers, as native English speakers had substantially high accuracy rates for our word stimuli (89.7% of the words with an accuracy rate of over 90%).

As what was found in our RT analysis, the predictive power of these frequency norms changed gradually along a continuum, with certain frequency norms outperforming others, as indicated by subsequent Vuong tests (see the details in

**Table 2** Predictive power of corpus-based frequency norms on L2 and L1 accuracy data ($N$ = 370 items)

| L2 English Speakers | | | US Native English Speakers | | |
|---|---|---|---|---|---|
| Corpus | McFadden $R^2$ (%) | AIC | Corpus | McFadden $R^2$ (%) | AIC |
| **CELEX** | 55.1 | 4,287.868 | **WL-twitter** | 29.2 | 1,225.316 |
| **WL-blog** | 53.0 | 4,488.429 | **WL** | 29.1 | 1,226.124 |
| **ANC** | 52.4 | 4,545.362 | **SUBTLEX-US** | 28.7 | 1,233.336 |
| **BNC** | 52.1 | 4,571.751 | **COCA** | 28.4 | 1,239.204 |
| **COCA** | 52.1 | 4,576.538 | **WL-blog** | 28.3 | 1,241.290 |
| **GB_AmE** | 51.1 | 4,669.939 | **HAL** | 27.4 | 1,255.932 |
| **WL** | 51.0 | 4,675.427 | **WL-news** | 27.2 | 1,258.821 |
| **SUBTLEX-UK** | 50.0 | 4,776.767 | **USENET** | 26.8 | 1,266.891 |
| **USENET** | 48.7 | 4,895.700 | **BNC** | 24.6 | 1,303.861 |
| **KF** | 48.7 | 4,897.835 | **ANC** | 24.3 | 1,309.283 |
| **HAL** | 48.1 | 4,949.800 | **SUBTLEX-UK** | 24.2 | 1,311.595 |
| **WL-twitter** | 46.8 | 5,075.021 | **CELEX** | 23.7 | 1,319.432 |
| **GB_BrE** | 46.8 | 5,077.230 | **GB_AmE** | 22.9 | 1,333.472 |
| **WL-news** | 46.6 | 5,100.425 | **KF** | 21.8 | 1,351.574 |
| **SUBTLEX-US** | 44.1 | 5,332.510 | **GB_BrE** | 19.5 | 1,391.625 |

(1) The frequency norm models are ranked in the order of the goodness of fit, with a larger McFadden $R^2$ value indicating a larger proportion of variance explained by the model, and a smaller AIC value indicating a better fit of the model for the data. (2) WL: WorldLex; GB: Google books

Appxs. Tables 14 and 15). For our L2 English participants, CELEX provided the best prediction, as it was (marginally) significantly better than most of the other frequency norms (*p*s < .1; see Appx. Table 14). The next four corpora (WorldLex–Blog, BNC, ANC, and COCA) seemed to be less good, but they still showed a (marginally) significant advantage than the last seven frequency norms (USENET, KF, HAL, WorldLex–Twitter, GB_BrE, WorldLex–News, and SUBTLEX-US) (*p*s < .1; see Appx. Table 14). A noteworthy aspect was that most of these best-fitting corpus frequency norms (CELEX, BNC, ANC, and COCA) were derived from the corpora with a mixture of written and spoken genres. Besides, the second-best corpus frequency norm WorldLex–Blog can also be considered to be based on the materials with a mixture of written and spoken genres since bloggers tend to use both formal (akin to written) and informal (akin to spoken) styles in writing blogs. These results suggested that corpus frequency norms derived from a mixture of both written and spoken genres tended to make better predictions about the accuracy of L2 English lexical processing. In contrast, for the accuracy data of US native English speakers, all the US English Web-based corpus frequency norms (WorldLex and its three subcorpora, HAL, and USENET), SUBTLEX-US, and COCA seemed to show a better performance than most of the other frequency norms, as Vuong tests showed that these frequency norms had a (marginally) significant advantage over the last four frequency norms (CELEX, two Google-book-based corpora, and the traditional written corpus KF) (*p*s < .1; see Appx. Table 15). These results indicated that overall, the Web-based and subtitle-based frequency norms in US English, together with the frequency norms from the large-scale balanced US English corpus COCA, provided better predictions for the accuracy data in L1 English lexical processing of US participants.

*In sum, our data analysis revealed that frequency norms from SUBTLEX-US and WorldLex–Blog better predicted the RT data of L2 English lexical processing, whereas the frequency norms from a mixture of written and spoken genres (such as CELEX, WorldLex–Blog, BNC, ANC, and COCA) better predicted the L2 accuracy data. These patterns were different from those found in L1 lexical processing of US native English speakers, in which COCA and SUTBLEX-US showed better predictions for the RT data, whereas the Web-based corpus frequency norms, SUBLTEX-US, and COCA predicted the accuracy data better.*

## Question 2: How do various corpus-based frequency norms differ in predicting L2 English lexical processing among learners of different L2 proficiencies?

To examine whether the results reported in the last section would be different for learners of different L2 proficiencies, we further selected two subgroups from our L2 participants based on participants' QPT scores (60 scores in total, range: 30–54) and the QPT classification criterion of English proficiency (Geranpayeh, 2003), with 23 as intermediate learners (scoring 30–42 in the QPT), and 24 as advanced learners (scoring 48–54 in the QPT). The RT and accuracy data from the two subgroups of learners were re-calculated and re-analyzed with the same methods described previously.

As is shown in Tables 3 and 4, similar patterns were observed for both intermediate and advanced learners. For the RT data, SUTBLEX-US and WorldLex–Blog showed the largest predictive power for either subgroup of learners, consistent with the results reported in the previous section. With regard to the accuracy data, the five best frequency norms found in the previous section (CELEX, WorldLex–Blog, BNC, ANC, and COCA) also showed the largest predictive power for both subgroups. *In all, the results suggest that L2 proficiency may not modulate the comparative predictive power of different corpus-based frequency norms on L2 lexical processing*.

## Question 3: How do wordform frequency measures compare with lemma frequency measures in predicting L2 English lexical processing?

So far we have focused on the predictive power of wordform frequency measures on L2 lexical processing. In this section, we compare the lemma frequency measure[7] with the corresponding wordform frequency measure from the same corpus. For this purpose, we examined only six corpora (CELEX, BNC, COCA, ANC, SUBTLEX-UK, and SUBTLEX-US), because only these corpora provided the part-of-speech information that was necessary for the calculation of lemma frequencies. For the RT data, as is shown in Table 5, the lemma frequencies showed much less predictive power than did the wordform frequencies, which was confirmed by Vuong tests (*p*s < .001 for all the corpora). But for the accuracy data, the lemma frequencies seemed to perform better, as indicated by the increase of the McFadden $R^2$

[7] In the present study, we followed Keuleers et al.'s (2010, p. 646) practice by summing up the lemma frequencies of all the possible interpretations of a given wordform. For example, for the lemma frequency of the item "play," we summed up the frequencies of all the inflected forms of the verb "play" (play $_{verb}$, plays $_{verb}$, playing $_{verb}$, played $_{verb}$) and the noun "play" (play $_{noun}$, plays $_{noun}$). For another item "thought," we calculated its lemma frequency by summing up the frequencies of the inflected forms of the verb "think" (think $_{verb}$, thinks $_{verb}$, thinking $_{verb}$, thought $_{verb}$) and the noun "thought" (thought $_{noun}$, thoughts $_{noun}$). It should be noted that we only calculated those inflected forms that shared the same syntactic categories with the given word form. For instance, the item "surprise" has an inflected form "surprised" (the past tense form and past participle), but "surprised" can be an adjective as well. In that case, the frequency of the adjective "surprised" was excluded during the calculation of the lemma frequency for the item "surprise." Besides, we did not take polysemy/homonymy into account during the calculation of lemma frequencies, so we only calculated the total frequencies of all the multiple meanings of an item. The derived lemma frequencies were then transformed into Zipf values.

**Table 3** Predictive power of corpus-based frequency norms on RT data from intermediate and advanced L2 learners ($N$ = 282 items)

| Intermediate L2 Learners | | | Advanced L2 Learners | | |
|---|---|---|---|---|---|
| Corpus | $R^2$ (%) | AIC | Corpus | $R^2$ (%) | AIC |
| **WL-blog** | 71.5 | – 650.843 | **SUBTLEX-US** | 73.4 | – 667.075 |
| **SUBTLEX-US** | 70.8 | – 643.277 | **WL-blog** | 70.1 | – 633.716 |
| **CELEX** | 69.7 | – 633.492 | **WL-twitter** | 69.9 | – 631.912 |
| **WL** | 69.4 | – 630.552 | **WL** | 69.7 | – 630.205 |
| **COCA** | 69.0 | – 627.175 | **COCA** | 69.1 | – 624.912 |
| **WL-twitter** | 68.0 | – 617.575 | **USENET** | 68.4 | – 618.749 |
| **SUBTLEX-UK** | 67.7 | – 615.266 | **CELEX** | 68.3 | – 617.518 |
| **BNC** | 67.6 | – 614.647 | **SUBTLEX-UK** | 68.1 | – 615.273 |
| **HAL** | 67.3 | – 611.561 | **HAL** | 67.4 | – 609.615 |
| **ANC** | 66.7 | – 606.495 | **BNC** | 66.4 | – 601.265 |
| **USENET** | 66.5 | – 604.771 | **ANC** | 66.1 | – 598.261 |
| **WL-news** | 65.9 | – 599.888 | **WL-news** | 65.1 | – 590.347 |
| **GB_AmE** | 65.2 | – 594.201 | **KF** | 64.4 | – 584.780 |
| **KF** | 64.9 | – 591.818 | **GB_AmE** | 63.8 | – 580.373 |
| **GB_BrE** | 63.6 | – 581.884 | **GB_BrE** | 61.9 | – 565.704 |

(1) The frequency norm models are ranked in the order of the goodness of fit, with a larger $R^2$ value indicating a larger proportion of variance explained by the model, and a smaller AIC value indicating a better fit of the model for the data. (2) WL: WorldLex; GB: Google books; (3) only items with an accuracy rate of 66.7% for both low- and high-proficiency L2 English

values. Vuong tests further showed that lemma frequencies had a (marginally) significant edge over wordform frequencies for only three corpora (BNC, $p$ = .028; ANC, $p$ = .099; COCA, $p$ = .053).

