Behav Res (2012) 44:991–997
DOI 10.3758/s13428-012-0190-4

Adding part-of-speech information to the SUBTLEX-US word frequencies

Marc Brysbaert · Boris New · Emmanuel Keuleers

Published online: 7 March 2012

Abstract The SUBTLEX-US corpus has been parsed with the CLAWS tagger, so that researchers have information about the possible word classes (parts-of-speech, or PoSs) of the entries. Five new columns have been added to the SUBTLEX-US word frequency list: the dominant (most frequent) PoS for the entry, the frequency of the dominant PoS, the frequency of the dominant PoS relative to the entry's total frequency, all PoSs observed for the entry, and the respective frequencies of these PoSs. Because the current definition of lemma frequency does not seem to provide word recognition researchers with useful information (as illustrated by a comparison of the lemma frequencies and the word form frequencies from the Corpus of Contemporary American English), we have not provided a column with this variable. Instead, we hope that the full list of PoS frequencies will help researchers to collectively determine which combination of frequencies is the most informative.

Keywords SUBTLEX · Word frequency · Part-of-speech information · Subtitles · Lexical decision

Whereas throughout most of the twentieth century, collecting a corpus of texts and tagging it with part-of-speech (PoS) information required a massive investment in time and manpower, nowadays it can be done in a matter of days on the basis of digital archives and automatic parsing algorithms. As a result, researchers in psycholinguistics are becoming more aware of quality differences between word frequency measures (Balota et al., 2007; Brysbaert, Buchmeier, et al., 2011; Brysbaert & New, 2009). The use of an appropriate word frequency measure for research was demonstrated by comparing the widely used Kučera and Francis (1967) frequency counts to the best available frequency measure, which explained 10% more variance in naming and lexical decision times of English words. For all languages for which these data are available, word frequency estimates based on a corpus of some 30 million words from film and television subtitles turn out to be the best available predictor of lexical decision and naming times (Brysbaert, Buchmeier, et al., 2011; Brysbaert, Keuleers, & New, 2011; Cai & Brysbaert, 2010; Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011; Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010; New, Brysbaert, Veronis, & Pallier, 2007).

A second way to improve the available word frequency measures would be to add PoS information, or information about the word classes of the entries. Having information on the number of times that a word is observed in a representative corpus is essential, but at the same time is limited in many respects. For a start, researchers are often interested in a particular type of words (e.g., nouns, verbs, or adjectives). This is the case, for instance, when eye movement researchers want to insert words in carrier sentences. In these cases, all words must be of the same syntactic class, and selection is much more efficient if such information is included in the master list from which the words are selected. The same is true for researchers investigating the cortical regions involved in the processing of different types of words, such as nouns or verbs (e.g., Pulvermüller, 1999; Yang, Tan, & Li, 2011). They, too, would prefer to have syntactic information from the outset, so that they can select on this variable, rather than having to clean lists manually after the initial selection.

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

B. New
Université René Descartes,
Paris, France

Also, when researchers present words in isolation, it is a good idea to match the various conditions on syntactic class. Otherwise, syntactic class could turn out to be a confounding variable. For instance, many very-high-frequency words are syntactic function words (articles, determiners, or prepositions). They differ in important aspects from content words, because they come from a limited set (new function words in a language are extremely rare) and are often used in many different constructions (which is the reason of their high frequency). Therefore, researchers would not like to see these words unequally distributed across conditions.

A related concern is that systematic differences may exist between different types of content words. For instance, Baayen, Feldman, and Schreuder (2006) reported faster lexical decisions to monosyllabic verbs than to monosyllabic nouns. This again suggests that researchers may want to match their words on this variable, even though Sereno and Jongman (1997, Exp. 1) reported exactly the opposite finding (i.e., longer lexical decision times for verbs—both mono- and disyllabic—than for nouns).

Finally, many English words are classified under several different PoSs. For instance, the entries "play" and "plays" may be either nouns or verbs. The same is true for "playing," which in addition can be an adjective. Having access to word frequencies that are disambiguated for PoS would allow researchers not only to better select their stimuli in this respect, but also to do research on the topic. For instance, Baayen et al. (2006) found faster lexical decision times to verbs that were also frequently used as nouns.

