

SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles

EMMANUEL KEULEERS AND MARC BRYLSBAERT
Ghent University, Ghent, Belgium

AND

BORIS NEW
Université Paris Descartes and CNRS UMR 8189, Paris, France

We present a new database of Dutch word frequencies based on film and television subtitles, and we validate it with a lexical decision study involving 14,000 monosyllabic and disyllabic Dutch words. The new SUBTLEX frequencies explain up to 10% more variance in accuracies and reaction times (RTs) of the lexical decision task than the existing CELEX word frequency norms, which are based largely on edited texts. As is the case for English, an accessibility measure based on contextual diversity explains more of the variance in accuracy and RT than does the raw frequency of occurrence counts. The database is freely available for research purposes and may be downloaded from the authors' university site at <http://crr.ugent.be/subtlex-nl> or from <http://brm.psychonomic-journals.org/content/supplemental>.

One of the most important predictors of word processing times is the frequency with which words have been encountered. In large-scale studies, word frequency (WF) reliably explains the largest percentage of variance of any predictor of word processing times (e.g., Baayen, Feldman, & Schreuder, 2006; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Yap & Balota, 2009). Therefore, psycholinguists have invested time in the collection of WF measures. The first list of word frequencies widely used in language research was published in English by Thorndike and Lorge (1944; see Bontrager, 1991, for a review of older frequency lists including German ones). Its main motivation was educational (helping teachers decide which words should be taught to pupils). A few decades later, Kučera and Francis (1967; KF) published a list (also for American English) that would become the frequency measure of choice for language researchers up to the present (Brylsbaert & New, 2009).

For the Dutch language, van Berckel, Brandt Corstius, Mokken, and van Wijngaarden (1965) collected word frequencies based on a newspaper corpus of about 50,000 words. Although this list contained additional statistical information, such as ngram sequences up to three letters, about the Dutch language, it did not gain wide adoption. The first publicly available frequency list for Dutch was edited by Uit den Boogaart (1975), who published frequencies of “written and spoken Dutch” based on a corpus of 605,733 words from written sources and 121,569 words from spoken sources. This book was superseded in 1993, when the Centre for Lexical Information (CELEX)

published frequencies based on a 42-million-word corpus of written texts collected by the Institute for Dutch Lexicology (Baayen, Piepenbrock, & van Rijn, 1993). In addition to the frequencies of the different forms (e.g., *play*, *plays*), the CELEX database also contained the frequencies of the words as different parts of speech (*play* as a noun vs. *play* as a verb) and the frequencies of the headwords or lemmas (e.g., the frequency of the nominal lemma *play* consisting of the summed frequency of the word form *play* as a noun and the word form *plays* as a noun). Since its publication, CELEX has been the primary source of word frequencies and other lexical information for the Dutch language.¹

For a long time, face validity was the main factor in assessing the quality of a frequency measure for research in word recognition. Two criteria were of importance: the representativeness of the sources and the size of the corpus. On both criteria, CELEX scored well. Special care had been taken to select texts from a wide variety of documents produced by the Dutch-speaking community, and the size of the corpus was larger than what was available in most other languages. However, in the past 2 years, researchers have started to measure the validity of word frequencies for research into word recognition processes by correlating them with word processing times for thousands of words. This research has revealed considerable quality differences between existing frequency measures that all score well on the face-validity criteria. Next, we summarize these developments before we return to the Dutch language.

E. Keuleers, emmanuel.keuleers@ugent.be



Edited Texts May Not Be the Best Source of Information for Word Frequencies

When researchers started comparing the correlations among different word-frequency measures, lexical decision times, and word-naming times, they discovered that the much used KF norms were not performing as well as other, less popular frequency measures (Balota et al., 2004; Brylsbaert & New, 2009; Burgess & Livesay, 1998; Zevin & Seidenberg, 2002). For instance, Balota et al. (2004, Figure 7) observed that KF explained only 26% of the variance in the lexical decision times of student participants, which was 9% less than the best frequency measure tested.

A first source that yielded better frequency measures was the Internet. It is much easier to obtain a large corpus from the Internet than from published texts (which sometimes have to be scanned). In addition, word use on the Internet is more varied than the formal language used in edited texts. Burgess and Livesay (1998) showed that a frequency measure (called HAL) based on a few hundred million words taken from Internet discussion groups accounted for more variance in lexical decision times than the KF frequencies. A similar finding was reported by Balota and colleagues (e.g., Balota et al., 2004), who subsequently recommended the HAL frequencies for further research (e.g., Balota et al., 2007). More recent Internet-based frequency measures are based on even larger corpora that contain up to 500 billion words (Brants & Franz, 2006; Shaoul & Westbury, 2009).

