SUBTLEX- AL: Albanian word frequencies based on film subtitles

Rrezarta Avdyli, Fernando Cuetos

Abstract

Recently several studies have shown that word frequency estimation based on subtitle files explains better the variance in word recognition performance than traditional words frequency estimates did. The present study aims to show this frequency estimate in Albanian from more than 2M words coming from film subtitles. Our results show high correlation between the RT from a LD study (120 stimuli) and the SUBTLEX- AL, as well as, high correlation between this and the unique existing frequency list of a hundred more frequent Albanian words. These findings suggest that SUBTLEX-AL it is good frequency estimation, furthermore, this is the first database of frequency estimation in Albanian larger than 100 words.

Keywords*: word frequency, subtitles, Albanian, word recognition*

Word frequency is one of the most important psycholinguistic variables. Any study involving word processing, as memory, reading, writing, speaking, and all other basic psychological processes have to consider this variable, either in normal samples (adults or children), or patients (aphasia, Alzheimer's dementia, dyslexia, Parkinson´s disease). For this reason, researchers require a good estimation of frequency, which allows them to select words for their experimental or clinical purposes. Given the importance of this variable and its implication in many fields of study, nowadays there are frequency dictionaries in many languages.

Data collection traditionally have been done by choosing a good number of written sources which could represent in the best way the language at the moment of constructing the new measure.

The following dictionaries are some examples of these confections, English (Baayen, Piepenbrock, and Gulikers, 1995, Kucera and Francis, 1967, Thorndike and Lorge, 1944), Italian (De Mauro, Vedovelli, and Voghera, 1993), French (Imbs, 1971, New, Pallier, Brysbaert, and Ferrand, 2004), Hungarian (Furedi and Kelemen, 1989), or Spanish (Alameda and Cuetos, 1995; Juilland and Chang-Rodríguez, 1964; Sebastian-Gallés, Martí, Carreiras, and Cuetos, 2000).

At present, the Internet is available by a simple mouse click and it contains hundreds of billions of words and can be used for all manner of language research, which facilitates the way of obtaining larger corpora than from published texts, which in some cases have to be scanned.

The first frequencies corpus made by using online Usernet groups was done by Burgess and Livesay (1998) called Hyperspace Analog to Language (HAL) with more than 130 million words. In these groups, the Internet users participate in discussions on a variety of topics without much supervision or editing. Burgess and colleagues showed that HAL accounted for more variance in lexical decision times than the KF frequencies (Kučera and Francis, 1967). A similar finding was reported by Balota et al., (2004) who subsequently recommended the HAL frequencies for further research, also because of its higher correlations with behavioral data. According to these findings, other researchers have proposed the outcome of the Internet search engines as another interesting estimate of word frequency (Blair, Urland, & Ma, 2002; New, Pallier, Brysbaert, & Ferrand, 2004). So, the Internet was one of the first sources that yielded better frequency measures. In addition, word use on the Internet is more varied than the formal language used in edited texts.

However, in recent years, several 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 written texts. The first authors who observed this frequency effect were New, Brysbaert, Veronis, and Pallier (2007) in French, compiling a corpus of 52 million of French words coming from 9,474 different films and television series. They found that subtitle frequencies explained more of the variance in lexical decision reaction times (RTs) than written frequency measures did. As a consequence, they added a subtitled frequencies to their third version of Lexique project (see www.lexique.org).

Inspired by this initiative of using internet files as possible source of information, there were created a number of corpora based on film and television subtitles. Brysbaert and New (2009) replicated the New et al., findings in English and found that their subtitle frequency measure did better than Internet-based frequencies in predicting word naming and lexical decision performance. They hypothesized that this was because film and television language approximates everyday word use better than written sources do.

Brysbaert and New (2009) also reported two important aspects of word frequency estimates for experimental research of word processing. The first one is the size of the corpus from 16 to 30 million words is enough for a good frequency estimation. The second one is the representativeness of the materials on which the norms are based. The more natural the language use is, the better the frequency norms account for lexical decision times.

The same procedure has been performed in Dutch by Keuleers, Brysbeart and New (2010), building SUBTELEX-NL by 40M words. This new measure also in Dutch demonstrates the superiority of word frequency based on subtitles to predict RT´s: 8% more than the standard dictionary of CELEX (Baayen, et al., 1995; Baayen, Piepenbrock and van Rijn, 1993) shown by a lexical decision study.

In a Mandarin Chinese, Cai and Brysbaert, (2010) constructed a subtitled based frequency dictionary by compiling a corpus of 46,8M one character words and 33.5M words of two or more characters. To show its validity, they compared it with other existing written frequency measures, such as the CCL (Center for Chinese Linguistics), LCSMCS dictionary (Liu, Shu, and Li, 2007) and the Lancaster Corpus of Mandarin Chinese (LCMC's) (Xiao, Rayson, and McEnery, 2009) and proved the superiority of Subtlex-CH-WF in predicting RT´s measured by a lexical decision task.