Syntactic ambiguities are a particular problem when they involve an inflected form and a lemma form, as is the case for many past and present participles of verbs. Researchers probably would not be inclined to include such words as "played" and "playing" in a list of base words (e.g., for a word-rating study) because these words are inflected forms of the verb "to play." However, the same is intuitively not true for "appalled" and "appalling." These words seem to be adjectives in the first place. Again, rather than having to rely entirely on intuition, it would be good also to have information about the relative PoS frequencies of these words.

Below, we first report how PoS information was obtained for the SUBTLEX-US word frequencies and then present some initial analyses.

Method

The SUBTLEX-US corpus is based on subtitles from films and television programs and contains 51 million word tokens coming from 8,388 different subtitle files (Brysbaert & New, 2009). To extract PoS information, we used the CLAWS ("constituent likelihood automatic word-tagging system") algorithm. This algorithm is a PoS tagger developed at Lancaster University (available at http://ucrel.lancs.ac.uk/claws/). We chose this tagger because it is one of the few developed by a team of computational linguists over a prolonged period of time and is optimized for word frequency research in both written and spoken language. CLAWS was the PoS tagger used and improved in creating the British National Corpus, a major research effort to collect a representative corpus of 100 million words and make this corpus available in tagged form (Garside, 1996). It is also the tagger used in an equivalent American initiative to make tagged spoken and written language available to researchers (the Corpus of Contemporary American English: Davies, 2008; see also below).

Even though major research efforts have been invested in the CLAWS tagger, it is important to realize that its output is not completely error-free (just like the outputs of its alternatives). Performance checks have indicated that it achieves 96%–97% overall accuracy, or 98.5% accuracy if judgments are limited to the major grammatical categories (Garside, Leech, & McEnery, 1997; see also the more detailed information available at http://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm). Therefore, users must be aware that, although most of the time the CLAWS gives accurate information, it is better to consider the output as useful guidelines rather than as a set of dictionary definitions (below, we will describe a few examples of errors that we spotted). As far as we know, at present there are no better alternatives to CLAWS (even human taggers disagree about the correct interpretation on some 2% of instances, and the costs that such an effort would involve would be prohibitive).

The CLAWS algorithm parses sentences and assigns the most likely syntactic roles to the words in six steps (Garside, 1996):

1. First, the input text is read in and divided into individual tokens, and sentence breaks are established.
2. A list of possible grammatical tags is assigned to the words on the basis of a lexicon.
3. For the words in the text not found in the lexicon, a sequence of rules is applied to assign a list of suitable tags.
4. Libraries of template patterns are used to adapt the list of word tags from Steps 2 and 3 in light of the immediate context in which each word occurs (e.g., "the play" vs. "I play").
5. The probability of each potential sequence of tags is calculated (as an index of how grammatically well-formed the sentence would be), and the sequence with the highest probability is selected.
6. The input text and the associated information about the tags are returned.

The algorithm uses a set of over 160 tags, which we reduced to the following main syntactic categories: noun, verb, adjective, adverb, pronoun, article, preposition,

conjunction, determiner, number, letter, name (or proper noun), interjection, and unclassified. For each word in the SUBTLEX-US frequency list, we calculated five values:

- The syntactic category with the highest frequency
- The frequency of this category
- The relative frequency of the dominant category
- Other categories assigned to the word
- The frequencies of these categories

Output

Table 1 shows the outcome of the PoS-tagging process for some entries related to "appal(l)" and "play." It illustrates the ways in which words are used in different roles with different frequencies. For instance, "playing" is used most often as a verb (observed 7,340 times in the corpus), but also as an adjective (101 times) and a noun (67 times). Examples from the corpus are "I was playing [V] with it first!," "I mean, if somehow we could level the playing [A] field, then, um, maybe I could find a way to come back," and "The only person my playing [N] is bothering is you." Table 1 also clearly shows that "appalled" and "appalling" are predominantly used as adjectives, whereas "played" and "playing" are predominantly used as inflected verb forms.