A second source of good frequency estimates for psycholinguistic research was textbooks aimed at primary and secondary school children. This source gained importance in research on the age-of-acquisition effect in visual word recognition, which demonstrates that words learned early in life keep a processing advantage over words learned later in life, even when corrected for the best possible frequency norms (for reviews, see Ghyselinck, Lewis, & Brylsbaert, 2004; Johnston & Barry, 2006; Juhasz, 2005; see also Cortese & Khanna, 2007, for the most recent evidence on this for English monosyllabic words). The database most often used for childhood frequencies in English is the Zeno database (Zeno, Ivens, Millard, & Duvvuri, 1995). It is based on 17 million words from a wide range of texts written for children from grades 1–12. Even though it is a rather small corpus (certainly in comparison with the Internet corpora), it correlates as highly with word processing times as the best collection of Internet-based word frequencies (Balota et al., 2004; Brylsbaert & New, 2009). This illustrates that, although the size of the corpus is an important element, the language register on which the frequency estimate is based is as important (in this case, children's books vs. Internet Web sites). On the basis of simulations with the British National Corpus, Brylsbaert and New estimated that, when used to predict word processing times, larger corpora yield significantly better frequency estimates up to a corpus size of about 16 million words, but that, for larger corpus sizes, the gains become vanishingly small if the corpus has been well sampled.²

Finally, film and television subtitles turned out to be another interesting source of word frequencies. New, Brylsbaert, Veronis, and Pallier (2007) observed this first for French, where subtitle frequencies explained more of the variance in lexical decision reaction times (RTs) than did frequency measures based on a selection of written materials (including books, newspapers, or Internet sources). Brylsbaert and New (2009) subsequently replicated the finding in English and found that their subtitle frequency measure did better than did Zeno and Internet-based frequencies in predicting word naming and lexical decision performance (RTs and percentages of error). Brylsbaert and New hypothesized that this was because film and television language approximates everyday word use better than written sources do.

Contextual Diversity Rather Than Raw Frequency of Occurrence

Another recent development has been the finding that the number of times a word occurs in a corpus is less informative than the number of documents in which the word appears (Adelman, Brown, & Quesada, 2006). Adelman et al. called this new measure *contextual diversity* (CD). The advantage of CD over the measure WF was confirmed by Brylsbaert and New (2009) for subtitles: Frequencies based on the number of films in which a word appeared accounted for 1%–3% more of the variance in lexical decision performance than did frequencies based on the raw number of occurrences.

The Collection of New Data for Dutch

The introduction of the CELEX database was of critical importance for psycholinguistic research in the Dutch language. CELEX offers extremely valuable information on lexical characteristics, such as phonology and morphology. However, given the developments outlined above, it seems necessary (A) to validate the CELEX frequencies on a sufficiently large sample of word processing data and (B) to compare the CELEX frequencies with a subtitle-based frequency measure.

Next we describe the subtitle frequency measure we collected for Dutch and the lexical decision megastudy that we ran to validate the frequencies. We begin with the subtitle frequency measure.

SUBTLEX-NL

Subtitles are increasingly available on the Internet, because they can easily be integrated in digital films. Between March 10 and March 19, 2009, a computer program written specifically for this purpose processed a large number of Dutch subtitles found on an Internet site grouping contributions made available by individual Internet users (www.ondertitels.nl). Disregarding duplicates, the program processed 43,729,424 words coming from 8,443 subtitles, of which the majority (5,966) were translated subtitles of American films and television series (we used the Internet Movie Database [www.imdb.com] to determine the countries of origin).

The number of words on which these word frequencies were based is slightly smaller than what has been assembled for French and English (50 million words), but it is well above the required 16 million words and is large enough to allow estimates per million with 1-digit precision (see also note 2). In addition to the number of times each word was encountered (WF), we also calculated the number of films or television shows in which it appeared (CD). In total, there were 8,070 contexts (subtitles covering different parts of the same film were counted as one context).³

Similar to what was done for the French subtitle frequencies and the CELEX database, we wanted to have information about the various grammatical functions of words in addition to the frequencies of the word forms themselves. This will allow users to calculate various types of frequencies (e.g., the frequency of the word form *play* as a noun—as opposed to a verb—and the frequency of the lemma *play*_{noun}, consisting of the summed frequencies of the word forms *play*_{noun} and *plays*_{noun}). To this end, we used the Tadpole program (available at <http://ilk.uvt.nl/tadpole/>), an integrated Dutch morphosyntactic analyzer and part-of-speech tagger (van den Bosch, Busser, Canisius, & Daelemans, 2007). The output of the Tadpole program allowed us to calculate WF and CD for the lemmas, defined as the sum of all inflected forms associated with a particular part of speech (e.g., *play* as a noun consisting of *play*_{noun} + *plays*_{noun}, and *play* as a verb consisting of *play*_{verb} + *plays*_{verb} + *played*_{verb} + *playing*_{verb}).

