frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 16 December 2010
doi: 10.3389/fpsyg.2010.00218

Subtitle-based word frequencies as the best estimate of reading behavior: the case of Greek

Maria Dimitropoulou[1,2], Jon Andoni Duñabeitia[1], Alberto Avilés[1,2], José Corral[1,2] and Manuel Carreiras[1,3,4]*

[1] *Basque Center on Cognition, Brain and Language, Donostia, Spain*
[2] *University of La Laguna, Tenerife, Spain*
[3] *Ikerbasque, Basque foundation for Science, Bilbao, Spain*
[4] *Departamento de Filología Vasca, University of the Basque Country, Bilbao, Spain*

Edited by:
Marc Brysbaert, University of Ghent, Belgium

Reviewed by:
Fernando Cuetos, Universidad de Oviedo, Spain
Emmanuel Keuleers, Ghent University, Belgium

****Correspondence:***
Manuel Carreiras, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 20009-Donostia, Spain.
e-mail: m.carreiras@bcbl.eu

Previous evidence has shown that word frequencies calculated from corpora based on film and television subtitles can readily account for reading performance, since the language used in subtitles greatly approximates everyday language. The present study examines this issue in a society with increased exposure to subtitle reading. We compiled SUBTLEX-GR, a subtitled-based corpus consisting of more than 27 million Modern Greek words, and tested to what extent subtitle-based frequency estimates and those taken from a written corpus of Modern Greek account for the lexical decision performance of young Greek adults who are exposed to subtitle reading on a daily basis. Results showed that SUBTLEX-GR frequency estimates effectively accounted for participants' reading performance in two different visual word recognition experiments. More importantly, different analyses showed that frequencies estimated from a subtitle corpus explained the obtained results significantly better than traditional frequencies derived from written corpora.

Keywords: frequency estimates, cultural variations, subtitles, word recognition

INTRODUCTION

The extensive evidence showing that word frequency dramatically influences most aspects of language processing has led psycholinguists to control for this variable in their experimental materials. Hence, word databases presenting frequency estimates are a widely used tool in languages subject to psycholinguistic research (e.g., English: Baayen et al., 1995; French: New et al., 2001; Spanish: Sebastián-Gallés et al., 2000). Most commonly, frequency databases have been obtained by counting the number of appearances of the words within corpora compiled from written texts mostly found in books and/or newspapers and periodicals of diverse topics. However, despite their frequent use in research and the effort to include recent texts from various sources, these corpora present two critical disadvantages. First, very often the topics chosen are not related to everyday life, and second, the linguistic style they tap onto is not representative of the most frequent usage of language. In an effort to overcome these limitations a number of recently created word corpora have used subtitles of films and television series to find samples of written input that are more representative of the one that individuals are usually exposed to. The present study describes the creation and the subsequent validation of a word frequency database entirely based on subtitle files available in the Internet for Modern Greek, and shows how controlling for Greek subtitle-based frequencies appears to be more suited than controlling for regular text-based frequencies.

Burgess and Livesay (1998) were the first authors who turned to the Internet to find the raw materials for a word frequency database. These authors gathered texts from online Usernet groups, and created the Hyperspace Analog to Language (HAL) with more than 130 million words. Following this initiative, a number of corpora have been created based on another type of digitalized texts available online: film and television subtitles. New et al. (2007) collected almost 9,500 French subtitle files from films and television series to create a word corpus of more than 50 million words. They validated the new subtitle-based corpus by comparing it to a French written corpus (Lexique 2; New et al., 2004) and also to a spoken language corpus (Corpus de Référence du Français Parlé, CRFP; Equipe DELIC, 2004). Despite the fact that subtitles are written transcripts of oral language, the subtitle-based frequency counts correlated more highly with those of the written corpus than those of the spoken corpus (0.73 and 0.85 for the spoken and written language corpora, respectively). More importantly, New and colleagues showed that the subtitle-based frequency norms effectively predicted the reading behavior of native French readers by calculating the proportion of reaction time variance explained by the new corpus based on lexical decision times from a set of 234 French words taken from Bonin et al. (2001, Experiment 3) and from another set of 240 French words. The authors found that the new subtitle-based frequency measures explained more variance than the spoken language corpus, on the one hand, and as much variance as the written language corpus, on the other. In other words, New et al. showed that word frequencies obtained from a subtitle-based corpus were as good as the traditional written word frequency estimations and better than oral frequency estimations at predicting reading behavior.