To further examine the locus of the difference between the lemma and wordform frequencies, we conducted separate analyses for the stem forms and inflected forms, because one recent study (Gimenes et al., 2016) indicated

**Table 4** Predictive power of corpus-based frequency norms on accuracy data from intermediate and advanced L2 learners ($N$ = 370 items)

| Intermediate L2 Learners | | | Advanced L2 Learners | | |
|---|---|---|---|---|---|
| Corpus | McFadden $R^2$ (%) | AIC | Corpuss | McFadden $R^2$ (%) | AIC |
| **CELEX** | 51.2 | 2,884.744 | **CELEX** | 49.2 | 2,395.525 |
| **WL-blog** | 49.3 | 2,998.194 | **WL-blog** | 47.3 | 2,484.067 |
| **ANC** | 48.9 | 3,016.351 | **COCA** | 46.5 | 2,521.665 |
| **BNC** | 48.7 | 3,028.349 | **ANC** | 46.3 | 2,530.330 |
| **COCA** | 48.3 | 3,052.884 | **BNC** | 46.1 | 2,542.223 |
| **GB_AmE** | 47.6 | 3,094.084 | **WL** | 45.3 | 2,577.854 |
| **WL** | 47.6 | 3,095.207 | **GB_AmE** | 45.3 | 2,579.326 |
| **SUTBLEX-UK** | 46.4 | 3,166.504 | **SUTBLEX-UK** | 44.6 | 2,610.045 |
| **USENET** | 45.7 | 3,208.700 | **USENET** | 43.0 | 2,685.435 |
| **HAL** | 45.7 | 3,209.618 | **KF** | 43.0 | 2,686.636 |
| **KF** | 45.6 | 3,213.257 | **WL-twitter** | 41.9 | 2,734.427 |
| **GB_BrE** | 43.9 | 3,310.068 | **HAL** | 41.8 | 2,741.899 |
| **WL-news** | 43.5 | 3,334.005 | **WL-news** | 41.1 | 2,771.989 |
| **WL-twitter** | 43.4 | 3,341.953 | **GB_BrE** | 41.1 | 2,775.357 |
| **SUBTLEX-US** | 41.1 | 3,475.715 | **SUBTLEX-US** | 39.2 | 2,863.989 |

(1) The frequency norm models are ranked in the order of the goodness of fit, with a larger McFadden $R^2$ value indicating a larger proportion of variance explained by the model, and a smaller AIC value indicating a better fit of the model for the data. (2) WL: WorldLex; GB: Google books

**Table 5** Comparisons of predictive power between wordform and lemma frequencies on L2 RT and accuracy data

| Corpus | RT ($N$ = 273 items) | | | | Accuracy ($N$ = 331 items) | | | |
|---|---|---|---|---|---|---|---|---|
| | Wordform | | Lemma | | Wordform | | Lemma | |
| | $R^2$ (%) | AIC | $R^2$ (%) | AIC | McFadden $R^2$ (%) | AIC | McFadden $R^2$ (%) | AIC |
| **CELEX** | 75.1 | – 637.181 | 58.4 | – 497.191 | 54.0 | 3,860.401 | 54.8 | 3,795.005 |
| **BNC** | 73.0 | – 615.577 | 61.7 | – 519.837 | 50.8 | 4,132.046 | 56.6 | 3,646.462 |
| **ANC** | 71.2 | – 597.476 | 60.8 | – 513.568 | 51.1 | 4,102.549 | 54.6 | 3,813.413 |
| **COCA** | 74.4 | – 629.997 | 62.7 | – 526.910 | 51.0 | 4,116.592 | 56.0 | 3,700.478 |
| **SUB-UK** | 72.7 | – 611.983 | 61.0 | – 515.059 | 49.3 | 4,253.889 | 50.4 | 4,167.497 |
| **SUB-US** | 75.9 | – 646.591 | 57.9 | – 494.238 | 42.8 | 4,800.862 | 44.8 | 4,629.608 |

(1) SUB-: SUBTLEX-.; (2) We only analyzed those items whose lemma frequency was not equal to its wordform frequency (i.e., the lemma of the item should include at least two different wordforms), with 331 items included in the accuracy analyses. As to the RT analysis, we additionally excluded those items with accuracy rates of larger than 66.7%, and thus 273 items remained

that lemma and wordform frequencies might play different roles in the processing of the stem and inflected forms. The results indicated that for the RT data (shown in Table 6), the "inferior" performance of lemma relative to wordform frequencies lay in the inflected forms ($p$s < .05 for all corpora, indicated by Vuong tests), whereas lemma and word frequencies had similar predictive power for the RTs of the stem forms ($p$s > .05 for all corpora). In contrast, for the accuracy data of the stem forms (shown in Table 7), the lemma frequencies (marginally) significantly outperformed the wordform frequencies for most of the corpora (BNC, $p$ = .014; ANC, $p$ = .017; COCA, $p$ = .003; SUBTLEX-UK, $p$ = .033; SUBTLEX-UK, $p$ = .051), but for those of the inflected forms, the lemma and wordform frequencies did not differ significantly ($p$s > .05 for all corpora). *In general, it was found that the lemma frequencies did not outperform the corresponding wordform frequencies in terms of their predictive power for RTs, but an advantage for the lemma frequencies was observed in predicting the accuracy data of stem forms.*

## Question 4: How does the wordform frequency measure compare with the contextual diversity measure in predicting L2 English lexical processing speed?

Most previous L1 research has shown that the contextual diversity measure was a better predictor of lexical processing speed than was the wordform frequency measure (e.g., Adelman et al., 2006; Brysbaert & New, 2009), but the issue has not been investigated in the context of L2. We thus explored how contextual diversity and wordform frequency measures differed in predicting L2 lexical processing speeds. We obtained

**Table 6** Comparisons of predictive power between wordform and lemma frequencies on L2 RT data

| Corpus | Stem Forms ($N$ = 164 items) | | | | Inflected Forms ($N$ = 109 items) | | | |
|---|---|---|---|---|---|---|---|---|
| | Wordform | | Lemma | | Wordform | | Lemma | |
| | $R^2$ (%) | AIC | $R^2$ (%) | AIC | $R^2$ (%) | AIC | $R^2$ (%) | AIC |
| **CELEX** | 74.4 | – 393.937 | 67.7 | – 355.840 | 75.8 | – 250.855 | 65.0 | – 210.812 |
| **BNC** | 70.9 | – 372.934 | 71.9 | – 379.013 | 74.4 | – 244.843 | 65.3 | – 211.700 |
| **ANC** | 67.9 | – 356.839 | 68.3 | – 358.846 | 73.7 | – 241.700 | 66.2 | – 214.422 |
| **COCA** | 71.6 | – 376.733 | 72.1 | – 380.144 | 77.0 | – 256.318 | 69.0 | – 224.045 |
| **SUB-UK** | 69.2 | – 363.840 | 72.3 | – 380.988 | 77.7 | – 259.755 | 65.7 | – 213.019 |
| **SUB-US** | 72.1 | – 380.025 | 71.6 | – 377.799 | 79.2 | – 267.535 | 62.3 | – 202.611 |

SUB-: SUBTLEX-

**Table 7** Comparisons of predictive power between wordform and lemma frequencies on L2 accuracy data

| Corpus | Stem forms ($N$ = 207 items) | | | | Inflected forms ($N$ = 124 items) | | | |
|---|---|---|---|---|---|---|---|---|
| | Wordform | | Lemma | | Wordform | | Lemma | |
| | McFadden $R^2$ (%) | AIC | McFadden $R^2$ (%) | AIC | McFadden $R^2$ (%) | AIC | McFadden $R^2$ (%) | AIC |
| **CELEX** | 62.3 | 2,290.105 | 64.5 | 2,160.020 | 48.2 | 1,196.984 | 41.3 | 1,355.264 |
| **BNC** | 58.2 | 2,539.445 | 62.8 | 2,262.085 | 47.5 | 1,213.206 | 46.5 | 1,237.244 |
| **ANC** | 58.8 | 2,502.827 | 63.0 | 2,247.660 | 45.0 | 1,271.519 | 43.7 | 1,299.474 |
| **COCA** | 57.2 | 2,598.549 | 62.5 | 2,281.223 | 50.1 | 1,154.898 | 46.0 | 1,248.761 |
| **SUB-UK** | 55.9 | 2,673.785 | 59.0 | 2,488.876 | 45.5 | 1,258.215 | 39.1 | 1,406.185 |
| **SUB-US** | 49.6 | 3,056.390 | 51.6 | 2,937.631 | 45.6 | 1,256.161 | 36.2 | 1,470.315 |

SUB-: SUBTLEX-

the log-transformed contextual diversity measures from two subtitle-based corpora, SUBTLEX-UK and SUBTLEX-US, and WorldLex in both its entirety and its three subcorpora (Blog, Twitter, and News). As is shown in Table 8, as compared with the wordform frequency measure from the same corpus, the extra variance explained by most contextual diversity measures was quite small (0.1% to 1% increase in $R^2$). Subsequent Vuong tests showed that only the contextual diversity measure from SUBTLEX-US significantly outperformed its corresponding wordform frequency measure ($p$ = .024). This suggested that *the advantage of the contextual diversity measure over the wordform frequency measure in predicting L2 lexical processing speed is limited.*