A Validation of the CELEX and the SUBTLEX-NL Frequency Measures

Because visual lexical decision is particularly sensitive to word frequencies (Balota et al., 2004; Brysbaert & New, 2009; Cortese & Khanna, 2007; Yap & Balota, 2009), it is a particularly informative task to validate a frequency measure. In English, there are two databases of lexical decision performance: that collected by Balota, Cortese, and Pilotti (1999), which consists of data from 30 younger and 30 older adults who made lexical decisions to 2,905 monosyllabic words, and that collected as part of the Elexicon Project (Balota et al., 2007; available at <http://elexicon.wustl.edu/>), which contains lexical decision RTs and accuracies for over 40,000 English words collected from hundreds of participants. Because a similar database was not available in Dutch, we decided to make one. Given that most psycholinguistic research is based on mono- and disyllabic words, we limited our study to these words.

Method

Stimuli. The study involved mono- and disyllabic words. For the most part, the stimuli were taken from the CELEX database, because this gave us valuable information about lexical characteristics, such as syllabic structure. We started with all mono- and disyllabic words with a frequency of 1 per million or higher in CELEX and included some extra low-frequency words we had needed in our previous research (e.g., *wilg* [willow]). Next, we included the major inflected forms of the selected set,⁴ regardless of their frequency. This resulted in a total of 14,037 words.

The Wuggy pseudoword generator (Keuleers & Brysbaert, 2010) was used to construct a corresponding pseudoword for each word in the experiment. Each pseudoword differed from the reference word by one subsyllabic segment (i.e., the onset, nucleus, or coda) per syllable. This implied that a one-syllable nonword differed in one position from its reference word and that a two-syllable nonword differed in two positions from its reference word. An advantage to this approach is that longer pseudowords still look very word-like but cannot be tied to a specific word, in contrast to other approaches, in which only one or two letters of the reference word are changed, independent of its length (e.g., Balota et al., 2007). Each nonword was generated by changing the position in the syllable that resulted in the smallest possible change in syllable frequency and in the transition frequencies of the syllables and subsyllabic segments. In this way, high-frequency morphological affixes of words tended to be maintained in their nonword counterparts (changing these affixes would almost always result in a larger change in transition frequencies compared with changing other segments). As a result, the pseudo-morphological structure of the nonwords very much resembled the morphological structure of the words.

Participants. Participants were 39 students and employees (32 female, 7 male) from Ghent University. Each participant responded to all 28,074 test trials. Participants needed 14–20 h to complete the experiment, at their own pace, over a 6-week period. They were paid €200 upon successful completion. Four more participants did not finish the experiment because their performance consistently dropped below 80% correct. They were paid €5 per hour (the different payment rates for successful vs. unsuccessful completion were made clear before the participants gave their informed consent).

Design. The words and nonwords were assigned randomly to 56 blocks of 500 stimuli (a different permutation was generated for each participant). Each block took about 15–17 min to finish and was subdivided into five parts of 100 stimuli each. Between each part, participants were asked to press on the space bar to continue. Although most participants continued immediately, they all reported that they liked the interruptions, because these increased their control and provided them with information about the progress in the block.

Stimuli were presented centrally on a computer screen in white lowercase letters against a black background (Times Roman, 18 pts. bold). A trial started with the presentation of two vertical fixation lines slightly above and below the center of the screen, with a gap between them wide enough to clearly present a horizontal letter string. Participants were asked to fixate the gap as soon as the lines appeared. After 500 msec, the stimulus was presented in the gap with the center between the vertical lines; the vertical lines remained on the screen. The stimulus stayed on the screen until the participant made a response or for a maximum of 2 sec. Participants used their dominant hand for word responses and their other hand for nonword responses (using response buttons of an external response box connected to a USB port). After the response, there was an interstimulus interval of 500 msec before the next trial started. The screen was blank in this interval. At the end of each block, participants received feedback about their accuracy in the block.

Participants booked time slots at one of four computers integrated in a network (so that the data could be stored centrally and participants did not always have to sit at the same computer). Participants entered their participation code, and, after verification, the computer automatically allocated the correct block to them (the experiment was programmed using the Tscope library; Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006). Participants were not allowed to run more than seven blocks in a row (about 2 h).

Results

The two dependent variables were accuracy (percentage correct, PC) and reaction time (RT) of the correct trials. Mean accuracy of the participants was 84% ($SD = 4.1$) for the words and 94% ($SD = 5.6$) for the nonwords.

Table 1
Percentages of Variance in Accuracy and Reaction Time (RT)
Explained by the Different Frequency Measures

Measure		Accuracy (%) (<i>N</i> = 12,964)	RT (%) (<i>N</i> = 11,386)
CELEX	Log	13.4	25.9
	Log + log ²	15.4	26.2
CELEX _{CD}	Log	18.8	25.2
	Log + log ²	19.1	26.8
SUBTLEX _{WF}	Log	17.3	33.9
	Log + log ²	22.0	34.9
SUBTLEX _{CD}	Log	20.6	35.1
	Log + log ²	25.3	35.2

Note—Because of the large number of observations, differences in explained variance as small as .1 are statistically significant.

