Preprint of paper published as: Keuleers, E., Brysbaert, M., & New, B. (2010).
SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles.
Behavior Research Methods, 42(3), 643–650. doi:10.3758/BRM.42.3.643

# SUBTLEX-NL:

## A new frequency measure for Dutch words based on film subtitles

Emmanuel Keuleers and Marc Brysbaert

Ghent University, Ghent, Belgium

Boris New

Université Paris Descartes and CNRS UMR 8189, Paris, France

Address for correspondence :

Emmanuel Keuleers

Ghent University

Department of Experimental Psychology

Henri Dunantlaan 2

B-9000 Gent, Belgium

Tel : +32 9 264 64 06

Email : emmanuel.keuleers@ugent.be

**Abstract**

A new database of Dutch word frequencies based on film and television subtitles is presented and validated with a lexical decision study involving fourteen thousand monosyllabic and disyllabic Dutch words. The new SUBTLEX frequencies explain up to 10% more variance in accuracies and reaction times of the lexical decision task than the existing CELEX word frequency norms, which are largely based on edited texts. As in English, an accessibility measure based on contextual diversity explains more of the variance in accuracy and RT than the raw frequency of occurrence counts. The database is freely available for research purposes.

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

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 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 word frequency 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 this date (Brysbaert & 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 about the Dutch language, such as ngram sequences up to 3 letters, it did not gain wide adoption. The first publicly available frequency list for Dutch was edited by Uit den Bogaart (1975), who published frequencies of "written and spoken Dutch" on the basis of 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 two 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. Below, we summarize these developments before we return to the Dutch language.

**Edited texts may not be the best source of information for word frequencies**

When researchers started comparing the correlation between different word-frequency measures and 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; Brysbaert & 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 are 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, & Brysbaert, 2004; Johnston & Barry, 2006; Juhasz, 2005; also see 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 grade 1 to grade 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 internet-based word frequencies (Balota et al., 2004; Brysbaert & 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 equally important (in this case, children's books vs. internet websites). On the basis of simulations with the British National Corpus, Brysbaert and New (2009) 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, Brysbaert, Veronis, and Pallier (2007) observed this first for French, where subtitle frequencies explained more of the variance in lexical decision RTs than frequency measures based on a selection of written materials (including books, newspapers, or internet sources). Brysbaert and New (2009) subsequently replicated the finding in English and found that their subtitle frequency measure did better than Zeno and internet-based frequencies in predicting word naming and lexical decision performance (RTs and percentages of error). Brysbaert and New (2009) hypothesized that this was because film and television language approximates everyday word use better than written sources.

**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 it appears (Adelman, Brown, & Quesada, 2006). Adelman et al. called this new measure "contextual diversity" (CD) as opposed to the traditional word frequency (WF) measure. The advantage of CD over WF was confirmed by Brysbaert 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 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. To this date CELEX offers extremely valuable information on lexical characteristics, such as phonology, morphology, etc. 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 to a subtitle-based frequency measure.

Below we describe the subtitle frequency measure we collected for Dutch and the lexical decision megastudy 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 to 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 decide from which country the films originated).

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 large enough to allow estimates per million with one digit precision (see also footnote 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*$_{\text{noun}}$, 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 morpho-syntactic 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$_{\text{noun}}$ + plays$_{\text{noun}}$, and *play* as a verb consisting of play$_{\text{verb}}$ + plays$_{\text{verb}}$ + played$_{\text{verb}}$ + playing$_{\text{verb}}$).

## 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. The first is a database collected by Balota, Cortese, and Pilotti (1999). It consists of data from 30 younger and 30 older adults who made lexical decisions to 2,905 monosyllabic words. The second is 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 monosyllabic 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, as 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], also those with a frequency lower than 1 per million. This resulted in a total of 14,037 words.

The Wuggy pseudoword generator (Keuleers & Brysbaert, in press) 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 differred in one position from its reference word and that a two-syllable nonword differed in two positions from its reference word. One of the advantages of this approach is that longer pseudowords still look very-word like but can not be tied to a specific word, in contrast to other approaches, where 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 to 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 hours 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 randomly assigned to 56 blocks of 500 stimuli (a different permutation was generated for each participant). Each block took about 15-17 minutes to finish and was subdivided in 5 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. Five hundred milliseconds later 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 s. 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 ms 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 4 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 T-scope library; Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006). Participants were not allowed to run more than 7 blocks in a row (about two hours).