Given the outstanding performance of the French subtitle-based frequencies in predicting French reading behavior, similar subtitle-based databases have been created for American-English ($SUBTLEX_{US}$; Brysbaert and New, 2009), Dutch (SUBTLEX-NL; Keuleers et al., 2010) and Mandarin Chinese (SUBTLEX-CH; Cai and Brysbaert, 2010). Each of these corpora was based on eight to nine thousand subtitle files and included more than 30 million

words. As Brysbaert and New showed, corpora of this size (ranging from 16 to 30 million words) can provide reliable estimates for both high and low frequency words. In order to empirically validate the new frequency estimates these studies performed multiple regression analyses on lexical decision and word naming data ($SUBTLEX_{US}$ and SUBTLEX-CH) and compared the proportion of variance explained by the subtitle-based corpora to that explained by existing corpora based on other written and spoken sources. For the $SUBTLEX_{US}$ validation these data were taken from the 40,000 words on which the Elexicon has lexical decision data and from the 2,046 words on which the Elexicon provides word naming data (Balota et al., 2004, 2007). For the SUBTLEX-CH validation the lexical decision data were taken from an experiment with 400 words and from two previously performed Chinese lexical decision studies with 200 words. Finally, for the SUBTLEX-NL validation, Keuleers et al. collected lexical decision data on 14,037 mono- and disyllabic Dutch words. Across the three studies, the results of the multiple regression analyses confirmed the validity of the subtitle-based corpora in predicting the performance of the participants in classical reading tasks and, in some cases the new frequency measures outperformed some of the up to now most widely used frequency norms. For instance in English $SUBTLEX_{US}$ frequencies explained more variance than the Kučera and Francis (1967) frequency norms and this was also the case in Dutch, with SUBTLEX-NL explaining 10% more variance than the CELEX Dutch frequency estimates (Baayen et al., 1993, 1995). Hence, similar to what was found in French, American-English, Dutch, and Chinese subtitle-based frequencies also provided an effective estimation of reading performance. The higher correlation with word recognition tasks suggest that the linguistic style presented in subtitled movies or television series such as those used in these studies is highly representative of the language experience of young adults.

Studies that have explored the validity of subtitles as a source to extract word frequencies have also pinpointed two further noteworthy effects. Brysbaert and New (2009) and Keuleers et al. (2010) showed that the lemma frequency, defined as the summed frequency of all the inflected forms of a particular word, was a less effective predictor of reading performance than word form frequency, the frequency of the distinct word forms (e.g., play, plays etc), suggesting that experimental item selection could be mainly based on the latter. Furthermore, both studies as well as the study by Cai and Brysbaert (2010) found that the frequency measure which best accounted for the reading performance was the "contextual diversity" (CD) and not the traditionally defined word frequency (i.e., number of appearances of a word within the corpus). CD was first defined by Adelman et al. (2006) as the number of documents a word appears in a corpus. In the case of subtitle-based corpora CD is defined as the number of films or television shows a word is encountered in.