Mean RT was 659 msec (*SD* = 189) for the words and 680 msec (*SD* = 192) for the nonwords.

For each word, PC and RT were calculated by taking the mean of the 39 participants. To get an estimate of the reliability of the measures, we computed the split-half correlations and corrected them for length using the Spearman–Brown formula⁵ for 100 random splits of the data (each time, 20 participants were randomly assigned to the first group and 19 were assigned to the second group). Mean corrected test–retest reliability was .79 (*SD* = .0056) for RTs and .96 (*SD* = .0012) for accuracy.⁶

The analyses reported below include only those word forms that have a frequency above 0 in both CELEX and SUBTLEX-NL. Furthermore, words that were judged to be nonwords by more than a third of the participants were not included in the RT analyses. The words that were excluded because of the accuracy threshold were mostly low-frequency words, but there were also some very short, high-frequency function words (e.g., *ten*, *der*, *bent*, *per*) and a surprising number of names, indicating that some participants did not consider these as words. A total of 12,964 words remained for the accuracy analyses and 11,386 words for the RT analyses.

Table 1 displays the percentages of variance in accuracy and RT accounted for by the different frequency measures based on the word forms (e.g., the word form *play*, irrespective of the lemma it belonged to, or the word form *plays*, irrespective of the lemma it belonged to). Frequency measures were log₁₀ transformed. Because Balota et al. (2004; see also Baayen et al., 2006) found that the relationship between log frequency and word processing performance is not completely linear (in particular, a floor effect seems to have been reached for words with a frequency above 100 per million), we report regression analysis both for log(frequency) and for log(frequency) + log²(frequency).

The first two lines in Table 1 show the results for the CELEX frequency measure: first when log(frequency) is entered as a predictor in the regression, then when both log(frequency) and log²(frequency) are entered. As can be seen, log(frequency) explained 13% of the variance in accuracy and 26% of the variance in RT. Adding log²(frequency) significantly increased the percentage of

variance explained in both accuracy and RT, in line with the findings reported for English.

The next two lines in Table 1 show the results for the CD measure in CELEX. The CELEX database includes such a measure, but it is not listed in the word-form database used by most researchers. However, it can be found in the Dutch and German corpus types database, where it is called *dispersion*. Compared with the WF measure, the CD measure explains substantially more of the variance in accuracy, but not in RTs.

The third entry in Table 1 shows the results for the SUBTLEX_{WF} measure, again, with the predictors log(frequency) and log(frequency) + log²(frequency). In line with the previous findings for French and English, the SUBTLEX_{WF} measure explains some 4% more of the accuracy data and nearly 8% more of the variance in RTs than the CELEX measure does.⁷

The last two lines in Table 1 show the results for SUBTLEX_{CD}. As expected, the CD measure explains 1%–3% more variance in accuracy and RT relative to the WF measure.

To examine the usefulness of lemma frequencies in explaining lexical decision performance, we entered them as an extra variable to the regressions of Table 1. A choice to be made here was how to define the lemma frequency of the stimuli presented in the experiment. Formally, a word's lemma frequency is defined as the sum of the frequencies of all the inflected forms of the root form. However, since inflection is defined only within a grammatical class (e.g., noun, verb), it is unclear which lemma frequency to use for stimuli that belong to more than one grammatical class. Take, for instance, the form *delen*. As the infinitive form of the verb *delen* (*to divide*, *to share*), its lemma frequency should include the frequencies of all the inflectional forms of the verb (e.g., the present and past tenses, the past participle). However, *delen* is also the plural of the noun *deel* (*part*, *share*). Should this lemma frequency be added or not? We opted for the former, because word-form frequencies are also summed over syntactic categories and, therefore, we defined the lemma frequency of a presented word form as the sum of the lemma frequencies of all its possible interpretations (e.g., the lemma frequency of *delen* was defined as the sum of the lemma frequency of *delen*_{verb} and the lemma frequency of *deel*_{noun}).⁸

Table 2 lists the percentages of variance in accuracy and RT explained when lemma frequency is added to the predictors in Table 1. Importantly, for CELEX the CELEX lemma frequency was used, whereas for SUBTLEX the SUBTLEX lemma frequencies were used. As can be seen in Table 2, lemma frequency added up to nearly 10% of extra variance explained in the accuracy data and up to 2% in RTs. The gains were larger for the SUBTLEX frequencies than for the CELEX frequencies. This is further testimony to the quality of the SUBTLEX measure.

A final noteworthy aspect of Table 2 is that the extra contribution of lemma frequencies is quite small for RTs. This means that, for most practical purposes (e.g., the selection of lists of stimuli matched on frequency), researchers can limit their efforts to word-form frequencies.

