Psicológica (2011), 32, 133-143.

SUBTLEX-ESP: Spanish word frequencies based on film subtitles

Fernando Cuetos*[1], Maria Glez-Nosti*, Analía Barbón*
& Marc Brysbaert**

**University of Oviedo, Spain; **Ghent University, Belgium*

Recent studies have shown that word frequency estimates obtained from films and television subtitles are better to predict performance in word recognition experiments than the traditional word frequency estimates based on books and newspapers. In this study, we present a subtitle-based word frequency list for Spanish, one of the most widely spoken languages. The subtitle frequencies are based on a corpus of 41M words taken from contemporary movies and TV series (screened between 1990 and 2009). In addition, the frequencies have been validated by correlating them with the RTs from two megastudies involving 2,764 words each (lexical decision and word naming tasks). The subtitle frequencies explained 6% more of the variance than the existing written frequencies in lexical decision, and 2% extra in word naming.

Word frequency, together with age of acquisition, is considered to be the most important variable in word comprehension and production: Words encountered often in life are processed more efficiently than words rarely encountered. Any study involving the perception or the production of words, be they on healthy individuals or on clinical samples (aphasia, Alzheimer's dementia, dyslexia, etc.), have to consider this variable. Therefore, researchers require good dictionaries that allow them to select words according to their frequency. Any language without a good word frequency measure is seriously disadvantaged when it comes to psycholinguistic research.

[1] This investigation was funded by grant MCI-PSI2009-09299 from the Spanish Government. We thank Rafael González Nosti for his help in the preparation of the corpus. Address for correspondence: Fernando Cuetos. Facultad de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003, Oviedo, Spain. Phone: 34985103283. Email: fcuetos@uniovi.es

Given the importance of word frequency, it is surprising to see how little attention language researchers have devoted to the quality of their measures. For instance, in a review of the literature Brysbaert and New (2009) noted that much frequency research in English is based on the Kucera and Francis (1967) frequency measure, despite the facts that it is derived from a small and dated corpus, and has been criticized repeatedly. Brysbaert and New (2009) argued that this state of affairs emerged because researchers simply took over the measure used by their predecessors without examining its criterion validity. Indeed, until recently the quality of frequency lists has been judged mainly on face validity. Two important factors were the size of the corpus and the diversity of the sources used.

In the last years, however, researchers have started to investigate the validity of the word frequency estimates empirically by correlating them with word processing times, in particular word naming times and lexical decision times (Balota et al., 2004; Brysbaert & New, 2009; Burgess & Livesay, 1998; Cai & Brysbaert, 2010; Ferrand, New, Brysbaert, Keuleers, Bonin, Meot, Augustinova, & Pallier, 2010; Keuleers, Brysbaert, & New, 2010; New, Brysbaert, Veronis, & Pallier, 2007; Zevin & Seidenberg, 2002). The picture emerging from these studies has not been entirely positive for the existing measures. The following shortcomings were noticed:

1. Frequency lists based on a corpus smaller than 10 million words correlate less with word processing times. This is particularly due to the inferior estimates of the low-frequency words.
2. At the same time, the gains due to the corpus size level end at 30-50 million words. It is not the case that a corpus of 1 billion words always gives better frequency measures than a corpus of 30 million words. From sizes of 30-50 million on, the language register on which the corpus is based becomes more important than the size of the corpus.
3. Book sources are interesting, but do not yield the highest correlations with word processing times, arguably because the edited language of books is not the language people are exposed to in daily life.
4. There are historical changes in word use, so that frequencies based on "old" (pre-1990) sources are less correlated with student performance in psychological experiments.

A first improvement in the frequency lists occurred when researchers started to use large corpora of unedited language from the internet (Burgess & Livesay, 1998). However, in recent years it has been discovered that an

even better source comes from subtitles. New et al. (2007) observed that French word frequencies taken from film and television subtitles predicted visual word recognition times better than the existing frequencies taken from written texts or from the Internet. The reason for this superiority was sought in the fact that written texts may not reflect the language used by people in daily life, because writers try to polish their language by using a more educated and refined register, which leads to an underestimation of many common words and an overestimation of words rarely used in everyday life. Written texts also tend to exaggerate lexical variation in order to avoid word repetition, which does not occur in spoken language. Finally, subtitles are closer to the language used by the students who usually take part in the laboratory experiments.

