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

ADDING PART OF SPEECH INFORMATION TO THE SUBTLEX-US WORD FREQUENCIES

Marc Brysbaert [1], Boris New [2], Emmanuel Keuleers [1]

[1] Ghent University, Ghent Belgium

[2] Université René Descartes, Paris, France

Address: Marc Brysbaert
Department of Experimental Psychology
Ghent University
Henri Dunantlaan 2
B-9000 Gent
Belgium
Tel. +32 9 264 94 25
Fax. +32 9 264 64 96
marc.brysbaert@ugent.be

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; PoS) of the entries. Five new columns are added to the SUBTLEX-US word frequency list: the dominant (most frequent) PoS for the entry, the frequency of the dominant PoS, the relative frequency of the dominant PoS to the entry's total frequency, all PoS observed for the entry, and the frequencies of these PoS. 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 do not provide 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.

Whereas in most of the twentieth century, collecting a corpus of texts and tagging it with Part of Speech 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 in word frequency measures (Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007; Brysbaert & New, 2009; Brysbaert, Buchmeier, Conrad, Jacobs, Bolte, & Bohl, 2011a). The use of an appropriate word frequency measure for research was demonstrated by comparing the widely used Kucera 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 et al., 2011a; Brysbaert, Keuleers, New, 2011b; Cai & Brysbaert, 2010; Cuetos, Glez-Nosti, Barbon, & Brysbaert, 2011; Dimitropoulou, Dunabeitia, Aviles, Corral, & Carreiras, 2010; New, Brysbaert, Veronis, & Pallier, 2007).

A second way to improve the available word frequencies measures is to add Part-of-Speech (PoS) information, or information about the word classes of the entries. Having information on the number of times a word is observed in a representative corpus is essential but at the same time limited in many respects. For a start, researchers are often interested in a particular type of words (e.g., nouns, verbs, adjectives). This is the case, for instance, when eye movement researchers want to insert words in carrier sentences. Then, all words must be of the same syntactic class and selection is much more efficient if such information is included in the master list to select from. The same is true for researcher investigating the cortical

regions involved in the processing of different types of words, such as nouns or verbs (e.g., Pulvermuller, 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.

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, 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 often are 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 there may be systematic differences 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, Experiment 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 classify under several Parts of Speech. For instance, the entries "play" and "plays" may be a noun or a verb. The same is true for "playing", which in addition can be an adjective. Having access to word frequencies that are disambiguated for Part of Speech allows researchers not only to better select their stimuli in this respect, but also to do

research on the topic. 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 words like “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 algorithm (CLAWS = Constituent Likelihood Automatic Word-tagging System). This algorithm is a part-of-speech tagger developed at Lancaster University (available at http://ucrel.lancs.ac.uk/claws/). We chose the tagger because it is one of the few developed by a team of computational linguists over a prolonged period of time and optimized for word frequency research both in written and spoken language. CLAWS was the PoS-tagger used and improved for the British National Corpus, a major research effort to collect a representative corpus of 100M words and make it 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 outcome is not completely error-free (just like that of its alternatives). Performance checks indicate that it achieves 96-97% overall accuracy and 98.5% accuracy if judgment is limited to the major grammatical categories (Garside, Leech, & McEnery, 1997; see also the more detailed information available on the website: 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 a set of dictionary definitions (below we will see a few examples of errors we spotted). As far as we know, at present there are no better alternatives (even humans disagree about the correct interpretation on some 2% of the instances; not to mention the prohibitive costs such an effort would involve).

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

1. First, the input running text is read in, 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 the 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 CLAWS 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:

- The syntactic category with the highest frequency
- The frequency of this category
- The relative frequency of the dominant category
- The 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 way 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.

Table 1: Processed outcome of the CLAWS algorithm for some words related to “appal(l)” and “play”. The respective columns contain: (1) the word, (2) the most frequent Part of Speech, (3) the frequency of the dominant 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 frequencies of the PoSs. Frequencies are based on the SUBTLEX-US corpus, which includes 51M words.

appal	Verb	2	1.00 Verb	2
appalled	Adjective	49	0.83 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	14646	0.81 Verb;Noun;Name	14646;3417;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	0.78 Noun;Name	169;47
playboys	Noun	48	0.94 Noun;Name	48;3
played	Verb	2843	0.99 Verb;Adjective	2843;26
player	Noun	1926	1.00 Noun	1926
players	Noun	872	1.00 Noun;Verb	872;1
playful	Adjective	59	1.00 Adjective	59
playfully	Adverb	7	0.88 Adverb;Name	7;1
playing	Verb	7340	0.98 Verb;Adjective;Noun	7340;101;67
plays	Verb	1163	0.77 Verb;Noun	1163;356

When reading the figures, 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 like “a”. This entry is not only classified as an article (943,636 times) or 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

uses 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 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 algorithm does more than merely counting 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 entries like 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 the 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 PoS.