Cuetos, González-Nosti, Barbon and Brysbaert, (2011) reported essentially the same findings in Spanish. Their subtitled frequency measure SUBTLEX-ES based on a 40M words, explained nearly 7% more variance in lexical decision times, and 2 % more of the variance on the naming times, than the existing standard written frequency dictionaries LEXESP (Sebastian-Gallés, et al., 2000).

The Modern Greek measure was made by Dimitropoulou et al., (2010) consisting of more than 27 million words. As it was shown in the studies above, Dimitropoulou et al., (2010) also found that SUBTLEX-GR frequency estimates outperformed the traditional frequencies GreekLex (Ktori et al., 2008) in two different visual word recognition experiments. In Experiment 1, SUBTLEX-GR explained more than 28% of the reaction time variance. Whereas in Experiment 2, where the Greek names of the Snodgrass and Vanderwart (1980) picture set were presented, SUBTLEX-GR explained up to 48.6 of reaction time and 30.5% of error rate variance.

These studies indicate that subtitle-based corpora of frequency estimates in the different languages predict better the reading behavior than frequency estimates collected by written sources (New et al., 2007; Brysbaert and New, 2009; Cai and Brysbaert, 2010; Keuleers et al., 2010, Cuetos et al., 2011, Dimitropoulou et al., 2010).

Inspired by these findings and by the lack of any frequency estimation measures in Albanian, and because of the apparent facility of access on the

subtitle files, we decided to compile a subtitle-based corpus. In Kosovo, Albanian speaking region, and even in Albania, it is used to subtitle almost all the foreign audiovisual material except for Disney and children's programs that are dubbed, as it was seen for the Modern Greek (Dimitropolou et al., 2010). Thus, we may expect that the Albanian subtitled based frequency measure would be a good representative estimation of the everyday language usage of young Albanian speaking adults.

The explanation about the lack of any scientific assessment materials in Albanian language it can be due to the socio-cultural and historical circumstances in the last two decades in Balkans and in concrete the context of Kosovo, mainly Albanian speaking region. For this reason, the development of scientific activity it has been limited and the elaboration of research materials as the following occupied a secondary place. However, after twelve years of the Kosovo war, the interest about the scientific fields has increased, nonetheless the lack of assessment materials difficult this labor. Although much work has been done it is still needed much more investigation, especially related with psycholinguistic nature.

Considering this, it seems necessary to create new assessment or support research measures in many fields, but as we are interested in reading and writing process, a word frequency estimates dictionary it is prerequisite to follow researching in this lines.

In the literature reviewed until know we din not find any frequency estimation done in Albanian except a list of a 100 more frequent Albanian words (Spahiu, 2010).

Materials and Methods

Corpus collection

Most of the subtitle corpus was downloaded from specialized websites as www.opensubtitles.com, www.all4divx.com, www.subs.to and www.findsubtitles.com. The vast majority of files were translations made from English, but there were also movie files from other European languages as German, Italian, and French. Duplicate files were removed. This resulted in a total of 426 movies from the years 1995 to 2009. There were no whole television series subtitled in Albanian, therefore, they are not included. 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, proper names, numbers and the words observed only once in the corpus, the final corpus consisted of a total of 2,666,025 words. Eighty per cent of the corpus came from English-language films and the rest from other languages.

We are aware about the small size of our corpus, but in spite of an extended use of the subtitles in Albanian, we did not find many files available to download during the period of the data collection (September- December 2009). Probably this is due to the majority of films and television series subtitles are owned by the State and Private TV Productions. This is the reason why we have just downloaded the subtitle files which were 'free' and without any previous requirement of copyright (because they were uploaded by Internet individual users or group users).

Validation of the Subtlex-Albanian

Given that we could not find any experimental data with the lexical decision task or naming in Albanian, probably as a consequence of a lack of this measure in Albanian also because anyone is carrying psycholinguistic, educational or clinical data using word frequencies, we decided to run a small visual lexical decision validation experiment of 120 words and 120 pseudowords.

As visual lexical decision is particularly sensitive to word frequencies (Balota et al., 2004; Brysbaert & New, 2009; Cortese & Khanna, 2007; Yap & Balota, 2009), then it is a particularly informative task to validate a frequency measure. The lexical decision task is commonly used when studying word processing in psycholinguistics. During the performance of this task, participants have to decide as fast as possible if a stimulus is a word or a pseudoword. The lexical decision task, which contains 240 stimuli, 120 words (frequency average 11.50 x million, length average 5.88) and 120 pseudowords (length average 5,883) formed by changing one letter (in the beginning, middle or final) of the words in such a way that the resulting pseudoword was a legal Albanian letter string. All the stimuli were nouns, and there were not included derivatives, compound words, and inflected verb forms. The reading times were measured by using the SuperLabPro softwere. The stimuli were presented centrally in computer screen in a black lowercase letters (System, 40 pts, bold) against white background, and before any trial a fixation point for 500ms was presented. The experiment started with the instructions in the screen, then two trials with one of each kind of stimuli, and finally the 240 experimental stimuli. The stimuli stayed on the screen until the participant made a response. The participants were instructed to make lexical decisions by pressing as fast as possible one of the two keys on the keyboard. The experiment took about 15 minutes. All the participants (fifty five) were students of the University of Prishtina, Kosovo.