In the first study reporting subtitle frequencies, New et al (2007) found that these frequencies (based on a corpus of 52 million words) together with the length explained 50% of the variance in lexical decision times, 4% more than the variance explained by the best frequency measure taken from written texts. Further studies found even greater gains, because the popular written frequencies were not optimal: over 10% in English relative to the much used Kucera and Francis (1967) frequency list (Brysbaert & New, 2009), 8% in Dutch relative to the Celex frequencies (Keuleers, Brysbaert & New, 2010), and 15% for Chinese two-character words (Cai & Brysbaert, 2010; the difference for single-character words was much smaller).

Looking at the situation for the Spanish language, it is clear that the current frequency lists do not look optimal given the above developments. Despite the fact that Spanish is one of the most widely spoken languages in the world and has a thriving research community on word processing, there are only two word frequency lists, based on rather small corpora of published texts. The first list was published by Alameda and Cuetos (1995). It was built on a corpus of 2 million words coming from different types of texts written between 1978 and 1993. Fifty percent corresponded to novels, 25% to newspapers, 15% to literary essays, and 10% to scientific magazines. The second list is LEXESP, compiled by Sebastian, Martí, Carreiras, and Cuetos (2000). It is an extension of the Alameda & Cuetos list and is based on a corpus of 5,020,930 words of texts written between 1978 and 1995. Forty percent of the words come from novels, 30% from newspapers and the rest from essays and magazines.

To improve the existing Spanish situation, we (1) compiled a new frequency list, SUBTLEX-EXP, based on corpus of 41.5 million words from contemporary subtitles, and (2) we validated the various frequency measures by correlating them with word naming and lexical decision times

for a total of 2,764 words. We expected the new list to do better than the existing ones.

METHOD

Collection of the subtitle frequencies. A total of 41,577,673 words from movies and TV series, all after 1990, were collected. Most of the subtitle corpus was downloaded from the specialized websites www.argenteam.com, www.subdivx.com and www.solosubtitulos.com . Duplicate files and series and movies made before 1990 were removed. This resulted in a total of 3,523 movies (20,253,754 words) and 257 TV series (21,323,919 words). Twenty percent of the corpus (8,315,535 words) came from the years 1990-1999 and the remainder (33,262,138 words) from 2000 to 2009. The majority of the files came from English speaking films and series (American, British and Australian), with a total of 38,598,518 words. The Spanish speaking films and series made 1,222,111 words; the remaining 1,757,044 words came from movies made in non English or Spanish-speaking countries such as France, Germany, Russia, Brazil, Denmark, Norway and Italy.

All files were combined into one big corpus file, which was analyzed with a proprietary program to count the number of times each word appeared in the corpus. After removing the symbols, isolated letters, foreign or invented words, imitations of sounds, unusual proper names, numbers and the words observed only once in the corpus. The final corpus consisted of a total of 39,935,628 words.

Written text frequencies. The frequencies of the written texts were taken from the two existing dictionaries: Alameda and Cuetos (1995), and LEXESP (Sebastian et al., 2000).

Reaction times. The frequencies were validated by correlating them with the reaction times from a lexical decision (LD) experiment and a word naming experiment. The lexical decision times were taken from the mega-study of González-Nosti, Rodriguez-Ferreiro, Barbón and Cuetos (submitted). This study involved a total of 2,764 words, containing nouns, verbs and adjectives, between three and ten letters long, selected from LEXESP with an average length of 6.5 letters and 2.8 syllables. Compounds words, derivatives and inflected verb forms were not included. The 2,764 words were supplemented with 2,764 pseudowords formed by changing one letter of the words in such a way that the resulting

pseudoword was a legal Spanish letter string. The stimuli were divided into six blocks of 922 items. Blocks were presented in random order. Also the items in each block were presented in a random order. Words were presented and responses collected with the use of the DMDX software (Foster & Foster, 2003). Before each item, an asterisk was presented for 500ms in the center of the screen. Thirty-five undergraduates studying psychology from the University of Oviedo participated in the experiment. Participants were asked to complete all six blocks, one per day. No participant had reading problems.

The word naming times were taken from a mega-study with the same words ran by Davies et al (submitted) using the same lab and the same DMDX application. The number of participants in this experiment was 25. Responses latencies were registered by the DMDX software voice-key function. One experimenter sat with participants to record errors. Each of the 6 session lasted about 30 minutes.