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 it as a single entry. PoS-related information is secondary and this is communicated best by putting it on a single line.

Application: verbs vs. nouns

As a first application, we can examine whether response times to verbs and nouns differ, as 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 the latter project, lexical decision times and naming times were gathered for over forty thousand 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 noun or verb only ($F(3, 16909) = 488$, $MSe = 11221$). 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 (non-linear regression using cubic splines), word length in number of letters (non-linear regression using cubic splines), word length in number of phonemes, the orthographic Levenshtein distance to the 20 closest words, and the 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 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 third column of Table 2. Still, the residual scores of the categories differed significantly from each other ($F(3, 16909) = 22.9$, $MSe = 5543$), 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 Sereno and Jongman (1997) and different from 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 RTs from the Elexicon Project for verbs and nouns according to the CLAWS PoS-information (only entries that were known to two thirds of the participants). Noun = all instances of the entry in the corpus were classified as nouns; Verb = all instances of the entry were classified as verbs; Noun+Verb: the majority of instances were classified as noun, the remainder as verb; Verb+Noun: most of the instances were classified as verb, the remainder as noun.

	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)

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

Historically, researchers 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 Kucera (1982) were the first to do so. In 1967 they published a word frequency list on the basis of the Brown corpus without information about the word classes (Kucera & 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 is 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, & 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 "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 was important, but the combined frequency of the singular and the plural noun (i.e., the lemma frequency). This conclusion was questioned 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 Kucera 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 Sereno and Jongman (1997).

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

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 http://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, movie scripts), popular magazines, newspapers, and academic journals. It is regularly updated and at the time of purchase (fall 2011) contained 425 million words. Frequencies can be downloaded or purchased for word forms (depending on the level of detail wanted) and purchased for lemmas; they are known as the COCA word frequencies.

We used the lemma frequency list as provided by COCA and added the word form frequencies from COCA and SUBTLEX-US. Frequencies of homographs were summed. So, the lemma frequency of the word "play" was the sum of the lemma frequency 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 the letter sequence "play" had been counted in Brysbaert & 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, Brysbaert, & Rastle, in press). Some basic cleaning was done to get rid of questionable entries. Only entries accepted by the MS Office Spell Checker American spelling 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. Analyses were based on log(frequency count+1) and consisted of non-linear regression (as in Brysbaert & New, 2009).

Table 3 : Percentage 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. Non-linear regression analysis on entries accepted by the MS Office Spell Checker (American English, Version 2007). All values are statistically significant (N = 26,073 for the Elexicon Project, and N = 14,765 for 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

As can be seen in Table 3, for the COCA frequencies we replicate 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 words like “playing”. In the COCA lemma frequencies, in line with the linguistic definition, the counts are limited to those of the noun “playing” (both in singular and plural form) 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 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 that the word "plays" does not figure in the COCA lemma list, because it is either part of 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 there is good evidence that frequencies of inflected forms affect 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 in general are not very helpful to select 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 Martin, Kostic, and Baayen (2004). However, it is clear that the issue is unlikely to be settled in a single paper like this one. Therefore, we feel that we would send the wrong signal by including a single lemma frequency in our database. It seems more in line with current knowledge to limit the PoS information to the various frequencies as 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 convergence will emerge about which equivalent to lemma frequency (if any) provides the best information for word recognition research. This can then be added to the SUBTLEX-US database.

Further interesting in Table 3 is 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 (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 the lemma frequencies, but it still shows that in particular word frequencies based on academic journals tend to predict the least amount of variance.

Table 4 : Percentage 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

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 http://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 including 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 and you/the ratios decrease the less socially oriented the source is, and (critically) also differs between the SUBTLEX-US corpus and the COCA spoken corpus.

Table 5: Relative frequencies of the words "the", "I", and "you" in various language registers. The more social the language register, the more frequent the pronouns "you" and "I". The more descriptive the language, the more frequent the article "the".

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

Summary and availability

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

All in all, we have added 5 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 51M words).
3. The percentage of the dominant PoS relative to the total frequency count of the word according to CLAWS (this allows 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 material to this article and can also be downloaded from the website: 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.

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.

Baayen, R.H., Piepenbrock, R., & Rijn, H. van (1993). *The CELEX Lexical Database* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., & Böhl, A. (2011a). 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. (2011b). 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.

Cai, Q. & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies ased on film subtitles. *PLOS ONE, 5*, e10729.

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

Davies, M. (2008) *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

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

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

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

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

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (in press). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*.

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

Martin, F.M.D., Kostic, A., & Baayen, R.H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94*, 1-18.

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.

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

Pulvermuller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences, 22*, 253+.

Sereno, J.A., & Jongman, A. (1997). Processing of English inflectional morphology. *Memory & Cognition, 25*, 425-437.

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.