**Table 8** Comparisons of predictive power between wordform frequency and contextual diversity measures on L2 RT data ($N$ = 303 items)

| Corpus | Wordform | | Contextual diversity | |
|---|---|---|---|---|
| | $R^2$ (%) | AIC | $R^2$ (%) | AIC |
| **SUB-UK** | 71.7 | – 677.267 | 72.1 | – 681.134 |
| **SUB-US** | 75.3 | – 717.916 | 76.3 | – 730.926 |
| **WL** | 73.2 | – 693.538 | 73.3 | – 694.719 |
| **WL-Blog** | 74.2 | – 704.791 | 74.3 | – 706.401 |
| **WL-Twitter** | 72.2 | – 682.693 | 72.2 | – 682.324 |
| **WL-News** | 68.8 | – 647.593 | 69.0 | – 649.189 |

SUB-: SUBTLEX-.; WL: WorldLex

## General discussion

In the present study, we compared the predictive power of a large set of corpus-based frequency norms on L2 English lexical processing. In general, we found that SUBTLEX-US and WorldLex–Blog better predicted the L2 RT data, whereas frequency norms from the corpora with a mixture of written and spoken genres (CELEX, WorldLex–Blog, BNC, ANC, and COCA) better predicted the L2 accuracy data. These patterns were replicated with the data from learners of different L2 English proficiencies, but were somewhat different from those found in L1 lexical processing. In addition, we observed some limited advantages of the lemma frequency and contextual diversity measures in predicting L2 lexical processing over the wordform frequency measure.

Besides, the present study seemed to demonstrate a more prominent word frequency effect in L2 than in L1 for both RT and accuracy data, consistent with previous findings (e.g., Lemhöfer et al., 2008; van Wijnendaele & Brysbaert, 2002). Unlike in L1 data, we found a lack of floor effect for high-frequency words in our L2 RT data, and a weaker ceiling effect in our L2 accuracy data. These differences between L1 and L2 conform to the lexical entrenchment hypothesis, which suggests that less exposure to a language leads to a stronger word frequency effect (see more details in Brysbaert, Lagrou, & Stevens, 2017; Cop, Drieghe, & Duyck, 2015; Diependaele et al., 2013; Kuperman & van Dyke, 2013; Monaghan et al., 2017).

### Comparison between different corpus-based frequency norms

The most important finding of the present study is that some corpus-based frequency norms outperformed others

in terms of their predictive power on the L2 RT data. The overall good performance of SUBTLEX-US in predicting the L2 and L1 RT data was consistent with previous findings (e.g., Brysbaert & Cortese, 2011; Brysbaert & New, 2009). This implies that our contemporary L2 English learners, like L1 native speakers, may enjoy more frequent exposure to English movies and TV programs. This was confirmed by a questionnaire in our previous pilot study, which showed that many L2 English learners in China frequently watched English movies and soap operas (about four TV episodes and two movies each month on average). The good performance of blog-based frequency norms in predicting the L2 RT data also suggests an increasing influence of Internet-based materials on contemporary L2 English learning for our L2 learners. However, the overall good performance of subtitle-based or web-based frequency norms in predicting the RTs needs to be interpreted with caution, because these frequency norms may be potentially confounded by other lexical factors that specifically contribute to the speed of responses in a lexical decision task, such as the emotionality (e.g., arousal or valence) of a word (Baayen et al., 2016; Heister & Kliegl, 2012). When other tasks (such as normal reading with eye movement recording) were employed, the superiority of subtitle-based frequency norms may disappear, as was shown by Baayen et al. (2016). Therefore, future studies are needed to see whether the findings in this study can be extended to other types of tasks.

In contrast, for our L2 accuracy data, the present study found that frequency norms from the corpora with a mixture of written and spoken genres tended to exhibit better performance. This was consistent with Shi's (2015) study, which also observed a slight advantage of the ANC frequency norm in predicting the error rates of the spoken word recognition task. As accuracy measure here can be regarded as an indication of word prevalence (i.e., how many people in a population know a word; see Brysbaert, Stevens, Mandera, & Keuleers, 2016), our results seemed to suggest that frequency norms from corpora with both written and spoken genres may be a better indicator of how well a word is known among L2 learners.

Moreover, comparison with L1 data revealed unique patterns for our L2 participants. For example, COCA was found to best predict L1 RTs, but this was not found for our L2 speakers. Besides, frequency norms from Web-based and subtitle-based corpora appeared to show better performance in predicting the L1 accuracy data, whereas frequency norms from the corpora with a mixture of written and spoken genres showed better predictions for our L2 accuracy data. The different patterns between L1 and L2 may be explained by the difference in the English input received by our L2 learners and native speakers, which suggests that participants' language experience is important in determining the predictive power of specific corpus-based frequency norms. Moreover, these results indicate that researchers need to be cautious about applying existing L1 corpus-based frequency norms in L2, as the same norms may not have equally good predictive power on both L1 and L2 lexical processing.

In addition, the present study observed similar patterns of the predictive power of corpus-based frequency norms on the data from learners of different L2 proficiencies. This is probably because our L2 participants may have received similar L2 English input, despite their different English proficiencies and lengths of English learning. Future studies are required to examine whether our results can be replicated with the data from other participants receiving different types of L2 input.

It should be noted that the present study observed quite different patterns for the RT and accuracy analyses. For example, in general, SUBTLEX-US was found to make the best predictions for our L2 RT data, but not for our L2 accuracy data. This is consistent with some previous L1 research (e.g., Brysbaert, Buchmeier, et al., 2011; Brysbaert, Keuleers, & New, 2011; Soares et al., 2015) in which subtitle-based frequencies were found to be consistently better at predicting the RT data but not the accuracy data. The observed discrepancy between RT and accuracy analyses may be due to the fact that RT and accuracy are two different kinds of measures and reveal different aspects of lexical processing. It is possible that different aspects of lexical processing are sensitive to different types of corpus-based frequency norms. Future research is needed to clarify this issue.

## Wordform frequency versus lemma frequency and contextual diversity measures

The present study only found a limited advantage of the lemma frequency measure over the wordform frequency measure in predicting L2 lexical processing. Consistent with some previous findings in the L1 (e.g., Brysbaert & New, 2009), the present study failed to find any superiority of lemma frequency over wordform frequency in predicting our L2 RT data. Instead, we found worse performance of lemma frequency in predicting the RTs of inflected forms. These results were different

from the findings of Gimenes et al.'s (2016) L1 study, which observed a stronger effect of lemma frequency on the processing of singular nouns, medium- and low-frequency plural nouns in English than wordform frequency. The discrepancy between these findings may be due to L2 learners' less sensitivity to morphological internal structure and more reliance on whole-word storage during L2 lexical processing (Clahsen, Felser, Neubauer, & Silva, 2010), which was further supported by the less predictive power of lemma frequency on our L2 RT data of the inflected forms. However, we did observe an advantage of lemma frequency of some corpora in predicting the L2 accuracy data, but this advantage was only confined to stem forms.

With respect to the contextual diversity measure, the present study only found an advantage of the contextual diversity measure in predicting RTs for only one corpus (SUBTLEX-US), but the extra contribution of the contextual diversity measure was quite small (an increase of 1% in $R^2$ values as compared to the wordform frequency measure). This is consistent with previous studies (e.g., Adelman et al., 2006; Brysbaert & New, 2009), which showed that the extra variance explained by the contextual diversity measure was small (1%–3%). Besides, aligned with Gimenes and New (2016), our failure to discover an advantage of the contextual diversity measure for other corpora also supports the view of whether the contextual diversity measure outperforms wordform frequencies is corpus-dependent (Baayen et al., 2016).

The results of the present study provide important methodological implications for the selection of corpus-based frequency norms in L2 lexical processing research when word frequency needs to be manipulated or controlled. The RT analysis in the present study indicates that researchers may apply frequency norms of SUBTLEX-US or WorldLex–Blog to control for the word frequency effect on the speed of L2 lexical processing. In contrast, our accuracy analysis suggests that frequency norms from the corpora with a mixture of written and spoken genres (such as CELEX) may be more helpful for researchers to select the appropriate word stimuli for L2 learners. Besides, the limited usefulness of the lemma frequency and contextual diversity measures implies that for most practical purposes, researchers can primarily focus on the wordform frequency measure in L2 lexical research.

There are limitations regarding the generality of the findings in the present study. First, the participants in the present study were Chinese learners of English in Mainland China and caution should be taken when applying the results to situations involving L2 English learners with other L1 language backgrounds or other L2 English learning contexts. For instance, since Chinese and English are different in orthography (i.e., logographic vs. alphabetic) and share only a very limited number of cognates in pronunciation, a similar study conducted on learners of English from other alphabetic language backgrounds (e.g., French) could produce different results. The same is true for other groups of L2 English learners that receive different types of English input due to different pedagogical practices or sociopolitical environments. Thus, more research is needed for L2 learners with different L1 backgrounds or L2 learning contexts. Second, our results were based on the LDT data from a limited set of word items. Due to potential problems in item sampling, it is possible that the patterns of the comparative predictive power of different frequency norms may vary when a different set of items were analyzed, as shown by the additional analyses in Appendix D in which we compared the regression analyses of our L2 data and the British Lexicon Project data (Keuleers et al., 2012, http://crr.ugent.be/blp/) on the basis of only a subset of our stimuli. Thus, further validation of the present results calls for a mega-study approach that collects massive LDT data of thousands of items in the L2 (e.g., Lemhöfer et al., 2008).