RESULTS

The first analysis involved the calculation of Pearson correlations between the RTs of the 2,764 words (LD and word naming) and the three frequencies we had: A&C (Alameda & Cuetos), LEXESP, and SUBTLEX-ESP and length (number of letters). Table 1 shows the results. From this analysis it is clear that the correlation between the SUBTLEX frequencies and the word processing times are higher than those between the other two frequencies and the word processing times. As could be expected, LEXESP is slightly better than A&C (given that it is an extension of the latter). Figure 1 additionally gives a graphical display of the relationship between the SUBTLEX frequencies and the Naming and LD latencies.

Table 1. Correlations between frequencies (log transformed), word length, naming and lexical decision latencies.

	LD	Naming	SUBTLEX	LEXESP	A&C
Naming	.538				
SUBTLEX	-.638	-.411			
LEXESP	-.549	-.342	.783		
A&C	-.528	-.340	.741	.944	
Length	.370	.448	-.265	-.157	-.203

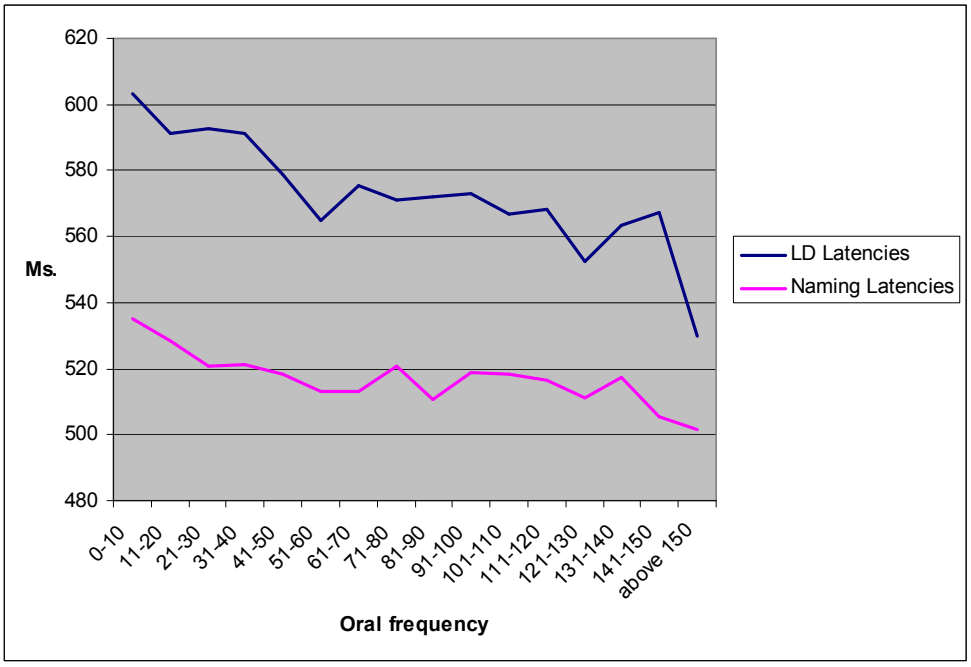

Figure 1. Partial effects of the SUBTLEX frequencies on the naming and lexical decision latencies.

The second analysis was a regression in which word length and word frequency were introduced as independent variables (see Table 2). Because there is a substantial correlation between word length and processing times (New, Ferrand, Pallier and Brysbaert, 2006), as shown in Figure 2, it is important to make sure that none of the correlations above are confounded by word length.

As could be expected from the correlations in Table 1, the SUBTLEX frequencies outperformed the existing frequencies based on written texts. The gain was nearly 7% for the lexical decision times and 2% for the naming times ($p < .001$). At the same time, the addition of Lexesp to SUBTLEX did not seem to make much difference (.9% in LDT and .3% in naming), certainly not if we take into account that the weights of the regression could be optimized in order to best predict the two datasets at hand (meaning that the weights would not be the same for a new set of stimuli or even a new sample of participants).

Table 2. Results of the regression analyses with the different frequency measures and word length. The first column of numbers shows the regression weights; the second column shows the percentage of variance accounted for by the model.