Results

Pearson's correlations were run between the 120 words used in DL experiment, with word length and word frequency of SUBTLEX-AL. Both variables were statistically significant, word frequency and word length, although word length ($r = 0.457$, $p < 0.01$) correlated more than word frequency ($r = 0.186$, $p < 0.05$).

We also correlated the frequency measures of the SUBTLX-AL and the unique written sources Albanian corpora of 100 more frequent words. Matching the frequencies in both measures, we found a high correlation ($r = 0.667$, $p < 0.01$).

Discussion

Our main objective of this investigation was the creation of a first frequency measure in Albanian language, which would serve as for further investigations, as mentioned in the introduction. Influenced by the developments of frequency measures in other languages from the subtitle files, we compiled an Albanian corpus as larger as we could, in a limit of years from 1995-2009. In spite of the Brysbeart and New (2009) profess that the optimal corpus size for a reliable estimation of low frequency words should be at least of a 16 million words whereas for high frequency words a corpus size of one million reaches a stable level, and even though our small corpus, we carried out the analysis because of the necessity of this measure in Albanian.

In the Pearson´s, correlation between the 100 more frequent words matched in Albanian it can be seen that with the increasing frequency of traditional written sources the SUBTLEX-AL also increases.

Comparing this frequency measure with the predecessors of this kind of instruments in the other languages, we are aware about the limitation that this presents, but his utility as a unique frequency measure it will be high, accounting that it is the first one in Albanian language.

Availability

The supplemental materials of the full SUBTLEX-AL database may be downloaded from http://www. unioviedo.es/neurociencias_cognitivas/ SUBTLEX-AL/

These file contain information about the words that were observed more than once in the corpus. There are 4 columns with self-explaining headings:

- Word
- Frequency count (on a total of 2.666.025 million words)
- Frequency per million: (21.715) this is the variable to be reported in manuscripts because of its easiest interpretation, as it is independent of the size of the corpus.

- Log10 (frequency count + 1): this is the variable to use when one wants to select or match stimuli on frequency.

Acknowledgments

This investigation was funded by grant MCI-PSI2009-09299 from the Spanish Government.

References

Alameda, J., & Cuetos, F. (1995). Diccionario de frecuencias de las unidades lingüísticas del castellano.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2)[cd-rom]. *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]*.

Baayen, R., Piepenbrock, R., & Van Rijn, H. (1993). The celex lexical database (cd-rom). *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA*.

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.

Blair, I., Urland, G., & MA, J. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers, 34*(2), 286-290.

Brysbaert, M., & New, B. (2009). Moving beyond Ku era 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.

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. *Behavior Research Methods Instruments and Computers, 30*, 272-277.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PloS one, 5* (6).

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.*

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

De Mauro, T., Vedovelli, F., & Voghera, M. (1993). *Lessico di frequanza dell´italiano parlato*. Milano: Etaslibri.

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

Füredi, M., & Kelemen, J. (1989). A mai magyar nyelv szépprózai gyakorisági szótára [A frequency dictionary of the literary language of Hungarian]. *Budapest, Hungary: Akadémiai Kiadó*.

Imbs, P. (1971). Études statistiques sur le vocabulaire français: Dictionnaire des fréquences. *Vocabulaire littéraire des XIXe et XXe siècles*.

Juilland, A., & Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*: Mouton.

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

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

Ktori, M., and Pitchford, N. J. (2008). Effect of orthographic transparency on letter position encoding: a comparison of Greek and English monoscriptal and biscriptal readers. *Lang. Cogn. Process.* 23, 258–281.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers, 36* (3), 516.

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

Thorndike, E., & Lorge, I. (1944). The teacher's word book of 30,000 words.

Sebastián-Gallés, N., Martí, M., Carreiras, M., & Cuetos, F. (2000). LEXESP: Una base de datos informatizada del español. *Universitat de Barcelona, Barcelona*.

Snodgrass, J. C., and Vanderwart, M. (1980). A Standardized Set of 260 Pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol. Hum. Learn.* 6, 174–215.

Spahiu, A. (2010). 100 fjalët më të shpeshta në gjuhën shqipe. Retrieved from http://www.shkenca.org/content/view/140/27/

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

Xiao, R., Rayson, P., & McEnery, T. (2009). A Frequency Dictionary of Mandarin Chinese.