When reading the figures in the table, it is good to keep in mind that a small number of entries should be considered incorrect, as indicated above. This becomes clear when we look at the results of a very-high-frequency word such as "a." This entry is not only classified as an article (943,636 times) and a letter (7 times), but also as an adverb (30,910), a noun (257), a preposition (50), an adjective (2), and unclassified (743). The high number of assignments as an adverb comes from situations in which the article precedes a sequence of adjectives, as in the sentences "It feels a [Adv] little familiar." and "I left it in a [Adv] little longer than I should've." The wrong assignments of "a" as an adjective come from the sentences "it would be good to start thinking the differences between the a [A] posteriori truths . . ." and "Yale preppies reuniting their stupid a [A] capella group."

Whereas assignment errors lead to easily recognizable noise for high-frequency words, they may result in misclassifications for low-frequency words. One of the most conspicuous examples we found in this respect is the word "horsefly," which occurred 5 times in the corpus and was consistently tagged as an adverb instead of as a noun, presumably because the word is not present in the CLAWS lexicon and the end letters *-ly* are interpreted as evidence for an adverbial role. Therefore, researchers using small sets of low-frequency words are advised to always manually check their stimuli to make sure that they are not working with materials that are manifestly parsed in the wrong way (as with "horsefly").

Attentive readers will further notice that the frequency counts of the CLAWS algorithm do not always fully agree with those of SUBTLEX-US. This is because the CLAWS

Table 1 Processed outcome of the CLAWS algorithm for some words related to "appal(l)" and "play"

Word	Dom_PoS	Freq_dPoS	Rel_Freq	All_PoS	All_Freq
appal	Verb	2	1.00	Verb	2
appalled	Adjective	49	.83	Adjective; Verb	49; 10
appalling	Adjective	99	1.00	Adjective	99
appallingly	Adverb	3	1.00	Adverb	3
appalls	Verb	1	1.00	Verb	1
appals	Verb	2	1.00	Verb	2
play	Verb	14,646	.81	Verb; Noun; Name	14,646; 3,417; 1
playable	Adjective	3	1.00	Adjective	3
playact	Noun	1	1.00	Noun	1
playbook	Noun	45	1.00	Noun	45
playbooks	Noun	2	1.00	Noun	2
playboy	Noun	169	.78	Noun; Name	169; 47
playboys	Noun	48	.94	Noun; Name	48; 3
played	Verb	2,843	.99	Verb; Adjective	2,843; 26
player	Noun	1,926	1.00	Noun	1,926
players	Noun	872	1.00	Noun; Verb	872; 1
playful	Adjective	59	1.00	Adjective	59
playfully	Adverb	7	.88	Adverb; Name	7; 1
playing	Verb	7,340	.98	Verb; Adjective; Noun	7,340; 101; 67
plays	Verb	1,163	.77	Verb; Noun	1,163; 356

The respective columns contain (1) the word, (2) the most frequent part of speech, (3) the frequency of the dominant part of speech (PoS), (4) the relative frequency of the dominant PoS versus the total frequency as calculated by CLAWS, (5) all PoSs taken by the word, in decreasing order, and (6) the respective frequencies of the PoSs. Frequencies are based on the SUBTLEX-US corpus, which includes 51 million words

algorithm does more than merely count the letter strings: It imposes some structure on the input. This becomes clear when we look at the SUBTLEX-US entries not observed in the CLAWS output. These are such entries as "gonna," "gotta," "wanna," "cannot," "gimme," "dunno," "isn," and "hes." The algorithm automatically corrects these entries and gives them their proper, full-length transcription. The alterations are small and mainly involve high-frequency words, so that for practical purposes they do not matter (i.e., they do not affect the correlation with RTs in typical word-processing tasks). Because the word form frequencies seem to be most important, at present we advise users to keep using the SUBTLEX-US frequencies, which are based on simply counting letter strings. The CLAWS total frequencies are used to calculate the relative frequencies of the dominant PoSs.