Table 2
Percentages of Variance Explained by Lemma Frequency
Together With Word-Form Frequency

Measure		Accuracy (<i>N</i> = 12,964)	RT (<i>N</i> = 11,386)
CELEX	Log	18.6 (5.2)	26.9 (1.0)
	Log + log ²	18.6 (3.3)	27.0 (0.8)
CELEX _{CD}	Log	21.1 (2.3)	25.5 (0.3)
	Log + log ²	21.1 (2.0)	27.4 (0.6)
SUBTLEX _{WF}	Log	26.7 (9.4)	35.9 (2.0)
	Log + log ²	26.9 (4.9)	35.9 (0.9)
SUBTLEX _{CD}	Log	28.0 (7.4)	36.1 (1.0)
	Log + log ²	28.7 (3.4)	36.2 (1.0)

Note—Between parentheses, the additional variance that lemma frequency explains relative to the variance explained by word-form frequency alone. RT, reaction time.

Availability

The SUBTLEX-NL frequencies are freely available for research purposes. We have summarized the frequency information in two files, which are available in the supplemental materials for this journal and at <http://crr.ugent.be/subtlex-nl>.

The first file, SUBTLEX-NL.master, is a text file, containing the outcome of the tagged analysis. Researchers familiar with the frequency lists made from the British National Corpus (<http://ucrel.lancs.ac.uk/bncfreq/>) will recognize the layout, since we chose to use a very similar format. Words are listed alphabetically, both as lemmas and as word forms. Figure 1 gives the information about the noun *deel* (part) and the verb *delen* (to divide/to share). The first line of the noun lemma *deel* includes four numbers: first the summed frequency (6,986) of all

different forms of the lemma, then the CD of the lemma (3,697), the summed frequency of all forms starting with a lowercase letter (6,801), and the CD of the lemma starting with a lowercase letter (3,596). Our previous work (Brysbaert & New, 2009) has shown that the distinction between words starting with a lowercase and an uppercase letter is interesting to filter out words that are often used as names. The frequency of these words tends to be overestimated, as can be concluded from the finding that their word processing times are more in line with their lowercase frequency than with their total frequency.

Below the lemma line for *deel*_{noun}, there are four lines with the constituting forms (each line starting with “@ @,” since these fields duplicate information from the lemma line). Each form is followed by the detailed part-of-speech tag assigned by the automatic analysis in Tadpole, its morphological analysis by the Tadpole system, and the four frequency values already described.

The next lines in Figure 1 describe all the relevant information for the verb lemma *delen*_{verb} (the abbreviation WW stands for *werkwoord*, the Dutch word for *verb*).

The SUBTLEX-NL.master file will be of use to anyone who wants to calculate word characteristics that go beyond the mere word forms (such as different definitions of lemma frequency, inflectional entropy, and so on). There are two versions of it: (1) with all the words, and (2) with the words that have a lemma CD above 2. The latter is substantially shorter and excludes many typos that are present in the database.

The second file (SUBTLEX-NL) is a simpler file, in the sense that it contains information only about the different letter strings in the corpus with a CD of more than 1.

deel	N	%	%	6986	3697	6801	3596
@	@	deel	soort,ev,basis,onz,stan	5974	3318	5805	3222
@	@	deel	soort,ev,basis,zijd,stan	2	2	2	2
@	@	dele	soort,ev,basis,dat	21	18	21	18
@	@	delen	soort,mv,basis	989	847	973	834
delen	WW	%	%	3122	2211	3007	2135
@	@	deel	pv,tgw,ev	245	238	180	175
@	@	deeld	vd,vrij,zonder	1	1	1	1
@	@	deelde	pv,verl,ev	151	141	149	140
@	@	deelden	pv,verl,mv	114	109	112	107
@	@	deelt	pv,tgw,met-t	330	300	316	287
@	@	delen	inf,nom,zonder,zonder-n	44	41	44	41
@	@	delen	inf,vrij,zonder	1589	1296	1585	1293
@	@	delen	pv,tgw,mv	307	289	307	289
@	@	delend	od,vrij,zonder	3	3	3	3
@	@	gedeeld	vd,prenom,zonder	6	6	6	6
@	@	gedeeld	vd,vrij,zonder	262	230	250	219
@	@	gedeelde	vd,prenom,met-e	70	54	54	44

Figure 1. Layout of the SUBTLEX-NL.master file. A line starting with a word signifies a lemma (e.g., *deel* as a noun [N] and *delen* as a verb [WW]). Lines starting with “@ @” indicate word forms. Each line includes the specific form and the part-of-speech tag assigned by the program Tadpole. The final four columns include word frequency (WF) and contextual diversity (CD) of the word and WF and CD of the word starting with a lowercase letter.

This is the file researchers will use when they simply want to know the frequency of their stimulus words. It exists as both a text file and an Excel file (again with all words or only with the words that have a CD above 2). People familiar with our English SUBTLEX-US database (Brylsbaert & New, 2009) will be familiar with its layout. We added only a column with lemma frequency (see Figure 2).