## Conclusion

In sum, the present study compared the predictive power of a large set of corpus-based frequency norms on L2 English lexical processing and found that different types of frequency norms had different amounts of predictive power on the RT and accuracy data in L2 lexical processing. Our results thus provide some methodological implications about the choice of corpus-based frequency norms in future L2 lexical research. But proper application of the corpus-based frequency norms in lexical research requires a full understanding of the nature of different frequency norms, which still needs further research.

**Acknowledgements** This research is supported by the National Social Science Foundation of China (15AYY002). The authors thank Emmanuel Keuleers and an anonymous reviewer for their comments on earlier drafts. The authors also thank Jiadan Lin, Fei Li, Fei Zhong and Hongming Zhao for helping recruit the participants and collect the data and James Campion for proofreading the final draft.

# Appendix A

**Table 9.** Basic information on the selected corpora

| Corpus | Corpus Size | | Genre | English Variety | Time |
|---|---|---|---|---|---|
| | Tokens | Types | | | |
| KF | +1 million (1,014,000) | +50 thousand (50,406) | Written texts | AmE | 1967 |
| CELEX | +17.9 million (18,580,121) | +66 thousand (66,372) | Written/ spoken texts | BrE + AmE | 1993/1995 |
| BNC | 100.1 million | +939 thousand (939,028) | Written/ spoken texts | BrE | 1980s–1993 |
| COCA | +520 million | +2 million | Written/ spoken texts | AmE | 1990–2015 |
| ANC | +22 million (22,164,985) | +239 thousand (239,208) | Written/ spoken texts | AmE | 1990–2005 |
| SUBTLEX – UK | 201.3 million | 333 thousand (332,987) | Film and TV subtitles | BrE | 2014 |
| SUBTLEX – US | 51 million | +74 thousand (74,286) | Film and TV subtitles | AmE | 2007/2009 |
| HAL | +131/160/437 million? | 97 thousand (97,261)/+3.4 million(3, 461,884) | Usenet posts | AmE | 1995–1996 |
| USENET | +7 billion (7,781,959,860) | +1.62 million | Usenet posts | AmE | 2005–2006 |
| WorldLex | 104.2 million | +800 thousand (800,072) | Blogs (38.1 million) Twitter (30.9 million) News (35.2 million) | AmE | 2012 |
| GB-BrE | 34 billion | +13 million | Books | BrE | 1810–2009 |
| GB-AmE | 155 billion | +13 million | Books | AmE | 1810–2009 |

BrE = British English, AmE = American English; GB = Google Books

**Retrieval of the corpus-based word frequency measures in the present study:**

**KF:** retrieved from the English Lexicon Project (http://elexicon.wustl.edu/)

**CELEX:** retrieved from WebCelex (http://celex.mpi.nl/)

**BNC:** calculated from Kilgarriff's BNC word frequency list( retrieved from www.kilgarriff.co.uk/bnc-readme.html)

**COCA:** retrieved from http://corpus.byu.edu/coca/

**ANC:** retrieved from www.anc.org/data/anc-second-release/frequency-data/

**SUBTLEX-UK:** retrieved from http://crr.ugent.be/archives/1423

**SUBTLEX-US**: retrieved from http://subtlexus.lexique.org/

**HAL:** retrieved from the English Lexicon Project (http://elexicon.wustl.edu/)

**WestburyLab USENET:** retrieved from WestburyLab (www.psych.ualberta.ca/~westburylab/downloads/wlfreq.download.html)

**WorldLex:** retrieved from http://worldlex.lexique.org/

**Google Books AmE:** retrieved from http://googlebooks.byu.edu/

**Google Books BrE:** retrieved from http://googlebooks.byu.edu/

## Appendix B

**Table 10.** Wordform frequency distribution of selected words used in the LDT

| Corpus | Number of Word Stimuli in Different Frequency Ranges | | | Zipf Mean (*SD*) | Zipf Range |
|---|---|---|---|---|---|
| | Zipf < 4 | 4 ≤ Zipf ≤ 5 | Zipf > 5 | | |
| **KF** | 106 | 163 | 101 | 4.47(0.76) | 2.97-6.24 |
| **CELEX** | 128 | 151 | 91 | 4.39(0.85) | 1.73-6.17 |
| **BNC** | 123 | 140 | 107 | 4.42(0.86) | 1.69-6.21 |
| **COCA** | 118 | 145 | 107 | 4.45(0.84) | 1.60-6.14 |
| **ANC** | 124 | 144 | 102 | 4.40(0.86) | 2.13-6.15 |
| **SUB-UK** | 141 | 135 | 94 | 4.29(0.97) | 1.00-6.36 |
| **SUB-US** | 167 | 122 | 81 | 4.15(0.98) | 1.29-6.60 |
| **HAL** | 124 | 126 | 120 | 4.42(0.93) | 1.06-6.15 |
| **USENET** | 150 | 115 | 105 | 4.27(0.95) | 0.97-6.48 |
| **WorldLex** | 128 | 141 | 101 | 4.37(0.90) | 1.58-6.23 |
| **GB_BrE** | 112 | 164 | 94 | 4.41(0.84) | 1.79-6.30 |
| **GB_AmE** | 139 | 153 | 78 | 4.28(0.83) | 1.76-6.14 |

(1) The word frequency was measured in terms of Zipf values proposed by van Heuven et al. (2014) ; SUB-: SUBTLEX; GB: Google books, BrE : British English; AmE: American English

**Table 11.** Pearson correlation coefficients between different corpus-based wordform frequency measures of the word stimuli in the LDT

| Frequency Measures | KF | CELEX | BNC | COCA | ANC | SUBT-UK | SUBT-US | HAL | USENET | WL | WL–Blog | WL–Twitter | WL–News | GB_BrE | GB_AmE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **KF** | 1.000 | .936 | .933 | .935 | .919 | .876 | .842 | .899 | .896 | .903 | .909 | .839 | .907 | .918 | .938 |
| **CELEX** | | 1.000 | .967 | .954 | .926 | .922 | .887 | .907 | .901 | .919 | .932 | .862 | .909 | .933 | .946 |
| **BNC** | | | 1.000 | .965 | .950 | .919 | .853 | .941 | .934 | .933 | .939 | .863 | .932 | .938 | .951 |
| **COCA** | | | | 1.000 | .969 | .931 | .894 | .951 | .945 | .967 | .963 | .906 | .968 | .899 | .936 |
| **ANC** | | | | | 1.000 | .901 | .863 | .948 | .941 | .943 | .938 | .882 | .946 | .891 | .928 |
| **SUBT-UK** | | | | | | 1.000 | .933 | .883 | .875 | .950 | .950 | .918 | .921 | .828 | .845 |
| **SUBT-US** | | | | | | | 1.000 | .852 | .842 | .921 | .919 | .922 | .868 | .790 | .813 |
| **HAL** | | | | | | | | 1.000 | .969 | .933 | .934 | .888 | .919 | .874 | .912 |
| **USENET** | | | | | | | | | 1.000 | .921 | .918 | .868 | .913 | .870 | .909 |
| **WL** | | | | | | | | | | 1.000 | .987 | .968 | .969 | .845 | .881 |
| **WL-Blog** | | | | | | | | | | | 1.000 | .950 | .943 | .865 | .896 |
| **WL-Twitter** | | | | | | | | | | | | 1.000 | .898 | .771 | .805 |
| **WL-News** | | | | | | | | | | | | | 1.000 | .849 | .888 |
| **GB_BrE** | | | | | | | | | | | | | | 1.000 | .986 |
| **GB_AmE** | | | | | | | | | | | | | | | 1.000 |

(1) WL: WorldLex, SUBT: SUBTLEX, GB: Google Books, BrE: British English, AmE: American English. (2) All correlations were highly significant ($p < .001$). The frequency measures used in the correlation analysis were Zipf values proposed by van Heuven et al. (2014)

**Table 12.** Pairwise model comparisons for L2 RT data ($p$ values of Vuong tests) ($N$ = 303 items)

| | WL–bl | WL | CELEX | COCA | WL–tw | SUB-UK | BNC | USENET | HAL | ANC | WL–n.s. | KF | GB_AmE | GB_BrE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SUB-US** | .232 | .091 | .098 | **.035** | **.027** | **.048** | **.014** | **.020** | **.012** | **.002** | **.001** | **.000** | **.000** | **.000** |
| **WL–bl** | | .079 | .175 | **.023** | .052 | .089 | **.011** | **.014** | **.009** | **.000** | **.000** | **.000** | **.000** | **.000** |
| **WL** | | | .408 | .151 | .151 | .221 | .066 | .063 | **.048** | **.003** | **.000** | **.001** | **.002** | **.000** |
| **CELEX** | | | | .308 | .376 | .304 | .098 | **.018** | .107 | **.012** | **.010** | **.000** | **.000** | **.000** |
| **COCA** | | | | | .493 | .402 | .145 | .130 | .137 | **.004** | **.000** | **.001** | **.001** | **.000** |
| **WL–tw** | | | | | | .410 | .249 | .264 | .211 | .067 | **.021** | **.027** | **.022** | **.004** |
| **SUB-UK** | | | | | | | .359 | .346 | .341 | .174 | .106 | .087 | .073 | **.018** |
| **BNC** | | | | | | | | .505 | .437 | .131 | .098 | .051 | **.019** | **.002** |
| **USNT** | | | | | | | | | .455 | .124 | .093 | **.041** | **.012** | **.001** |
| **HAL** | | | | | | | | | | .156 | .117 | .072 | **.031** | **.004** |
| **ANC** | | | | | | | | | | | .341 | .250 | .119 | **.022** |
| **WL–n.s.** | | | | | | | | | | | | .370 | .269 | .075 |
| **KF** | | | | | | | | | | | | | .332 | .059 |
| **GB_AmE** | | | | | | | | | | | | | | **.003** |

WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$ value in a cell indicates whether the model of corpus-frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus-frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

**Table 13.** Pairwise model comparisons for L1 ELP RT data ($p$ values of Vuong tests) ($N$ = 303 items)

| | SUB-US | CELEX | WL | WL–bl | WL–n.s. | BNC | GB_AmE | ANC | HAL | USENET | WL–tw | SUB-UK | GB_BrE | KF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **COCA** | .423 | .151 | **.046** | **.044** | **.027** | **.025** | .054 | **.007** | **.004** | **.002** | **.021** | **.016** | **.007** | **.003** |
| **SUB-US** | | .313 | .194 | .183 | .202 | .182 | .184 | .137 | .117 | .099 | **.016** | **.038** | **.045** | **.023** |
| **CELEX** | | | .411 | .366 | .348 | .115 | .169 | .180 | .144 | .130 | .145 | .084 | **.008** | **.010** |
| **WL** | | | | .402 | .372 | .325 | .333 | .209 | .160 | .168 | .056 | .110 | .083 | **.043** |
| **WL–bl** | | | | | .457 | .366 | .363 | .263 | .206 | .216 | .102 | .133 | .086 | **.048** |
| **WL–n.s.** | | | | | | .393 | .396 | .273 | .257 | .255 | .197 | .176 | .113 | .063 |
| **BNC** | | | | | | | .470 | .377 | .346 | .336 | .293 | .226 | **.047** | **.047** |
| **GB_AmE** | | | | | | | | .402 | .396 | .389 | .319 | .280 | **.003** | .069 |
| **ANC** | | | | | | | | | .495 | .490 | .353 | .342 | .184 | .174 |
| **HAL** | | | | | | | | | | .494 | .340 | .329 | .165 | .110 |
| **USENET** | | | | | | | | | | | .344 | .325 | .156 | .118 |
| **WL–tw** | | | | | | | | | | | | .479 | .344 | .287 |
| **SUB-UK** | | | | | | | | | | | | | .332 | .275 |
| **GB_BrE** | | | | | | | | | | | | | | .420 |

ELP: English Lexicon Project (Balota et al., 2007); WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$ value in a cell indicates whether the model of corpus frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

**Table 14.** Pairwise model comparisons for L2 accuracy data ($p$ values of Vuong tests) ($N$ = 370 items)

| | WL–bl | ANC | BNC | COCA | GB_AmE | WL | SUB-UK | USENET | KF | HAL | WL–tw | GB_BrE | WL–n.s. | SUB-US |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CELEX** | .231 | .123 | .059 | .073 | **.045** | .061 | **.027** | **.012** | **.006** | **.014** | **.005** | **.001** | **.000** | **.000** |
| **WL–bl** | | .398 | .367 | .316 | .253 | **.027** | .117 | **.040** | .066 | .064 | **.001** | **.036** | **.001** | **.005** |
| **ANC** | | | .454 | .423 | .285 | .249 | .173 | **.032** | .056 | **.045** | **.018** | **.034** | **.004** | **.007** |
| **BNC** | | | | .49 | .348 | .323 | .183 | .097 | .096 | .093 | .053 | **.016** | **.005** | **.012** |
| **COCA** | | | | | .345 | .255 | .161 | **.036** | .082 | .070 | **.022** | **.048** | **.000** | **.005** |
| **GB_AmE** | | | | | | .492 | .377 | .172 | .155 | .158 | .100 | **.007** | .075 | **.024** |
| **WL** | | | | | | | .314 | .159 | .200 | .170 | .006 | .113 | **.001** | **.014** |
| **SUB-UK** | | | | | | | | .333 | .348 | .300 | .146 | .201 | .055 | **.048** |
| **USENET** | | | | | | | | | .497 | .389 | .252 | .276 | .200 | .074 |
| **KF** | | | | | | | | | | .432 | .286 | .267 | .223 | .117 |
| **HAL** | | | | | | | | | | | .335 | .350 | .308 | .127 |
| **WL–tw** | | | | | | | | | | | | .498 | .459 | .188 |
| **GB_BrE** | | | | | | | | | | | | | .472 | .235 |
| **WL–n.s.** | | | | | | | | | | | | | | .225 |

WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$ value in a cell indicates whether the model of corpus-frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus-frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

**Table 15.** Pairwise model comparisons for L1 ELP accuracy data ($p$ values of Vuong tests) ($N$ = 370 items)

| | WL | SUB-US | COCA | WL–bl | HAL | WL–n.s. | USENET | BNC | ANC | SUB-UK | CELEX | GB_AmE | KF | GB_BrE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **WL–tw** | .489 | .427 | .381 | .326 | .262 | .238 | .174 | .099 | .077 | .055 | .067 | .079 | **.037** | **.030** |
| **WL** | | .418 | .291 | .087 | .211 | .076 | .081 | **.031** | **.014** | **.007** | **.017** | **.031** | **.009** | **.009** |
| **SUB-US** | | | .440 | .411 | .316 | .267 | .195 | .093 | .059 | **.041** | **.049** | **.047** | **.018** | **.013** |
| **COCA** | | | | .464 | .283 | .178 | .098 | **.012** | **.005** | **.019** | **.006** | **.018** | **.005** | **.005** |
| **WL–bl** | | | | | .349 | .237 | .194 | .053 | **.025** | **.018** | **.026** | **.042** | **.013** | **.011** |
| **HAL** | | | | | | .472 | .305 | .100 | .085 | .143 | .071 | .073 | **.046** | **.024** |
| **WL–n.s.** | | | | | | | .399 | .098 | **.049** | **.026** | **.041** | .062 | **.016** | **.013** |
| **USENET** | | | | | | | | .132 | .086 | .148 | .083 | .082 | **.042** | **.023** |
| **BNC** | | | | | | | | | .414 | .423 | .183 | .179 | .132 | **.023** |
| **ANC** | | | | | | | | | | .477 | .294 | .218 | .139 | **.031** |
| **SUB-UK** | | | | | | | | | | | .376 | .343 | .178 | .094 |
| **CELEX** | | | | | | | | | | | | .380 | .204 | **.047** |
| **GB_AmE** | | | | | | | | | | | | | .335 | **.002** |
| **KF** | | | | | | | | | | | | | | .203 |

ELP: English Lexicon Project (Balota et al., 2007); WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$ value in a cell indicates whether the model of corpus-frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus-frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

## Appendix C: Stimuli in the lexical decision task

**Word Stimuli:** absolute accrue acid admiral affair again against agreed allowed amazement American analysis anchor angriest announces another arisen ashes assign atmosphere attempt autonomous available avoid away awkward babies balconies bank being best bill bleak boast body bombers box bravest breed broader bump burial bus business buy calmer capable capture cart caught characterize checks children chosen cigar class classified coalition coincide coke combat comes comfy comparison computer conceal conference confession confide confirm conquer consulting contact core corridor corroborate country course covering creation credit crouch cultural curl currently danger day dazzle dealt dearer debut decent declares densest development different diminishes diplomat discovery dishes disliked distance distributed division dominant done down drawn drives driveway dug during easiest east eaten ecstatic edge edifice employer emulation enough entertainment environment episodes excellent excuse experience extracted factories fade fancier fans fantastic far fatter fear feet few fiercest file fire flagrant flirt fold forced frees fuel fullest funniest gathered geese generations gets glasses god gold good gospel government gross guys habit happen having healthier heard heavy hid highest himself honey hose human hybrid implies incredibly induction information installed international island issue its jaunt jewel jingle job keep kind knives knowledge known last later latex launch leagues lifeboat linking listed loathed looking luckier lumber maintains major manhood manor mean meetings mice minister mirror morning mouths move must nastier necessary nerve never nights noisier normal numbers office older opportunity orientation outbreak pamper particle patron peaches pebble perhaps perimeter periodic picking place played pleasure plunging plus possession pound presidential prevent probably problems products proposals prototype provides punched purchases purest quarter quieter reasonably received reef relating relationship requires response responsible reveal revival ribbon richest right rises roast room rough ruler running sadism sat saying second security seems selection sent sentencing separating services session shorter shout silence slam slaughter slow sometimes south square stakes stamp starting state statistics steady stories strain strongest student successor sung surprise surveillance tagging taxes tear telescope territories terror textual theme thieves those thought tile till today together tolerance torches total touchy toxicology tragedies trail transport trouble turned twist type uglier uncover under understand universities unless upbringing visitor vocal vulnerable walk wavy way weapon weekend weird welcome went wide withhold wonder work worn worthiest wound yarn year yes yonder zone