We prefer the format of Table 1 over the more frequently used format in which words are given separate lines for each PoS. It is our experience that the latter organization makes the search for good stimuli in psycholinguistic research harder. As we will argue later, word form frequency is the most important variable for psycholinguistic research, and therefore, it is good to have this frequency for a word as a single entry. PoS-related information is secondary, and this is communicated best by putting it on a single line.

Application: Verbs versus nouns

As a first application, we examined whether response times (RTs) to verbs and nouns differ, as had been suggested by Sereno and Jongman (1997) and Baayen et al. (2006), but with opposite results. To this end, we selected the entries from SUBTLEX that only took noun and verb PoS tags and that were recognized by at least two thirds of the participants in the lexical decision experiment of the Elexicon Project. In this project, lexical decision times and naming times were gathered for over 40,000 English words (Balota et al., 2007). The majority of the entries selected were used only as nouns (Table 2). The second most frequent category comprised entries that predominantly served as nouns, but in addition acted as verbs. Then followed the entries only used as verbs, and the verbs also used as nouns.

As can be seen in Table 2, the entries serving both as nouns and verbs were responded to faster than the entries serving as a noun or a verb only [$F(3, 16909) = 488$, $MSE = 11{,}221$]. However, the various categories also differed on a series of confounding variables. Therefore, we examined how much of the differences could be predicted on the basis of the SUBTLEX-US word form frequencies (nonlinear regression using cubic splines), word length in number of letters (nonlinear regression using cubic splines), word length in number of phonemes, orthographic Levenshtein distance to the 20 closest words, and phonological Levenshtein distance to the 20 closest words (see Balota et al., 2007, for more information on these variables). All variables had a significant effect, and together they accounted for 54% of the variance in RTs. They also accounted for most of the differences observed between the four categories, as can be seen in the RTpred column of Table 2. Still, the residual scores of the categories differed significantly from each other [$F(3, 16909) = 22.9$, $MSE = 5{,}543$], mainly due to the fact that the entries primarily used as nouns were processed faster than predicted on the basis of the confounding variables, whereas the entries primarily used as verbs were processed more slowly than predicted. This is in line with the findings of Sereno and Jongman (1997) and different from those of Baayen et al. (2006), possibly because an analysis limited to monosyllabic words does not generalize to the full corpus. The difference between nouns and verbs illustrates, however, that researchers should match their stimuli on PoS information in addition to word form frequency, word length, and similarity to other words.

Table 2 Lexical decision response times (RTs) from the Elexicon Project for verbs and nouns according to the CLAWS part-of-speech information (only entries that were known to two thirds of the participants)

	N	RT (*SD*)	RTpred (*SD*)	Residual
Noun	9,443	774 (113.4)	775 (81.2)	–1 (78.5)
Verb	2,189	767 (99.7)	761 (68.7)	6 (73.9)
Noun + Verb	3,788	691 (89.3)	700 (62.2)	–9 (62.2)
Verb + Noun	1,493	706 (94.2)	701 (77.2)	5 (65.3)

The Noun row indicates all instances of the entry in the corpus that were classified as nouns; Verb indicates all instances of the entry that were classified as verbs; for Noun + Verb, the majority of instances were classified as nouns, the remainder as verbs; for Verb + Noun, most of the instances were classified as verbs, the remainder as nouns

Does lemma frequency, as currently defined, add much to the prediction of lexical decision times?

Historically, researchers have added PoS information to word frequencies because they believed that a combined frequency measure based on the different word forms belonging to the same PoS category would be informative. Francis and Kučera (1982) were the first to do so. In 1967, they had published a word frequency list on the basis of the Brown corpus, without information about the word classes (Kučera & Francis, 1967). In 1982, they added PoS information and used the notion of lemma frequency. A lemma was defined as "a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling" (see also Knowles & Don, 2004). In this case, lemma frequency was the summed

frequency of a base word and all its inflections. For instance, the lemma frequency of the verb "to play" is the sum of the frequencies of the verb forms "play," "plays," "played," and "playing." Similarly, the lemma frequency of the noun "play" is the sum of the frequencies of the noun forms "play" and "plays." Lemma frequencies gained further attention because of their inclusion in the CELEX lexical database (Baayen, Piepenbrock, & van Rijn, 1993).