Lexical decision		
SUBTLEX	-39.66	Adjusted R^2 = .450
Length	14.80	
Lexesp	-33.30	Adjusted R^2 = .384
Length	19.27	
A&C	-30.20	Adjusted R^2 = .351
Length	17.53	
SUBTLEX	-19.65	Adjusted R^2 = .459
Lexesp	-7.02	
Length	15.47	
Word naming		
SUBTLEX	-18.92	Adjusted R^2 = .292
Length	21.96	
Lexesp	-16.98	Adjusted R^2 = .276
Length	24.65	
A&C	-15.58	Adjusted R^2 = .265
Length	23.71	
SUBTLEX	-8.86	Adjusted R^2 = .295
Lexesp	-3.91	
Length	22.26	

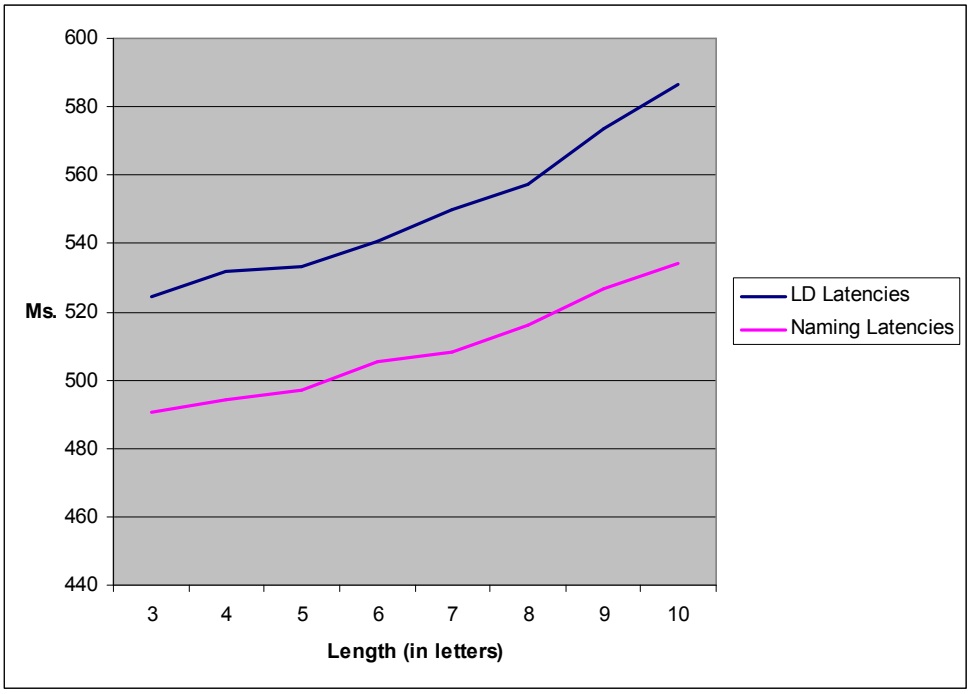

Figure 2. Partial effects of the word length on the naming and lexical decision latencies.

DISCUSSION

Inspired by the developments in other languages, we composed a list of Spanish word frequencies based on a reasonably large corpus (41.5 million words) of film subtitles. In line with previous studies, we found that the new word frequencies explained nearly 7% more of the variance in lexical decision times than the best existing measure based on written texts. The new index also explained 2% more of the variance in the word naming times. This is quite impressive, given that Spanish is a language with a very transparent orthography, as far as reading is concerned, which means that word naming can largely be based on non-lexical letter-sound correspondences and therefore, is less sensitive to word frequency.

Our previous research makes us confident that the better performance of the Spanish SUBTLEX frequency measure is not solely due to the size of the corpus on which it is based (41.5 million words against 5 million words for Lexesp). Part of the reason is that subtitles seem to be a better approximation of everyday word usage. In an unpublished study, Brysbaert and Keuleers compared the percentages of variance explained in Dutch

lexical decision times by the SUBTLEX frequencies (based on a corpus of 44M words) and newspaper frequencies (based on a corpus of 800M words). The SUBTLEX-frequencies significantly outperformed the newspaper frequencies, despite the fact that they were based on a much smaller corpus (tested on a sample of 14,000 words).

Ironically, without empirical validation it is unlikely that many researchers would have believed in the usefulness of subtitle-based word frequencies (the present authors included). Indeed, there are many reasons to believe why subtitles would be a less interesting language source. Subtitles are biased in various ways (the topics covered, the American dominance) and are not always a 100% accurate translation of what is said. There are also considerable differences in the extent to which people from various countries are used to reading subtitles (e.g., subtitling is very frequent in the Dutch-speaking countries, but less so in Spanish-speaking or English-speaking countries). Still, in all languages tested, subtitle frequencies outperform text-based word frequencies. Post hoc, the following arguments can be made. For a start, the situations depicted on the screen may be more representative of everyday life (interactions with objects and other people). Second, students may watch more television than they read books or newspapers and may be more familiar with "film language" than with "book language". Finally, it seems plausible that visual word recognition depends not only on the number of times the word has been seen or produced in print, but also on the number of times the word has been heard and used in speech.