The definitions of the different columns are as follows:

1. The word.
2. **FREQcount** is the number of times the word appears in the corpus (i.e., on the total of 43.8 million words).
3. **CDcount** is the number of films in which the word appears (i.e., it has a maximum value of (8,070).
4. **FREQlow** is the number of times the word appears in the corpus starting with a lowercase letter. This allows users to further match their stimuli.
5. **CDlow** is the number of films in which the word appears starting with a lowercase letter.
6. **FREQlemma** is the sum of the frequencies of all lemmas to which the word belongs.
7. **SUBTLEX_{WF}** is the WF per million words and has 4-digit precision. It is the measure that researchers would preferably use in their manuscripts, because it is a standard measure of WF independent of the corpus size.
8. **Lg10WF** is a value based on $\log_{10}(\text{FREQcount} + 1)$ and with 4-digit precision. Calculating the log frequency on the raw frequencies is the most straightforward transformation, because it allows researchers to give words that are not in the corpus a value of 0. One can easily lose 5% of the variance explained by taking $\log(\text{frequency per million} + 1)$, because, in this case, there is not much distinction between words with low frequencies. Similarly, adding values lower than 1 (e.g., $+1E-10$) is dangerous, because one may end up with a big gap between the words in the corpus and words for which there is no frequency measure (which will get a log value of -10). Also, if one uses $\log(\text{frequency per million})$, one obtains negative values for words with a frequency lower than 1 per million and one has to enter negative values for missing words.
9. **SUBTLEX_{CD}** indicates in what percent of the films the word appears, with 4-digit precision. For instance, the word *de* (the) has a **SUBTLEX_{CD}** of 100.00, because it

occurs in each film. In contrast, the word *afkorting* (abbreviation) has a **SUBTLEX_{CD}** of 1.7, because it appears in only 74 films.

10. **Lg10CD**. This value is based on $\log_{10}(\text{CDcount} + 1)$ and has 4-digit precision. As is shown in Table 1, overall this is the best value to match stimuli on.

Conclusion

In the present article, we presented a frequency measure for the Dutch language, based on subtitles, which is superior to the existing CELEX frequencies, as shown by a lexical decision validation study involving most known monosyllabic and disyllabic Dutch words. As in English we found that the CD measure outperforms the WF measure. For RTs, it explained 35% of the variance between words; for accuracy, this was 26%. For the latter variable, we saw a clear additional effect of the lemma frequency, and it will be interesting to examine the underlying processes. Compared with the CELEX frequencies for Dutch, the SUBTLEX-NL frequencies are an improvement of almost 10% in explained variance in RTs. Therefore, we think that the SUBTLEX-NL word frequencies will be of valuable use for language research in general, and for word recognition research in particular. Although the lexical information contained in the CELEX lexical database remains invaluable, the SUBTLEX-NL word frequencies should be preferred over the CELEX frequencies when selecting stimuli for experiments. Next, the SUBTLEX-NL word frequencies will allow researchers to optimally control and account for the effects of WF when other variables in word processing are under investigation. Finally, SUBTLEX-NL shares an important feature with CELEX, in that it has both lemma frequencies and word-form frequencies.

At the same time, our article shows how easy it has become to make a good WF list for a language. Whereas it took a big investment in time and manpower to compile the CELEX frequencies in the late 1980s, two recent developments made it possible for us to collect new lists of word frequencies in a matter of weeks. First, although the compilation of the corpus on which the CELEX frequencies are based involved the lengthy process of scanning printed

Word	FREQcount	CDcount	FREQlow	CDlow	FREQlemma	SUBTLEXWF	Lg10WF	SUBTLEXCD	Lg.0CD
ik	1744062	8054	778704	3125	1744527	39883.0316	6.2416	99.8017	3.9061
je	1600888	8060	1315051	6535	1600923	36608.9432	6.2044	99.8761	3.9064
het	1068396	8066	780771	5578	1913811	24431.9706	6.0287	99.9504	3.9067
de	1061177	8070	903872	6512	1063827	24266.8872	6.0258	100	3.9069
dat	965424	8063	715570	6107	965431	22077.2174	5.9847	99.9133	3.9066
is	947568	8067	891894	7636	1891610	21668.8882	5.9766	99.9628	3.9068
niet	801297	8066	755407	7498	801317	18323.9779	5.9038	99.9504	3.9067
een	785913	8069	691904	6864	868582	17972.1782	5.8954	99.9876	3.9069
en	611665	8063	422748	5936	611666	13987.4927	5.7865	99.9133	3.9066
wat	480650	8061	220264	3577	480680	10991.4551	5.6818	99.8885	3.9064
van	455234	8062	443772	7754	455234	10410.2446	5.6582	99.9009	3.9065
we	443891	8058	250759	4447	443894	10150.854	5.6473	99.8513	3.9063
ze	407143	8053	253203	4941	407155	9310.5045	5.6097	99.7893	3.906
hij	388131	8048	198012	3623	388131	8875.7401	5.589	99.7274	3.9057
in	385809	8066	353918	7197	385809	8822.6408	5.5864	99.9504	3.9067
maar	366703	8060	254294	5637	366704	8385.7267	5.5643	99.8761	3.9064

Figure 2. Layout of the SUBTLEX-NL file. See the text for the explanation of the column titles.

sources, written material is now ubiquitously available in digital format. In particular, the subtitles of popular films and television series seem to contain a representative sample of the language and come in handy packages (on average some 5,000 words per film or television episode). Second, it is easy to write software to reliably count the number of occurrences of words in text files.