Using the CELEX frequencies, Baayen, Dijkstra, and Schreuder (1997) published evidence that lemma frequency may be more informative than word form frequency. They showed that Dutch singular nouns with high-frequency plurals (such as the equivalent of English "cloud") were processed faster than matched singular nouns with low-frequency plurals (such as the equivalent of "thumb"). This seemed to indicate that not the word form frequency of the singular noun, but the combined frequency of the singular and plural forms (i.e., the lemma frequency), was important. This conclusion was put in question for English, however, when Sereno and Jongman (1997) examined the same issue and argued that for English, the frequency of the word form was more important than the lemma frequency. Possibly as a result of this finding, American researchers kept on using the Kučera and Francis (1967) word form frequencies rather than the 1982 lemma frequencies, even though New, Brysbaert, Segui, Ferrand, and Rastle (2004) published results for English closer to those of Baayen et al. (1997) than of Sereno and Jongman.

Brysbaert and New (2009) addressed the usefulness of word form frequency versus lemma frequency in a more general way by making use of the word-processing times of the English Lexicon Project (Balota et al., 2007). They observed that, across the 40,000 words, the CELEX word form frequencies accounted for slightly more variance in the RTs than did the CELEX lemma frequencies, and they thus advised researchers to continue working with word form frequencies rather than lemma frequencies. Similar conclusions were reached for Dutch (Keuleers, Brysbaert, & New, 2010) and German (Brysbaert, Buchmeier, et al., 2011).

To further assess the usefulness of lemma frequencies versus word form frequencies for general psycholinguistic research, we turned to a new, independent source of information. In recent years, Davies has compiled a Corpus of Contemporary American English (e.g., Davies, 2008; available at www.wordfrequency.info/). This corpus is based on five different sources with equal weight: transcriptions of TV and radio talk shows, fiction (short stories, books, and movie scripts), popular magazines, newspapers, and academic journals. It is regularly updated, and at the time of purchase (fall 2011) it contained 425 million words. Frequencies can be downloaded or purchased for word forms (depending on the level of detail wanted) and purchased for lemmas; these norms are known as the *COCA word frequencies*.

We used the lemma frequency list provided by COCA and added the word form frequencies from COCA and SUBTLEX-US. Frequencies of homographs were summed. Thus, the lemma frequency of the word "play" was the sum of the lemma frequencies of "play" as a verb (197,153 counts) and "play" as a noun (43,818 counts). Similarly, the COCA word form frequency of the word "play" was the sum of the frequencies of the word "play" classified as a verb (78,621), a noun (36,201), an adjective (36), a name (9), and a pronoun (5). For the SUBTLEX-US word form frequency, we simply took the number of times that the letter sequence "play" had been counted in Brysbaert and New (2009). We correlated the various frequencies with the standardized lexical decision times and the accuracy levels of the English Lexicon Project (Balota et al., 2007) and the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012). Some basic cleaning was done to get rid of questionable entries. Only entries accepted by the Microsoft Office spell checker (American spellings) were included. This excluded most names (which are not accepted if they do not start with a capital) and British spellings. All in all, the analysis based on the English Lexicon Project included 26,073 words; the analysis based on the British Lexicon Project comprised 14,765 words. Entries not observed in the SUBTLEX-US lists were given a frequency of 0. The analyses were based on log(frequency count + 1) and consisted of nonlinear regressions (as in Brysbaert & New, 2009).

As can be seen in Table 3, for the COCA frequencies we replicated the finding that lemma frequencies in general are not more informative than word form frequencies for typical psycholinguistic research, such as matching words in lexical decision experiments. This is surprising, given the results of Baayen et al. (1997) and New et al. (2004). Some further scrutiny suggests why the lemma frequencies, as currently defined, perform as they do. The main differences between lemma frequencies and word form frequencies have to do with such words as "playing." In the COCA lemma frequencies, in line with the linguistic definition, the counts are

Table 3 Percentages of variance accounted for by the COCA lemma frequencies and the word form frequencies in lexical decision performance in the Elexicon Project and the British Lexicon Project