**Nonword Stimuli:** abronote acclub acerican acoden acreed affept agict agische alempment allaved allempt alve amolydus anfliest angiances anscar arnwack assank aste atmosclode atrical atucker avarm awin awow baderitions balfer balture banfitories bave bercoties bertest beskness bidar biest bire bist blachant blawn blees blook blus bobberable bodassary bok brooner brudit buk bule bup burters cacadro cace caing cames capies cear cervently ceses chesent chiggren chofter clace climmified coindore colky coll commifor commoripen compaforate concorm concorsion congial connoiling conrocente contive coom corruner cortler coscat cose couffly counge cowind creach cudy cug cumnaral cusk dacking decheres decure deet deggerent delilortent dencatery denering denflabuted denit diaration dicaliches dimper diprosit dismarked distique dold dorest doy dranchter dreves dronkway duritant duser eash eches emprewer ench eneaux ensoction entersilemant epimirds ertonience esrallent etifect ettritic exfose expirulment exprented farches fas fass feaking feam fearlest femial fethier figging fike fintboat fithing flanter flianer foiced folt foltual fondestic forn fosted foughtier foving frieves fubbories furror gady gead geech gid gork goscal gots gols groke gurbers hadden hagil heen herfest hice himsoft hoke hopal houdy huden hunem hust huxes hyglad idaration idmonorous imichable implelled insercutousal invangation ippries isbard issau jakel jaste jub keed kile knoil knornenth lare laught leathes lemerance lemililogy lenviest limer loached logay luet lurger lutter maglon mahar mainbakes manreed martigra maught meard mecielic memaction memadeter mepanity middor mocacating mone moubenably mough mourds mubstases mucond mussion nelve nerstiest neventment nirths nuder nurmel oaden obtier oleuddation onser ontletably ootiest orleccunity ots outsleek owless paror pervare pilette pinger pizzle plearian plonting poppassion pouse predudes prepacly procolads proridenteer prorting prosies protonill provasts prument pulched rabing radics raister rantenting rebronsible recial recining reclores recoomed reems respaque retoxal revationtrat rew ribble rire rirge rirton ronding rooches rurned russest sankixes scayed seath sering sheat shecks shovest sid sirning slar sleed slirt smives smosses snose soast soatings soorer sorttives spoosure spow squank stade stanks stapesies stass stooty stotittism strail strinkest subseallance sucision suffissor sunt surquire sushes sut swace sype tatew tearty teef telebrain tennier thede thounce tofiller trablems trark tratebort treesle troben tuke tull tung tunkest twire uffile umipersities underchird unseger upshaining urter vangle vopel vurtier wal wans wask wassered wealt weasend weeson weize wenger wesh wethome wime wincier wirthoth wisinar woft wone wouse wromasterize yapper yed zove

## Appendix D: Analyses of the predictive power of different corpus-based frequency norms on our L2 data and the BLP L1 data

We conducted the same regression analyses as those reported in the main text on the RT and accuracy data from our L2 English lexical decision task and from the BLP L1 data (British Lexicon Project, Keuleers et al., 2012). To enable parallel comparisons between the two datasets, we excluded the multisyllabic items in our word stimuli because the BLP data only included mono- and disyllabic items. For the RT analyses, only those items that had an accuracy rate of 66.7% or more were included.

For the remaining L2 RT data, our analyses showed that SUBTLEX-UK provided the best predictions, as indicated by the R-square and AIC values in Table 16 and the *p* values of Vuong tests in Table 18. This was slightly different from what was reported in the main text, and the difference was probably due to item sampling, because the results in this appendix were based on only a subset of our word stimuli (only mono- and disyllabic items). Nevertheless, SUBTLEX-US and WorldLex–Blog were still the second- and the third-best frequency norms, which confirmed the patterns reported in the main text showing the good performance of these two frequency norms when multisyllabic words were included.

For the BLP L1 RT data, Table 16 showed that SUBTLEX-US and WorldLex–Blog had the best predictions. But this ran counter to what van Heuven et al. (2014) had found, that is, SUBLEX-UK predicted the BLP RT data better than SUBLTEX-US. Again, we believe that this gap was due to item sampling, as our results here were based on a rather limited set of items from the BLP, whereas the results from van Heuven et al. (2014) were based on a larger set. And yet, what is worth noticing is that the differences between the predictive powers of these frequency norms on the BLP RT data were rather small, as indicated by the large *p* values of the Vuong tests in Table 19. In particular, word frequency seemed to explain much less variance of the BLP RT data (26%–32%), suggesting that there might be a large floor effect for the BLP RT data of the selected mono- and disyllabic items, and that the ranking of predictive powers of different frequency norms may not be taken very seriously.

With regard to the L2 accuracy data, despite some slight variations, our analyses generally confirmed the patterns reported in the main text and showed that frequency norms from corpora with a mixture of written and spoken genres (such as CELEX, BNC, ANC, and COCA) tended to make better predictions, as indicated by Table 17 and the *p* values of Vuong tests in Table 20.

For the BLP L1 accuracy data, we found different patterns from those in the L2. As is indicated by Table 17 and the Vuong tests in Table 21, the two subtitle-based frequency norms (SUBTLEX-UK and SUBLEX-US) had the best predictions, consistent with what was found in van Heuven et al. (2014).

To sum up, on the basis of a subset of our stimuli, the patterns in L2 English reported in the main text were roughly confirmed, although there were variations probably due to the item sampling of only mono- and disyllabic words.

**Table 16.** Predictive power of corpus-based frequency norms on L2 and L1 BLP RT data ($N$ = 226 mono- and disyllabic items)

| L2 English Speakers | | | Native English Speakers | | |
|---|---|---|---|---|---|
| Corpus | $R^2$ (%) | AIC | Corpus | $R^2$ (%) | AIC |
| **SUBTLEX-UK** | 68.9 | -566.138 | **WL–Blog** | 32.1 | -650.249 |
| **SUBTLEX-US** | 67.7 | -540.402 | **SUBTLEX-US** | 32.0 | -649.717 |
| **WL–Blog** | 67.3 | -531.736 | **WL** | 31.8 | -649.097 |
| **CELEX** | 66.9 | -529.311 | **CELEX** | 30.8 | -646.059 |
| **WL** | 65.6 | -526.553 | **WL–Twitter** | 30.8 | -645.995 |
| **COCA** | 65.3 | -517.709 | **COCA** | 30.8 | -645.763 |
| **BNC** | 65.3 | -515.823 | **SUBTLEX-UK** | 30.8 | -645.743 |
| **WL–Twitter** | 64.4 | -515.727 | **WL–News** | 30.2 | -644.086 |
| **HAL** | 63.1 | -509.865 | **ANC** | 29.8 | -642.576 |
| **USENET** | 62.4 | -501.930 | **BNC** | 29.3 | -641.018 |
| **ANC** | 61.9 | -497.582 | **USENET** | 28.6 | -638.989 |
| **WL–News** | 59.3 | -494.498 | **GB_AmE** | 28.6 | -638.862 |
| **KF** | 59.0 | -479.502 | **HAL** | 28.5 | -638.650 |
| **GB_AmE** | 58.5 | -478.045 | **GB_BrE** | 26.5 | -632.418 |
| **GB_BrE** | 56.8 | -475.264 | **KF** | 26.0 | -630.741 |

(1) The frequency norm models are ranked in the order of the goodness of fit, with a larger $R^2$ value indicating a larger proportion of variance explained by the model, and a smaller AIC value indicating a better fit of the model for the data. (2) WL: WorldLex; GB: Google books

**Table 17.** Predictive power of corpus-based frequency norms on L2 and L1 BLP accuracy data ($N$ = 272 mono- and disyllabic items)

| L2 English Speakers | | | Native English Speakers | | |
|---|---|---|---|---|---|
| Corpus | McFadden $R^2$ (%) | AIC | Corpus | McFadden $R^2$ (%) | AIC |
| **CELEX** | 58.7 | 1872.131 | **SUBTLEX-UK** | 31.5 | 937.656 |
| **COCA** | 56.5 | 2772.561 | **SUBTLEX-US** | 29.6 | 963.293 |
| **ANC** | 55.9 | 2918.432 | **WL–Blog** | 29.3 | 967.003 |
| **BNC** | 55.8 | 2960.992 | **WL** | 27.3 | 994.126 |
| **SUBTLEX-UK** | 55.4 | 2964.224 | **WL–Twitter** | 25.1 | 1021.450 |
| **WL–Blog** | 55.1 | 2993.498 | **COCA** | 24.4 | 1033.659 |
| **WL** | 54.1 | 3010.663 | **BNC** | 23.8 | 1041.810 |
| **GB_AmE** | 53.2 | 3076.624 | **CELEX** | 23.5 | 1046.190 |
| **USENET** | 52.0 | 3137.111 | **USENET** | 23.2 | 1049.330 |
| **HAL** | 51.4 | 3220.120 | **ANC** | 23.1 | 1050.394 |
| **WL–News** | 51.3 | 3263.532 | **WL–News** | 22.0 | 1066.261 |
| **SUBTLEX-US** | 50.9 | 3264.344 | **HAL** | 22.0 | 1066.518 |
| **GB_BrE** | 48.9 | 3296.499 | **KF** | 18.7 | 1111.094 |
| **WL–Twitter** | 48.8 | 3429.720 | **GB_AmE** | 16.0 | 1147.893 |
| **KF** | 48.2 | 3432.715 | **GB_BrE** | 14.9 | 1162.691 |

(1) The frequency norm models are ranked in the order of the goodness of fit, with a larger McFadden $R^2$ value indicating a larger proportion of variance explained by the model, and a smaller AIC value indicating a better fit of the model for the data. (2) WL: WorldLex; GB: Google books