An interesting point regarding these sets of findings is that the predictive capacity of the frequencies based on subtitles seems to be affected by the extent of exposure to subtitle reading. Even though the subtitle-based word frequency databases effectively accounted for the reading behavior in all the languages examined so far, the SUBTEX-NL database was the one to provide a more marked additional benefit in predicting the lexical decision performance of Dutch readers as compared to the most widely used Dutch text-based word frequency database, CELEX (10% additional variance in reaction times; Keuleers et al., 2010). Critically, out of the countries for whose official languages subtitle-based word corpora have been created, only Belgium has a strong subtitle-reading tradition. Note that in France (as in a large number of European countries) foreign films as well as most foreign television programs are dubbed not subtitled. This is also the case in China, where there is also the limitation of more marked cultural differences with the countries of origin of most of the films and television shows to which the subtitle files used corresponded (mostly European or American). Similarly, since most of the viewing public in the USA watches English language films and programs produced by their huge domestic film and TV industry, there is no need for any translation (subtitling, dubbing or voice-over). Hence, considering that subtitle-based corpora are able to predict reading behavior better than text-based corpora in populations with limited exposition to subtitles, it should therefore be expected that in populations in which subtitle reading is much more extended, word frequency norms taken from subtitles could be better predictors of reading performance than word frequency estimates based on written corpora (as shown in Dutch). The present study aims at further exploring this issue in a language whose users are constantly exposed to subtitle reading: Modern Greek.

There are several reasons to believe that subtitle-based frequency norms could provide a valid estimate of word frequencies for Modern Greek. First, in Greece almost all foreign audiovisual material, with the exception of children's films and television programs, is broadcast subtitled. Moreover, considering that the national film production is quite limited, Greeks and Cypriots are mainly exposed to subtitled foreign films. This, of course, represents an important difference as compared to other countries with a relatively larger film industry (e.g., France, Spain, and Germany). Second, due to historical-linguistic factors regarding the recent evolution of the Greek language, there is a notable discrepancy between the linguistic style used by young adults and that found in most written texts. From its well-studied ancient version, the official language of the Greek state changed to a previous, transitional version of Modern Greek, before the establishment of the currently used Modern Greek in 1976 (see Mackridge, 2009, for a detailed description of the differences between the transitional version of the Greek language and Modern Greek). Note, however, that this cultured vocabulary is rarely used in everyday written or oral communication and even less by young adults (Kazazis, 2001). Thus, within a corpus based on subtitles, terms from this rarely used vocabulary will appear in a limited fashion, compared to a corpus based on newspaper articles and books. Considering this, it seems plausible that a Modern Greek subtitle-based frequency database will be more representative of the language usage of young adults. In order to put this hypothesis to the test we first created SUBTLEX-GR, a word frequency corpus entirely based on Greek subtitle files available on the Internet, with a total of more than 27 million tokens. After assembling the corpus, two lexical decision experiments were conducted aiming to examine the psychological validity of the new subtitle-based frequencies and to compare their predictive value to that of an existing written frequency database of Modern Greek.

Due to the increasing amount of psycholinguistic research that is being conducted in Modern Greek (e.g., Orfanidou and Sumner, 2005; Ktori and Pitchford, 2008, 2009; Dimitropoulou et al., in press-a,b; Georgiou et al., 2010), efforts have started to focus on providing appropriately controlled tools and norms (e.g., Ktori et al., 2008; Dimitropoulou et al., 2009; Protopapas and Vlahou, 2009). So far, the only published frequency norms for Modern Greek correspond to GreekLex, a database recently created by Ktori et al. (2008) containing more than 35,000 different entries taken from a 47 million-word corpus of written texts (Hellenic National Corpus; HNC, Mikros et al., 2001; Hatzigeorgiu et al., 2005). The earliest sources on which the HNC is based date from 1990 and were mainly gathered from newspapers (61.3%), books (9.4%) periodicals (5.9%) and other written texts (23.1%) covering a relatively large variety of topics. GreekLex presents word (and lemma) frequency estimates for a total of 29,900 non-inflected Greek word forms, extracted from the 2007 version of the HNC corpus. For these words the authors provided word length values and they also calculated the number of orthographic neighbors *(N;* Coltheart et al., 1977). Finally, Ktori et al. also provided transposition, addition, and deletion neighbor counts (see Davis et al., 2009; Duñabeitia et al., 2009b,c). The present study presents a word frequency database for Modern Greek based on Greek subtitle files available on the Internet and examines whether these subtitle-based frequency counts offer valid estimates of the reading performance of Greeks, for whom subtitles represent a remarkable portion of their daily reading material.