	Elexicon Project		British Lexicon Project	
	zRT	Acc	zRT	Acc
COCA lemma	36.2	19.7	40.8	28.6
COCA word form	43.0	27.5	47.8	40.9
SUBTLEX word form	48.1	22.6	47.6	39.2

Nonlinear regression analysis on entries accepted by the Microsoft Office spell checker (American English, 2007 Version). All values are statistically significant (N = 26,073 for the Elexicon Project, and N = 14,765 for the British Lexicon Project). zRT, response time z score; Acc, percentage accuracy

Table 4 Percentages of variance accounted for by the various language registers included in the COCA corpus, based on lemma frequencies

	Elexicon Project		British Lexicon Project	
	zRT	Acc	zRT	Acc
COCA lemma total	36.2	19.7	40.8	28.6
COCA (lemma spoken)	34.2	21.1	40.4	31.1
COCA (lemma fiction)	41.3	14.8	39.7	26.1
COCA (lemma magazines)	37.0	18.6	38.8	26.2
COCA (lemma newspapers)	35.9	19.0	38.4	27.9
COCA (lemma academic)	20.5	14.8	31.9	20.3

limited to those of the noun "playing" (in both singular and plural forms) and the adjective "playing," for a total of 2,686 counts. In contrast, the frequency of the word form "playing" does not include the plural noun "playings," but it does include the verb form "playing," giving a total of 53,512 counts. A similar situation occurs for the word "played" (COCA lemma frequency of 306 vs. word form frequency of 50,724). Because the verb forms "playing" and "played" are added to the verb lemma "play," the lemma frequency of this word (240,971) is much higher than the word form frequency (114,872). Also worth mentioning is the fact that the word "plays" does not figure in the COCA lemma list, because it is part of either the verb lemma "play" or the noun lemma "play."

It is clear that the contributions of base words and inflected forms require further scrutiny. On the one hand, good evidence exists that the frequencies of inflected forms affect the recognition of base words in at least one case (Baayen et al., 1997; New et al., 2004). On the other hand, it is also clear that lemma frequencies as currently defined are, in general, not very helpful for selecting the stimuli for word recognition experiments (Table 3). One way to improve the situation may be to try out different definitions of lemma frequency and see which one best predicts lexical decision times for various types of words (and in different languages). Another approach may be to use other measures of inflectional and morphological complexity, as proposed by Martín, Kostić, and Baayen (2004). However, it is clear that the issue is unlikely to be settled in a single study such as this one. Therefore, we felt that including a single lemma frequency in our database would send the wrong signal. It seemed more in line with current knowledge to limit the PoS information to the various frequencies provided by the CLAWS algorithm, so that researchers can collectively sink their teeth into the issue and try out different combinations of word frequencies. Hopefully, over time, convergent evidence will emerge about which equivalent to lemma frequency (if any) provides the best information for word recognition research. This could then be added to the SUBTLEX-US database.

Of further interest in Table 3 is the finding that the COCA frequencies, despite being based on a larger and more diverse corpus, do not predict word-processing times better than the SUBTLEX-US frequencies do (although they are better at predicting which words are known). This once again illustrates the importance of the language register. Further evidence is obtained when we look at the performance of the various frequency sources used in COCA (Table 4). Unfortunately, we only have this information for lemma frequencies, but it still shows that, in particular, word frequencies based on academic journals tend to predict the least amount of variance.

Attentive readers may wonder why the COCA spoken frequencies are not equivalent to the SUBTLEX-US frequencies, given that they are both based on transcriptions of spoken materials. To answer this question, it is important to keep in mind that the language registers of the two corpora differ. In the COCA corpus, the spoken sources are talk shows on radio and television, whereas in the SUBTLEX corpus, they are subtitles from films and television series, which typically refer to social interactions. This difference can clearly be shown by looking at the frequencies of the words "I," "you," and "the." In a recent Internet discussion about the most frequent word in English (held on the Corpora List and available at www.hit.uib.no/corpora/), it became clear that the relative frequencies of these three words differ systematically between corpora. Whereas the word "the" is the most frequent in all corpora that include descriptions, "I" and "you" tend to be more prevalent in corpora centered on social interactions, such as SUBTLEX-US (and some of Shakespeare's plays). Table 5 lists the frequencies of the three words in SUBTLEX-US and the various COCA subcorpora. As can be seen, the "I"/"the"