**Table 18.** Pairwise model comparisons for L2 RT data ($p$ values of Vuong tests) ($N$ = 226 mono- and disyllabic items)

| | SUB-US | WL-bl | CELEX | WL | COCA | BNC | WL-tw | HAL | USENET | ANC | WL-n.s. | KF | GB_AmE | GB_BrE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SUB-UK** | .109 | .072 | **.023** | **.018** | **.004** | **.003** | **.003** | **.001** | **.001** | **.000** | **.000** | **.000** | **.000** | **.000** |
| **SUB-US** | | .263 | .170 | .185 | **.028** | **.021** | **.027** | **.026** | **.009** | **.002** | **.000** | **.000** | **.000** | **.000** |
| **WL–bl** | | | .438 | .393 | .186 | .170 | .215 | .072 | .062 | **.036** | **.020** | **.003** | **.002** | **.004** |
| **CELEX** | | | | .425 | .055 | .067 | .156 | .055 | **.030** | **.011** | **.002** | **.000** | **.000** | **.000** |
| **WL** | | | | | .289 | .202 | .135 | .190 | .076 | **.030** | **.014** | **.003** | **.000** | **.000** |
| **COCA** | | | | | | .407 | .442 | .195 | .138 | .079 | **.024** | **.000** | **.002** | **.007** |
| **BNC** | | | | | | | .497 | .339 | .156 | .070 | **.013** | **.000** | **.000** | **.002** |
| **WL–tw** | | | | | | | | .375 | .153 | .070 | **.029** | **.005** | **.001** | **.000** |
| **HAL** | | | | | | | | | .325 | .243 | .173 | **.022** | **.032** | **.044** |
| **USENET** | | | | | | | | | | .308 | .272 | .062 | **.048** | **.039** |
| **ANC** | | | | | | | | | | | .395 | .097 | .065 | .056 |
| **WL–n.s.** | | | | | | | | | | | | .093 | .087 | .067 |
| **KF** | | | | | | | | | | | | | .458 | .397 |
| **GB_AmE** | | | | | | | | | | | | | | .405 |

WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$ value in a cell indicates whether the model of corpus-frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus-frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

**Table 19.** Pairwise model comparisons for L1 BLP RT data ($p$ values of Vuong tests) ($N$ = 226 mono- and disyllabic items)

| | SUB-US | WL | CELEX | WL–tw | COCA | SUB-UK | WL–n.s. | ANC | BNC | USENET | GB_AmE | HAL | GB_BrE | KF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **WL–bl** | .473 | .360 | .254 | .297 | .155 | .302 | .123 | .091 | .077 | **.045** | .065 | .057 | **.030** | **.038** |
| **SUB-US** | | .465 | .317 | .266 | .293 | .288 | .270 | .197 | .173 | .111 | .184 | .098 | .088 | .051 |
| **WL** | | | .318 | .309 | .217 | .339 | .144 | .138 | .109 | .062 | .127 | .079 | .054 | **.047** |
| **CELEX** | | | | .497 | .479 | .486 | .401 | .304 | .122 | .147 | .155 | .165 | **.042** | .061 |
| **WL–tw** | | | | | .488 | .488 | .417 | .347 | .301 | .204 | .280 | .179 | .139 | .080 |
| **COCA** | | | | | | .499 | .337 | .198 | .196 | .112 | .186 | .106 | .074 | .052 |
| **SUB-UK** | | | | | | | .425 | .343 | .295 | .194 | .286 | .151 | .143 | .051 |
| **WL–n.s.** | | | | | | | | .383 | .330 | .238 | .283 | .237 | .138 | .100 |
| **ANC** | | | | | | | | | .390 | .238 | .324 | .209 | .143 | .091 |
| **BNC** | | | | | | | | | | .347 | .377 | .358 | .137 | .136 |
| **USENET** | | | | | | | | | | | .494 | .469 | .228 | .156 |
| **GB_AmE** | | | | | | | | | | | | .491 | .068 | .236 |
| **HAL** | | | | | | | | | | | | | .252 | .135 |
| **GB_BrE** | | | | | | | | | | | | | | .437 |

BLP: British Lexicon Project (Keuleers et al., 2012); WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$-value in a cell indicates whether the model of corpus-frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus-frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

**Table 20.** Pairwise model comparisons for L2 accuracy data ($p$ values of Vuong tests) ($N$ = 272 mono- and disyllabic items)

| | COCA | ANC | BNC | SUB-UK | WL–blog | WL | GB_AmE | USENET | HAL | WL–n.s. | SUB-US | GB_BrE | WL–tw | KF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CELEX** | .201 | .145 | .136 | .087 | .164 | .078 | **.026** | **.025** | **.021** | **.008** | **.002** | **.001** | **.005** | **.001** |
| **COCA** | | .352 | .401 | .322 | .270 | .095 | .092 | **.024** | **.039** | **.003** | **.027** | **.016** | **.006** | **.001** |
| **ANC** | | | .494 | .419 | .391 | .215 | .164 | .051 | **.044** | **.028** | **.036** | **.030** | **.008** | **.003** |
| **BNC** | | | | .428 | .415 | .282 | .210 | .130 | .118 | **.035** | .084 | **.005** | **.042** | **.009** |
| **SUB-UK** | | | | | .461 | .281 | .259 | .147 | .136 | **.039** | **.036** | **.031** | **.010** | **.012** |
| **WL–bl** | | | | | | .193 | .275 | .136 | .154 | .064 | .113 | .058 | **.005** | **.024** |
| **WL** | | | | | | | .385 | .223 | .206 | .055 | .148 | .088 | **.003** | **.034** |
| **GB_AmE** | | | | | | | | .313 | .282 | .283 | .219 | **.028** | .127 | **.042** |
| **USENET** | | | | | | | | | .387 | .420 | .364 | .194 | .177 | .092 |
| **HAL** | | | | | | | | | | .499 | .444 | .274 | .246 | .184 |
| **WL–n.s.** | | | | | | | | | | | .444 | .253 | .203 | .153 |
| **SUB-US** | | | | | | | | | | | | .291 | .252 | .225 |
| **GB_BrE** | | | | | | | | | | | | | .496 | .434 |
| **WL–tw** | | | | | | | | | | | | | | .441 |

WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$ value in a cell indicates whether the model of corpus-frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus-frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

**Table 21.** Pairwise model comparisons for L1 BLP accuracy data ($p$ values of Vuong tests) ($N$ = 272 mono- and disyllabic items)

| | SUB-US | WL-bl | WL | WL-tw | COCA | BNC | CELEX | USENET | ANC | WL-n.s. | HAL | KF | GB_AmE | GB_BrE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SUB-UK** | .292 | .224 | .135 | .072 | **.029** | **.042** | **.044** | **.048** | **.042** | **.023** | **.042** | **.007** | **.005** | **.004** |
| **SUB-US** | | .462 | .231 | .078 | .086 | .112 | .113 | **.047** | .090 | **.042** | .051 | **.009** | **.008** | **.006** |
| **WL–bl** | | | .099 | **.039** | **.038** | .075 | .085 | **.037** | .054 | **.019** | **.033** | **.008** | **.006** | **.005** |
| **WL** | | | | .107 | .122 | .167 | .169 | .065 | .110 | **.019** | .050 | **.018** | **.008** | **.006** |
| **WL–tw** | | | | | .408 | .373 | .360 | .244 | .304 | .170 | .172 | .080 | **.037** | **.029** |
| **COCA** | | | | | | .399 | .362 | .336 | .280 | .145 | .190 | .060 | **.004** | **.004** |
| **BNC** | | | | | | | .450 | .423 | .389 | .285 | .242 | .146 | **.011** | **.007** |
| **CELEX** | | | | | | | | .476 | .454 | .307 | .344 | .091 | **.003** | **.001** |
| **USENET** | | | | | | | | | .488 | .327 | .176 | .157 | **.031** | **.026** |
| **ANC** | | | | | | | | | | .319 | .304 | .152 | **.006** | **.005** |
| **WL–n.s.** | | | | | | | | | | | .497 | .151 | **.017** | **.010** |
| **HAL** | | | | | | | | | | | | .244 | **.043** | **.035** |
| **KF** | | | | | | | | | | | | | .238 | .173 |
| **GB_AmE** | | | | | | | | | | | | | | .096 |

BLP: British Lexicon Project (Keuleers et al., 2012); WL: WorldLex; SUB-: SUBTLEX-; bl: Blog; Tw: Twitter; n.s.: News; GB: Google books. The models in the columns and rows are listed in the order of goodness of fit. The $p$-value in a cell indicates whether the model of corpus-frequency measure represented by the row of the cell provided a significantly better fit for the data than the model of corpus-frequency measure represented by the column of the cell. Significant difference between models ($p < .05$) is marked in bold

## References


Adams, M. J. (1979). Models of word recognition. *Cognitive Psychology*, *11*, 133–176. doi:https://doi.org/10.1016/0010-0285(79)90008-2

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. doi:https://doi.org/10.1111/j.1467-9280.2006.01787.x

Adorni, R., Manfredi, M., & Mado Proverbio, A. (2013). Since when or how often? Dissociating the roles of age of acquisition (AoA) and lexical frequency in early visual word processing. *Brain and Language*, *124*, 132–141. doi:https://doi.org/10.1016/j.bandl.2012.11.005

Akbari, N. (2015). Word frequency and morphological family size effects on the accuracy and speed of lexical access in school-aged bilingual students. *International Journal of Applied Linguistics* doi:https://doi.org/10.1111/ijal.12113

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press. doi:https://doi.org/10.1558/sols.v2i3.471

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *Mental Lexicon*, *5*, 436–461. doi:https://doi.org/10.1075/ml.5.3.10baa

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290–313. doi:https://doi.org/10.1016/j.jml.2006.03.008

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*, 1174–1220. doi:https://doi.org/10.1080/02687038.2016.1147767

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia: Linguistics Data Consortium, University of Pennsylvania.