A significant convenience for this line of research is that subtitles are readily available on various Internet sites and in various languages. In the development of our WF database, we automatically processed thousands of these subtitles with relatively little effort. Although it is impossible to determine the origin of each subtitle file, most subtitles available on the Internet appear to fall into two categories: Either they are copies of the original subtitles available on DVD or other media, or they are translations or transcripts made by interested persons (fan-created subtitles, or “fansubs”). Although using these subtitles for our research is convenient and inexpensive, there are some legal and ethical issues to consider.

Providing subtitles for download without explicit permission from the rights holders may be a violation of copyright laws in several countries. For files taken directly from DVD, the rights holders must grant permission for publishing on an Internet site. Arguably, the rights holders’ major concern is that combining these subtitles with illegally downloaded copies of films allows people all over the world to watch the films with foreign-language subtitles, thus precluding the sale of a legally distributed film. Even fan-created subtitles may not be free from copyright restrictions, depending on whether they are considered transformative.

As of yet, we are not aware of court rulings in legal cases opposing Internet sites hosting subtitles to rights holders, although we have been made aware of some legal action being taken and of substantial threats from entertainment companies (Cassel, 2007; Enigmax, 2009). Furthermore, since different countries have different legal systems, they may also come to different conclusions regarding the legality of these sites.

To the best of our understanding, our use of the subtitles as described in this research is not a violation of copyright because (among other things) the WF database is only a statistical description of the subtitles. This is considered to be fair use of copyrighted material. However, in research benefiting from potentially illegal activity, ethical issues should also be considered. Much of the WF database could, in theory, be recreated without using the subtitle Internet sites. DVDs of movies and television shows could be purchased (or borrowed from a library) and the subtitles could be extracted for the analysis we describe in this article. Should subtitles not be available, we could create our own transcripts in a variety of languages. However, the working costs associated with such an approach would be prohibitive, and the end result would be essentially the same in content as accessing the subtitle Internet sites.

The increased availability of information on the Internet will likely cause researchers to frequently run into these kinds of issues. When using Internet material that may be subject to copyright issues for scientific research, the ben-

efits should be carefully weighed against the possible ethical and legal issues. We have tried to be transparent about these issues surrounding our research. In our opinion, three factors justify making our WF database available for scientific research. First, making word frequencies is fair use of copyrighted material, since it is clearly transformative: The list of frequencies bears no relation to the primary use of subtitles—to accompany a film. Second, the word frequencies have a clear scientific value, as shown by the validation study described above. Finally, the alternative, processing or transcribing subtitles on the basis of original media, is prohibitive in terms of working costs.

We think it is good practice to validate the obtained frequencies with lexical decision times. This is why we invested considerably in the collection of a large database. However, analyses by Burgess and Livesay (1998) and New et al. (2007) have suggested that differences in quality between various frequency counts can be detected with samples of only a few hundred words spread over the entire frequency range. So, it may not be necessary to collect data for thousands of words. A typical 1-h experiment with some 1,000 words and nonwords may be enough.

AUTHOR NOTE

Address correspondence to E. Keuleers, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium (e-mail: emmanuel.keuleers@ugent.be).

REFERENCES

- ADELMAN, J. S., BROWN, G. D. A., & QUESADA, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*, 814-823. doi:10.1111/j.1467-9280.2006.01787.x
- BAAYEN, R. H., FELDMAN, L. B., & SCHREUDER, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory & Language*, *55*, 290-313.
- BAAYEN, R. H., PIEPENBROCK, R., & VAN RIJN, H. (1993). The CELEX Lexical Database [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.
- BALOTA, D. A., CORTESE, M. J., & PILOTTI, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. *Abstracts of the 40th Annual Meeting of the Psychonomic Society*, *4*, 44.
- BALOTA, D. A., CORTESE, M. J., SERGENT-MARSHALL, S. D., SPIELER, D. H., & YAP, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283-316.
- BALOTA, D. A., YAP, M. J., CORTESE, M. J., HUTCHISON, K. A., KESSLER, B., LOFTIS, B., ET AL. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445-459.
- BONTRAGER, T. (1991). The development of word frequency lists prior to the 1944 Thorndike-Lorge list. *Reading Psychology*, *12*, 91-116. doi:10.1080/0270271910120201
- BRANTS, T., & FRANZ, A. (2006). Web 1T 5-Gram Corpus (Version 1). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.
- BRYBAERT, 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
- 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.
- CASSEL, D. (2007, May 17). Police raid Polish subtitle site [Online article]. Retrieved from <http://tech.blorge.com/Structure:%20/2007/05/17/police-raid-polish-subtitle-site/>.