THE SUBTLEX-GR CORPUS

A total of 6,032 Greek subtitle files from 5,508 television series and films provided by the www.opensubtitles.org website constituted the raw material for SUBTLEX-GR. Recent evidence from studies examining context diversity effects suggest that a sample of this size is large enough to capture the different contexts in which a word can appear (e.g., Adelman et al., 2006; Brysbaert and New, 2009). From these files we identified a total of 27,761,198 space-separated tokens that corresponded to 597,540 different types. The space delimited tokenization is considered to be a valid way to identify Greek words since Greek is one of the few languages in which multiple-word compounds are extremely scarce. This is possibly due to the fact that Greek is a language with a rich concatenating morphology and thus, most compounds can be represented in a single word (see Ralli, 2007; Koliopoulou, 2008). For these types we calculated the number of occurrences and the number of different contexts they appeared in CD. For a more detailed description of SUBTLEX-GR along with all the information regarding the different versions of the corpus that are available to researchers, see the Appendix.

COMPARING SUBTLEX-GR TO GReekLex

The quality of the new SUBTLEX-GR corpus and of the frequency counts provided was first examined by comparing SUBTLEX-GR to GreekLex (Ktori et al., 2008). Given that the entries listed in GreekLex resulted from cross-validating the HNC with a dictionary, we compared it to a restricted version of SUBTLEX-GR which resulted from cross-checking the entire corpus to a Modern Greek spell-checker (see the Appendix). This comparison aimed to identify possible differences between the two databases resulting from the different raw materials of the two corpora and to test whether the content of SUBTLEX-GR is representative of the Modern Greek language.

In order to compare the frequency values reported in the two databases, the distributions of the percentages of the summed frequency per million (the total frequency per million words in SUBTLEX-GR_restricted.txt) in terms of word length were plotted. This allowed for a direct comparison between the frequency distribution of SUBTLEX-GR and GreekLex, despite the large difference in the size of the two databases (see **Figure 1**). In general terms, the two distribution curves were clearly comparable. In both databases 3-letter words had the highest summed frequency, and 1–4 letter words accounted for the largest part of the summed frequencies. One relatively salient difference was that in GreekLex, 1-letter words represent 9.4% while in SUBTLEX-GR they represent only 2.7% of the total frequencies. This is due to the fact that apart from the 1-letter words, GreekLex also includes as separate entries the letters of the Greek alphabet. In contrast, given that the individual letters are not existing Greek words, they are not included in the restricted version of SUBTLEX-GR[1]. The other difference was found for the 5–7 letter words. Although their distribution was similar for the two databases, these words appeared to be more frequent in SUBTLEX-GR (a 10.7% difference). This difference probably reflects the fact that short and common words are more representative of oral language. Furthermore, these words could also be preferentially used in subtitles because they are easier to perceive, thus aiding to match the timing of the subtitle reading process to the rapid succession of the film and television scenes.

In order to further compare the two databases, the common words across the two databases were identified (see also New et al., 2007 for a similar procedure). Out of the 19,955 common words, the 14,527 words with a frequency per million higher than 0 (with one decimal precision, in line with the GreekLex lay out) were included

[1]The number of occurrences of each letter of the Greek alphabet as single-letter space-separated entries (not embedded within letter strings) can be found in the full version of SUBTLEX-GR. However, given that this count does not represent the number of times each letter appears within the corpus, researchers should bear in mind that this values do not represent the letter frequency. The actual letter frequencies can be found in a separate file called SUBTLEX-GR_letters.txt, at www.bcbl.eu/databases/subtlexgr (see Appendix, for further information).