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*, 419–463. doi:https://doi.org/10.1075/ml.2.3.06baa

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 and Performance*, *10*, 340–357. doi:https://doi.org/10.1037/0096-1523.10.3.340

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation lexical access and/or production? *Journal of Memory and Language*, *24*, 89–106. doi:https://doi.org/10.1016/0749-596X(85)90017-8

Balota, D. A., & Chumbley, J. I. (1990). Where are the effects of frequency in visual word recognition tasks? Right where we said they were! Comment on Monsell, Doyle, and Haggard (1989). *Journal of Experimental Psychology: General*, *119*, 231–237. doi:https://doi.org/10.1037/0096-3445.119.2.231

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316. doi:https://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A, Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. doi:https://doi.org/10.3758/BF03193014

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America*, *106*, 2074–2085. doi:https://doi.org/10.1121/1.427952

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates

in German. *Experimental Psychology*, *58*, 412–424. doi:https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology*, *64*, 545–559. doi:https://doi.org/10.1080/17470218.2010.503374

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, *45*, 422–430. doi:https://doi.org/10.3758/s13428-012-0270-5

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 27:1–8. doi:https://doi.org/10.3389/fpsyg.2011.00027

Brysbaert, M., Lagrou, E., & Stevens, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, *20*, 530–548. doi:https://doi.org/10.1017/S1366728916000353

Brysbaert, M., Mandera, P., & Keuleers, E. (2017). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*. Advance online publication. doi:https://doi.org/10.1177/0963721417727521

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. doi:https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*, 991–997. doi:https://doi.org/10.3758/s13428-012-0190-4

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 441–458. doi:https://doi.org/10.1037/xhp0000159

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. doi:https://doi.org/10.3758/BF03200655

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*, e10729. doi:https://doi.org/10.1371/journal.pone.0010729

Clahsen, H., Felser, C., Neubauer, K., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning*, *60*, 21–43. doi:https://doi.org/10.1111/j.1467-9922.2009.00550.x

Colombo, L., Pasini, M., & Balota, D. a. (2006). Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. *Memory & Cognition*, *34*, 1312–1324. doi:https://doi.org/10.3758/BF03193274

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1084–1096. doi:https://doi.org/10.1037/0278-7393.16.6.1084

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PLoS ONE*, e134008:1–38. doi:https://doi.org/10.1371/journal.pone.0134008

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, *22*, 1216–1234. doi:https://doi.org/10.3758/s13423-015-0819-2

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, *32*, 133–143. doi:https://doi.org/10.1371/journal.pone.0010729

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, *42*, 317–367. doi:https://doi.org/10.1006/cogp.2001.0750

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+). Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159–190. doi:https://doi.org/10.1075/ijcl.14.2.02dav

Davies, M. (2011a). Google Books (American English) Corpus (155 billion words, 1810–2009). Retrieved from googlebooks.byu.edu/

Davies, M. (2011b). Google Books (British English) Corpus (34 billion words, 1810–2009).

De Groot, A. M. B., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language*, *47*, 91–124. doi:https://doi.org/10.1006/jmla.2001.2840

R Development Core Team. (2017). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*, 843–863. doi:https://doi.org/10.1080/17470218.2012.720994

Dimitropoulou, M., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: the case of Greek. *Frontiers in Psychology*, 218:1–12. doi:https://doi.org/10.3389/fpsyg.2010.00218

Dong, Y., & Yuan, Y. (2008). The necessity of collecting baseline reaction time in priming experiments. Xinli Kexue (Psychological Science), *31*, 192–194.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, M. A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*, 1246–1258. doi:https://doi.org/10.3758/s13428-013-0326-1

Dufour, S., Brunellière, A., & Frauenfelder, U. H. (2013). Tracking the time course of word-frequency effects in auditory word recognition with event-related potentials. *Cognitive Science*, *37*, 489–507. doi:https://doi.org/10.1111/cogs.12015

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188. doi:https://doi.org/10.1017/S0272263102002024

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635. doi:https://doi.org/10.1016/S0022-5371(73)80042-8

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language*, *45*, 468–492. doi:https://doi.org/10.1006/jmla.2000.2784

Geranpayeh, A. (2003). A quick review of the English Quick Placement Test. *Research Notes*, *12*, 8–10.

Gimenes, M., Brysbaert, M., & New, B. (2016). The processing of singular and plural nouns in English, French, and Dutch: New insights from megastudies. *Canadian Journal of Experimental Psychology*, *70*, 316–324. doi:https://doi.org/10.1037/cep0000074

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*, 963–972. doi:https://doi.org/10.3758/s13428-015-0621-0

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and*

*Language*, *58*, 787–814. doi:https://doi.org/10.1016/j.jml.2007.07.001

Gollan, T. H., Slattery, T. J., Goldenberg, D., van Assche, E., Duyck, W., & Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *Journal of Experimental Psychology: General*, *140*, 186–209. doi:https://doi.org/10.1037/a0022256

Heister, J., & Kliegl, R. (2012). Comparing word frequencies from different German text corpora. In K.-M. Würzner & E. Pohl (Eds.), Lexical resources in psycholinguistic research (pp. 27–44). Potsdam: Universitätsverlag Potsdam.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*, *41*, 976–995. doi:https://doi.org/10.1111/cogs.12392

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, *41*, 401–410. doi:https://doi.org/10.1037/h0056020

Imai, S., Walley, A. C., & Flege, J. E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *The Journal of the Acoustical Society of America*, *117*, 896–907. doi:https://doi.org/10.1121/1.1823291

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, and Cognition*, *20*, 824–843. doi:https://doi.org/10.1037/0278-7393.20.4.824

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, *132*, EL74–80. doi:https://doi.org/10.1121/1.4731641

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*, 627–633. doi:https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: a new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*, 643–650. doi:https://doi.org/10.3758/BRM.42.3.643

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287–304. doi:https://doi.org/10.3758/s13428-011-0118-4

Kilgarriff, A. (2006). BNC data base and word frequency lists. Retrieved from www.kilgarriff.co.uk/bnc-readme.html

Kučera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence: Brown University Press. doi:https://doi.org/10.2307/302397

Kuperman, V., & van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 802–823. doi:https://doi.org/10.1037/a0030859

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 12–31. doi:https://doi.org/10.1037/0278-7393.34.1.12

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1–36. doi:https://doi.org/10.1097/MPG.0b013e3181a15ae8

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208. doi:https://doi.org/10.3758/BF03204766

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, *47*, 471–483. doi:https://doi.org/10.3758/s13428-014-0489-4

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–323. doi:https://doi.org/10.1177/00238309010440030101

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods*, *21*, 151–163. doi:https://doi.org/10.1037/met0000038

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182. doi:https://doi.org/10.1126/science.1199644

Monaghan, P., Chang, Y., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency , language exposure , and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21. doi:https://doi.org/10.1016/j.jml.2016.08.003

Monsell, S., Doyle, M. C., & Haggard, M. P. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*, 43–71. doi:https://doi.org/10.1037/0096-3445.118.1.43

Moulin, A., & Richard, C. (2015). Lexical influences on spoken spondaic word recognition in hearing-impaired patients. *Frontiers in Neuroscience*, 476:1–14. doi:https://doi.org/10.3389/fnins.2015.00476

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*, 661–677. doi:https://doi.org/10.1017/S014271640707035X

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report*, *10*, 357–376.

Pham, H. (2014). *Visual processing of vietnamese compound words: A multivariate analysis of using corpus linguistic and psycholinguistic paradigms (Unpublished PhD dissertation)*.University of Alberta, Edmonton.

Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*, 191–201. doi:https://doi.org/10.3758/BF03197692

Reppen, R., & Ide, N. (2004). The American National Corpus: Overall goals and the first release. *Journal of English Linguistics*, *32*, 105–113. doi:https://doi.org/10.1177/0075424204264856

Savin, H. B. (1963). Word frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, *35*, 200–206. doi:https://doi.org/10.1121/1.1918432

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Memory & Cognition*, *26*, 1270–1281. doi:https://doi.org/10.3758/BF03201199

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, 137:1–16. doi:https://doi.org/10.3389/fpsyg.2014.00137

Schneider, W., Eschman, A., & Zuccolotto, A. (2001). E-prime. Pittsburgh: Psychology Software Tools, Inc.

Shaoul, C., & Westbury, C. (2006). USENET Orthographic frequencies for 1,618,598 types. (2005–2006). Edmonton: University of Alberta. Retrieved from www.psych.ualberta.ca/~westburylab/downloads/wlallfreq.download.html

Shatzman, K. B., & Schiller, N. O. (2004). The word frequency effect in picture naming: Contrasting two hypotheses using homonym pictures. *Brain and Language*, *90*, 160–169. doi:https://doi.org/10.1016/S0093-934X(03)00429-2

Shi, L. (2014). Lexical effects on recognition of the NU-6 words by monolingual and bilingual listeners. *International Journal of Audiology*, *53*, 318–325. doi:https://doi.org/10.3109/14992027.2013.876109

Shi, L. (2015). English word frequency and recognition in bilinguals: Inter-corpus comparison and error analysis. *International Journal of Audiology*, *54*, 674–681. doi:https://doi.org/10.3109/14992027.2015.1030509

Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J. J., … Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology*, *68*, 680–696. doi:https://doi.org/10.1080/17470218.2014.964271

Univeristy of Cambridge Local Examination Syndicate. (2001). Quick Placement Test. Oxford: Oxford University Press.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176–1190. doi:https://doi.org/10.1080/17470218.2013.850521

van Wijnendaele, I., & Brysbaert, M. (2002). Visual word recognition in bilinguals: Phonological priming from the second to the first language. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 616–627. doi:https://doi.org/10.1037/0096-1523.28.3.616

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228–243. doi:https://doi.org/10.1037/a0027127

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333. doi:https://doi.org/10.2307/1912557

Whitford, V., & Titone, D. (2012). Second-language experience modulates first- and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review*, *19*, 73–80. doi:https://doi.org/10.3758/s13423-011-0179-5

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*, 502–529. doi:https://doi.org/10.1016/j.jml.2009.02.001

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*, 1–29. doi:https://doi.org/10.1006/jmla.2001.2834