- CORTESE, M. J., & KHANNA, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, **60**, 1072-1082.
- ENIGMAX (2009, February 5). Hackers hit anti-pirates to avenge sub-site takedown [Online article]. Retrieved from <http://torrentfreak.com/hackers-hit-anti-pirates-to-avenge-sub-site-takedown-090205/>.
- GHYSELINCK, M., LEWIS, M. B., & BRYLSBAERT, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, **115**, 43-67.
- JOHNSTON, R. A., & BARRY, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, **13**, 789-845.
- JUHASZ, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, **131**, 684-712.
- KEULEERS, E., & BRYLSBAERT, M. (2010). Wuggy: A multilingual pseudo-word generator. *Behavior Research Methods*, **42**, 627-633.
- KU ERA, H., & FRANCIS, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- NEW, B., BRYLSBAERT, M., VERONIS, J., & PALLIER, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, **28**, 661-677.
- SHAOL, C., & WESTBURY, C. (2009). A USENET corpus (2005-2009). Edmonton: University of Alberta. Retrieved from www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html.
- STEVENS, M., LAMMERTYN, J., VERBRUGGEN, F., & VANDIEREN-DONCK, A. (2006). Tscope: A C library for programming cognitive experiments on the MS Windows platform. *Behavior Research Methods*, **38**, 280-286.
- THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words*. New York: Columbia University, Teachers College.
- UIT DEN BOOGAART, P. C. (Ed.) (1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema Holkema.
- VAN BERCKEL, J., BRANDT CORSTIUS, H., MOKKEN, R., & VAN WIJNGAARDEN, A. (1965). *Formal properties of newspaper Dutch*. Amsterdam: Mathematisch Centrum Amsterdam.
- VAN DEN BOSCH, A., BUSSER, B., CANISIUS, S., & DAELEMANS, W. (2007). An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, & F. Van Eynde (Eds.), *Computational linguistics in the Netherlands: Selected papers from the Seventeenth CLIN Meeting* (pp. 99-114). Leuven.
- YAP, M. J., & BALOTA, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory & Language*, **60**, 502-529. doi:10.1016/j.jml.2009.02.001
- YARKONI, T., BALOTA, D., & YAP, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, **15**, 971-979.
- ZENO, S. M., IVENS, S. H., MILLARD, R. T., & DUVVURI, R. (1995). *The educator's word frequency guide*. Brewster, NJ: Touchstone Applied Science Associates.
- ZEVIN, J. D., & SEIDENBERG, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory & Language*, **47**, 1-29. doi:10.1006/jmla.2001.2834

NOTES

1. The CELEX database also contains an English and a German part.
2. The small gains above 16 million words became clear in the present analyses as well. Our original estimates were based on a subsample of 33 million words instead of the 43 million reported here. The differences in percentages variance explained never exceeded 0.5%.
3. We also calculated a different CD measure in which we grouped all film sequels and episodes of a television series, on the basis of the assumption that these files contained repeated information and that people were likely either not to have seen any episode or to have seen more than one. This definition made a total of 5,834 contexts. However, the correlation between this measure and the one mentioned in the article was .9976 and, hence, there were no significant differences between the measures.
4. For instance, for the verbs these were the different forms of the present and the past tense and the past participle.
5. $r_{\text{corr}} = (2 * r) / (1 + r)$, where r is the split-half correlation and r_{corr} is the correlation corrected for length.
6. We thank Kevin Diependaele for his help in computing these results.
7. The superiority of the SUBTLEX_{WF} measure is maintained when two other important variables in lexical decision times, word length and neighborhood size (operationalized as OLD20; see Yarkoni, Balota, & Yap, 2008), are entered in the regression. In combination with these variables, the log and log² of the CELEX frequencies explained 21.2% of the variance in PC and 27.3% of the variance in RT; the variance explained by SUBTLEX_{WF} in similar regression analyses was 30.9% of the variance in PC and 35.0% of the variance in RT. A similar advantage of SUBTLEX_{CD} over CELEX_{CD} was found.
8. Another advantage of summing the lemma frequencies across syntactic categories is that differences in tagging quality between CELEX and SUBTLEX-NL have little impact on the frequency estimates. Differences in output between taggers nearly always have to do with assigning the syntactic category to the word (e.g., is *play* used as a noun or a verb?).

SUPPLEMENTAL MATERIALS

The full SUBTLEX-NL database may be downloaded from <http://brm.psychonomic-journals.org/content/supplemental>.

(Manuscript received August 12, 2009;
revision accepted for publication March 27, 2010.)