Subtitle frequencies correlate roughly .70-.80 with written frequencies (see also table 1). It will be interesting to investigate what differences between both types of frequencies are responsible for the better prediction of word processing times. In the meantime, our results in various languages indicate that researchers are advised to control their stimuli on subtitle frequencies more than on written frequencies, if they want to use the best possible index of word frequency.

Availability

To give easy access to the new frequency measure, we have made a SUBTLEX-ESP text file and an Excel file of the word list. These files contain information about the words that were observed more than once in the corpus (the other "words" usually are typos and add unnecessary clutter to the list). There are 4 columns with self-explaining headings:

- Word
- Frequency count (on a total of 41,577,673 million words)

- Frequency per million: this is the variable easiest to interpret as it is independent of the size of the corpus (i.e., can easily be compared to the values of other corpora). This is the variable to be reported in manuscripts.
- Log10(frequency count + 1): this is the variable to use when one wants to select or match stimuli on frequency. By using the frequency count rather than the frequency per million, we are not losing any information by adding 1 (the latter is needed to have a log10 frequency value of 0 for the words not encountered in the list).

The SUBTLEX-ESP files are available as supplementary files to this article on the Psicologica website. They can also be found on the Internet at: http://www.unioviedo.es/neurociencias_cognitivas/data/

RESUMEN

SUBTLEX-ESP: Frecuencias de las palabras españolas basadas en los subtítulos de las películas. Estudios recientes han mostrado que las estimaciones de frecuencia de las palabras obtenidas de los subtítulos de películas y series de televisión predicen mejor los resultados de los experimentos de reconocimiento de palabras que la tradicional estimación de frecuencia basada en libros y periódicos. En este estudio presentamos una lista de frecuencias de las palabras basada en los subtítulos para el español, uno de los idiomas más extendidos en el mundo. La frecuencia de los subtítulos fue obtenida a partir de un corpus de 41 millones de palabras tomadas de películas y series de televisión (de entre los años 1990 y 2009). Además, las frecuencias fueron validadas al correlacionarlas con los tiempos de reacción de dos megaestudios realizados sobre 2764 palabras cada uno (con las tareas de decisión léxica y lectura en voz alta). La frecuencia de los subtítulos explicaban un 6% más de la varianza que las frecuencias escritas en la tarea de decisión léxica y un 2% extra en lectura en voz alta.

REFERENCES

Alameda, J.R. & Cuetos, F. (1995) *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Oviedo, Servicio de Publicaciones Universidad de Oviedo.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). *The CELEX lexical database, Release 2 (CD-ROM).* Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Balota, D.A., Cortese, M.J., Sergent-Marshal, S.D., Spieler, D.H. & Yap, M.J. (2004) Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General,* 133, 283-316.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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 (see also http://expsy.ugent.be/subtlexus).

Burgess, C. & Livesay, K. (1998) The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavioral Research Methods, Instruments & Computers.* 30, 272-277.

Cai, Q. & Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *Plos One*.

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.

Davies, R., Barbón, A. & Cuetos, F. (submitted) *Reading in transparent orthographies relies flexibly on lexical and sub-lexical knowledge: A mega-study of list composition effects in Spanish.*

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488-496.

Foster, K.I. & Foster, J.C. (2003) DMDX: A window display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers*, 35, 116-124.

Gonzalez-Nosti, M., Rodríguez-Ferreiro, J., Barbón, A. & Cuetos, F. (submitted). *Lexical decision in Spanish: Data from a mega-study.*

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

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

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

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin and Review, 13*, 45-52

Sebastian, N., Martí, M.A., Carreiras, M. & Cuetos, F. (2000) *LEXESP: Léxico informatizado del español*. Barcelona, University of Barcelona Press.

Thorndike, E.L. & Lorge, I. (1944) *The teacher's Word book of 30.000 words*. Teachers College, Columbia University.

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

(Manuscript received: 10 February 2010; accepted: 5 July 2010)