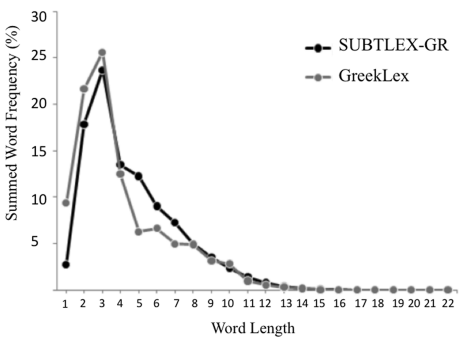

FIGURE 1 | Distribution of the percentages of summed word frequency as a function of word length (measured in number of letters) for the GreekLex and the SUBTLEX-GR entries.

in a correlation analysis that showed that the two frequency per million counts were highly correlated, $r = 0.835$ $p < 0.01$ (see also New et al., 2007 for a similar correlation between subtitle-based and written text-based corpora). In order to localize the origins of the apparently very small divergence between the frequency estimates of the two databases, we identified the 100 words whose frequency per million differed most across the two databases (see also New et al., 2007; Brysbaert and New, 2009). Calculating the ratio of the SUBTLEX-GR frequencies to the GreekLex frequencies, we identified the 50 words with much higher frequency in SUBTLEX-GR than in GreekLex. These were mainly words found in everyday oral interactions (e.g., συγνώμη [sorry], ενάξει [ok], έλα [come]) and words from street jargon or criminal/police contexts (e.g., μπάσος [cop], διατάζω [order]). The calculation of the ratio of the GreekLex frequencies to the SUBTLEX-GR frequencies revealed that the 50 words with much higher frequency values in the GreekLex corresponded mostly to the juridical and politico-economical lexical fields (e.g., υφυπουργός [vice minister], επιτόκιο [compound interest]). Furthermore, some of these words corresponded to Ancient Greek word types which do not belong to the Modern Greek grammar and are almost exclusively used in sophisticated linguistic contexts (e.g., ήτοι [namely], γίγνεσθαι [Ancient Greek infinitive form of the verb "to become"]). This set of comparisons showed that the discrepancies in the frequency estimates provided by the two databases clearly reflected the different composition of the two corpora of origin (i.e., subtitles and books and newspaper articles). However, more importantly, the two frequency counts highly correlated, suggesting that the final word frequency values were not determined by the precedence of the raw material (written texts or subtitles).

The comparison of the new subtitle-based Greek frequency database to the existing written frequencies showed that at a descriptive level the new database captures the wealth of the Greek language and that it is largely comparable to GreekLex. However, we also wanted to experimentally test whether both SUBTLEX-GR and GreekLex would effectively account for the reading performance of young Greeks and to explore whether either of the two databases provides a better estimate of their reading behavior. To this end, two lexical decision experiments were conducted. The lexical decision task, one of the most commonly used in the visual word recognition literature, has been repeatedly used to validate frequency counts (e.g., Burgess and Livesay, 1998; Zevin and Seidenberg, 2002; see also New et al., 2007; Brysbaert and New, 2009; Cai and Brysbaert, 2010; Keuleers et al., 2010). Furthermore, this task has consistently revealed frequency effects (i.e., faster response latencies for high frequency as compared to low frequency words; e.g., Scarborough et al., 1977; Morrison and Ellis, 1995; Balota et al., 2004), while the word frequency has been found to be a very good predictor of lexical decision latencies (e.g., Morrison and Ellis, 1995; Balota et al., 2004; Alija and Cuetos, 2006).