**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. Mean RT was 659 ms (SD = 189) for the words and 680 ms (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 formula[5] for 100 random splits of the data (each time, 20 participants were randomly assigned to the first group and 19 to the second group).

Mean corrected test-retest reliability was .79 (SD= .0056) for reaction times 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 also some very short, high-frequency function words (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). This is the word frequency measure traditionally used in (American) English. 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 be reached for words with a frequency above 100 per million), we report regression analysis both for log(frequency) and log(frequency) + log²(frequency).

---

Insert Table 1 about here

---

The first 2 lines of Table 1 show the results for the CELEX frequency measure: first when log(freq) is entered as a predictor in the regression, then when both log(freq) and log²(freq) are entered. As can be seen, log(freq) explained 13% of the variance in accuracy and 26% of the variance in RT. Adding log²(freq) significantly increased the percentage of variance explained in both accuracy and RT, in line with the findings reported for English.

The next two lines of Table 1 show the results for the contextual diversity measure in CELEX. Not many people know that the CELEX database includes such a measure, because it is not listed in the wordform database used by most researchers. However, it can be found in the corpus types database for Dutch and German, where it is called 'dispersion'. Compared to the WF measure, the CD measure explains substantially more of the variance in accuracy but not in RTs.

The third entry of Table 1 shows the results for the $SUBTLEX_{WF}$ measure, again with the predictors log(freq), and log(freq) + log²(freq). As can be seen, in line with the previous findings in 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.[7]

The last two lines of Table 1 show the results for $SUBTLEX_{CD}$. As expected, the CD measure explains 1 to 3 percent 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 only defined 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 (i.e., the present and past tenses, the past participle, etc). 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 (i.e., 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.

Insert Table 2 about here

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.

## 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 archive of 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, as 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 (6986) of all different forms of the lemma, then the CD of the lemma (3697), the summed frequency of all forms starting with

a lower case letter (6801), and the CD of the lemma starting with a lower case (3596). 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.

---

Insert Figure 1 about here

---

Below the lemma line for *deel*$_{noun}$, there are 4 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 only contains information about the different letter strings in the corpus with a CD of more than 1. This is the file researchers will use when they simply want to know the frequency of their stimulus words. It exists both as 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 (Brysbaert & New, 2009) will be familiar with its lay-out. We only added a column with lemma frequency (see figure 2).

---

Insert Figure 2 about here

---

The definition of the different columns is:

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 word frequency per million words and has four digit precision. It is the measure researchers would preferably use in their manuscripts, because it is a standard measure of word frequency independent of the corpus size.
8. **Lg10WF.** This value is based on log10(FREQcount+1) and has four digit precision. Calculating the log frequency on the raw frequencies is the most straightforward transformation, because it allows 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(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). In addition, if one uses log(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 how many percent of the films the word appears, with four 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 only appears in 74 films..
10. **Lg10CD.** This value is based on log10(CDcount+1) and has four digit precision. As shown in Table 1, overall this is the best value to match stimuli on.

## Conclusion

In this paper 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 to the CELEX frequencies for Dutch, the SUBTLEX-NL frequencies are an improvement of almost 10 % in explained variance in reaction times. 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. While 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 word frequency 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 wordform frequencies.

At the same time, our article shows how easy it has become to make a good word frequency 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, while the compilation of the corpus on which the CELEX frequencies are based involved the lengthy process of scanning printed 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 (free software is also available on the internet, e.g., Paul Nation's Range Program, available at http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx).

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 word frequency database, we automatically processed thousands of these subtitles with relatively little effort. While it is impossible to determine the origin of each subtitle file, most subtitles available on the Internet appear to fall in 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 (so-called 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 or not.

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 word frequency database is only a statistical description of the subtitles. This is considered "fair use" of copyrighted material. However, in research benefiting from potentially illegal activity, ethical issues should also be considered. Much of the word frequency 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 paper. 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 benefits 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 word frequency 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.

Finally, 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 & Livesay (1998) and New et al. (2007) suggest that differences in quality between various frequency counts can already be detected with samples of a few hundred words spread over the entire frequency range. So, maybe it is not necessary to collect data for thousands of words. A typical one-hour experiment with some 1,000 words and nonwords may already be enough.