Table 5 Relative frequencies of the words "the," "I," and "you" in various language registers

The more social the language register, the more frequently the pronouns "you" and "I" appear. The more descriptive the register, the more frequently the article "the" appears

Source	"the"	"I"	"you"	"I"/"the"	"you"/"the"
COCA (spoken)	4,190,341	1,623,705	1,472,529	0.39	0.35
COCA (fiction)	4,534,433	1,576,303	880,007	0.35	0.19
COCA (magazines)	4,878,925	648,344	517,144	0.13	0.11
COCA (newspapers)	4,648,992	506,030	271,095	0.11	0.06
COCA (academic)	5,549,547	204,916	79,063	0.04	0.01
SUBTLEX (films)	1,501,908	2,038,529	2,134,713	1.36	1.42

and "you"/"the" ratios decrease the less socially oriented that a source is, and (critically) also differ between the SUBTLEX-US corpus and the COCA spoken corpus.

Summary and availability

We parsed the SUBTLEX-US corpus with the CLAWS tagger so that we could provide information about the syntactic roles of the words. This will allow researchers to better match their stimulus materials or to select words belonging to specific syntactic categories. Unlike previous lists, we have not included lemma frequencies, because they do not yet seem to provide useful information for word recognition researchers.

All in all, we have added five columns to the Excel file containing the SUBTLEX-US word frequencies (Brysbaert & New, 2009). These are (see also Table 1):

1. The dominant PoS of the word according to the CLAWS output of the SUBTLEX-US corpus.
2. The frequency of the dominant PoS (on a total of 51 million words).
3. The percentage of the dominant PoS relative to the total frequency count of the word according to CLAWS (this will allow researchers, for instance, to select stimuli for which the dominant PoS constitutes more than 90% of all observed instances).
4. All PoS roles assigned to the word, in decreasing order of frequency.
5. All frequencies of the PoS roles. Together, these constitute the total frequency of the word according to the CLAWS algorithm.

The augmented SUBTLEX-US file (containing 74,286 entries) is available as supplementary materials for this article and can also be downloaded from http://expsy.ugent.be/subtlexus/.

References

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37,* 94–117. doi:10.1006/jmla.1997.2509

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:10.1016/j.jml.2006.03.008

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

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:10.3758/BF03193014

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

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology, 2,* 27.

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:10.3758/BRM.41.4.977

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

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica, 32,* 133–143.

Davies, M. (2008). The corpus of contemporary American English: 425 million words, 1990–present (Database). Available at http://corpus.byu.edu/coca/

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in Language Sciences, 1,* 218. doi:10.3389/fpsyg.2010.00218

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.

Garside, R. (1996). The robust tagging of unrestricted text: The BNC experience. In J. Thomas & M. Short (Eds.), *Using corpora for language research: Studies in honour of Geoffrey Leech* (pp. 167–180). London: Longman.

Garside, R., Leech, G., & McEnery, A. (Eds.). (1997). *Corpus annotation*. London: Longman.

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: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.

Knowles, G., & Don, Z. M. (2004). The notion of a lemma: Headwords, roots and lexical sets. *International Journal of Corpus Linguistics, 9,* 69–81.

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

Martín, F. M. D., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94,* 1–18. doi:10.1016/j.cognition.2003.10.015

New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language, 51,* 568–585. doi:10.1016/j.jml.2004.06.010

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied PsychoLinguistics, 28,* 661–677.

Pulvermüller, F. (1999). Words in the brain's language. *The Behavioral and Brain Sciences, 22,* 253–279.

Sereno, J. A., & Jongman, A. (1997). Processing of English inflectional morphology. *Memory & Cognition, 25,* 425–437. doi:10.3758/BF03201119

Yang, J., Tan, L. H., & Li, P. (2011). Lexical representation of nouns and verbs in the late bilingual brain. *Journal of Neurolinguistics, 24,* 674–682.