## References

Adelman, J.S., & Brown, G.D.A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115*, 214-227.

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science, 17*, 814–823.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*(2), 290–313.

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

Balota, D.A., Cortese, M.J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In *Abstracts of the 40th Annual Meeting of the Psychonomics Society* (p. 44). Los Angeles, CA: Psychonomic Society.

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., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (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: An International Quarterly*, 12, 91-116.

Brants, T., & Franz, A. (2006). *Web IT 5-gram Version 1*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

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*(4), 977-990.

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

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., & Brysbaert, M. (in press). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*.

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

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

Murray, W.S., & Forster, K.I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*, 721–756.

Murray, W.S., & Forster, K.I. (2008). The rank hypothesis and lexical decision: A reply to Adelman and Brown (2008). *Psychological Review, 115*, 240–252.

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

Shaoul, C. & Westbury C. (2009) A USENET corpus (2005-2009). Edmonton, AB: University of Alberta. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Stadthagen-Gonzalez, H., Bowers, J.S., & Damian, M.F. (2004). Age-of-acquisition effects in visual word recognition: Evidence from expert vocabularies. *Cognition, 93*, B11-B26.

Thorndike, E.L. & Lorge, I. (1944). *The teacher's word book of 30,000 words.* Teachers College, Columbia University, 1944.

Uit den Boogaart, P.C. (Ed.) (1975). *Woordfrequenties in Gesproken en Geschreven Nederlands.* Utrecht: Oosthoek, Scheltema and 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, G.-J., Canisius, S, and Daelemans, W. (2007). An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. Van Eynde (Eds.), *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting pp 99-114)*. Leuven, Belgium.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review, 15(5),* 971-979.

Yap, M.J. & Balota, D.A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*, 502-529.

Yap, M.J., Balota, D.A., Brysbaert, M., Shaoul, C. (submitted). Cora: A composite measure of word frequency.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science.

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

Table 1

*Percentages of variance in accuracy and RT explained by the different frequency measures.*

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

| | Accuracy (N = 12,964 ) | RT (N =11,386) |
|---|---|---|
| CELEX | | |
| Log | 13.4 | 25.9 |
| Log + log² | 15.4 | 26.2 |
| CELEXCD | | |
| Log | 18.8 | 25.2 |
| Log + log² | 19.1 | 26.8 |
| SUBTLEXWF | | |
| Log | 17.3 | 33.9 |
| Log + log² | 22.0 | 34.9 |
| SUBTLEXCD | | |
| Log | 20.6 | 35.1 |
| Log + log² | 25.3 | 35.2 |

Table 2

*Percentages of variance explained by lemma frequency together with word form frequency. Note.* Between brackets, the additional variance that lemma frequency explains relative to the variance explained by word form frequency alone).

| | Accuracy (N = 12,964 ) | RT (N =11,386) |
|---|---|---|
| CELEX | | |
| Log | 18.6 (5.2) | 26.9 (1.0) |
| Log + log² | 18.6 (3.3 ) | 27.0 (0.8) |
| CELEXCD | | |
| Log | 21.1 (2.3) | 25.5 (0.3) |
| Log + log² | 21.1 (2.0) | 27.4 (0.6) |
| SUBTLEXWF | | |
| Log | 26.7 (9.4) | 35.9 (2.0) |
| Log + log² | 26.9 (4.9) | 35.9 (0.9) |
| SUBTLEXCD | | |
| Log | 28.0 (7.4) | 36.1 (1.0) |
| Log + log² | 28.7 (3.4) | 36.2 (1.0) |

*Figure 1*. Lay-out of the SUBTLEX-NL.master file. A line starting with a word signifies a lemma (i.e., "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: WF of the word, CD of the word, WF of the word starting with a lowercase letter, and CD of the word starting with a lowercase letter.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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 2*: Lay-out of the SUBTLEX-NL file. See the text for the explanation of the column titles.

| Word | FREQcount | CDcount | FREQlow | CDlow | FREQlemma | SUBTLEXWF | Lg10WF | SUBTLEXCD | Lg10CD |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

## Footnotes

[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 .5%.

[3] We also calculated a different CD-measure in which we grouped all film sequels and episodes of a television series, based on 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] $r_{corr} = (2*r)/(1+r)$ (r = the split-half correlation; $r_{corr}$ = 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?).