In further detail, in Experiment 1 we opted for using a factorial design with words whose frequency per million as given by SUBTLEX-GR and by GreekLex was orthogonally manipulated. This factorial design provides a unique way of assessing the validity of the subtitle-based frequencies independently from the GreekLex frequencies, since it offers a way of directly examining whether SUBTLEX-GR can effectively account for the performance of the participants when they are faced with two critical sets of words: those with a large number of appearances in SUBTLEX-GR and a small number of appearances in GreekLex and vice versa. Using the alternative and more common approach of collecting data from a random sample of words irrespectively of their frequency in each of the two databases and performing exclusively multiple regression analyses would not answer the question of whether the SUBTLEX-GR frequencies were valid or not independently from the GreekLex frequencies, given that the two frequency counts are highly correlated (see above). In contrast, a main effect of the SUBTLEX-GR frequencies (i.e., significantly better performance for words with a high frequency value in SUBTLEX-GR than for words with a low frequency, e.g., Whaley, 1978; Balota et al., 2004) would provide clear-cut evidence of the validity of the new frequencies, irrespectively of the GreekLex frequency norms. In Experiment 2 the alternative procedure was followed in order to compare the quality of the two frequency databases. Native Greeks performed lexical decisions on a set of Greek words which corresponded to the names of the colored and texturized version of the Snodgrass and Vanderwart (1980) picture set, one of the most commonly used experimental sets in psycholinguistic research (see Dimitropoulou et al., 2009). This specific set of words consists exclusively of concrete nouns, all representing everyday objects. The collection of lexical decision data on a set of words with random frequency values has been the methodology followed so far by the existing subtitle corpora to empirically compare the quality of the new corpora to the that of other text-based ones (e.g., New et al., 2007; Cai and Brysbaert, 2010). In line with previous studies of subtitle-based corpora showing that these frequency counts account for at least as much variance of lexical decision latencies as text-based frequency norms, we expected that SUBTLEX-GR (as well as GreekLex) would effectively account for the lexical decision performance of the participants. Nevertheless, the question remained whether the increased exposure of this population to subtitle reading would influence the extent to which the new subtitle-based frequency norms would account for their reading performance as compared to the written frequencies taken from GreekLex. If the fact that the Greek readers are much more exposed to subtitle reading has influenced their reading behavior, then the subtitle-based frequencies should correlate more highly with their lexical decision performance and should explain a larger portion of reaction time variance than the written frequencies from GreekLex (see also Keuleers et al., 2010).

EXPERIMENT 1

MATERIALS AND METHODS

Participants

Thirty native Greek university students (mean age 26.1 ± 3.2) participated voluntarily in this experiment.

Materials

In order to select the critical words, the high frequency (HF hereafter; word frequency above 65 appearances per million) and the low frequency Greek words (LF hereafter; word frequency below 35 appearances per million) with frequency per million higher than 0 in both SUBTLEX-GR and GreekLex were identified. Based on these frequency cut-offs we created the following experimental conditions:

(i) HF SUBTLEX-GR/HF GreekLex (e.g., νερό [water]), (ii) HF SUBTLEX-GR/LF GreekLex (δώρο [gift]), (iii) LF SUBTLEX-GR/HF GreekLex (έτος [year]) and, (iv) LF SUBTLEX-GR/LF GreekLex (άρμα [float]). All the conditions were matched for word length (measured in number of letters) and number of orthographic neighbors (*N*). Furthermore, the four conditions were also matched in concreteness. The influence of this variable on the lexical decision performance has been repeatedly shown (e.g., James, 1975; Schwanenflugel and Shoben, 1983). However, since there are no concreteness norms available for Modern Greek, we collected concreteness ratings for a subset of the words that matched the frequency criteria across the two databases. To end up with a reasonable number of candidate items to be rated, a native speaker of Modern Greek eliminated (i) extremely concrete or extremely abstract words and (ii) words that were clearly related to corpus-specific semantic categories (see above). Following this step, we presented 16 native Greek speakers with the remaining 299 words that matched our frequency criteria across the two databases and asked them to rate them on a 7-point Likert scale (1 = totally abstract, 7 = totally concrete). The final 172 critical items were selected based on these concreteness ratings. An additional set of 172 ortho-phonologically legal non-words were also created for the purposes of the lexical decision task (e.g., κλίχα; see **Table 1** for a full description of the materials used).

Procedure

The experiment was run individually using DMDX software (Forster and Forster, 2003). Each trial consisted of the presentation of a fixation point for 300 ms, followed by a target letter string in lowercase. Each character of the targets had a width of 0.16 inches (note that Courier New font is a non-proportional font in which all letters occupy the same amount of space). Participants were instructed to make lexical decisions by pressing as fast as possible one of two buttons on the keyboard. After every response or 2000 ms, the next trial started. The order of item presentation was randomized across participants.

RESULTS

Incorrect responses (approximately 2.0% of the word data) and reaction times shorter than 250 ms and longer than 1500 ms (less than 1% of the word data) were excluded from the latency data analyses. Response times and error rates associated with each experimental condition are displayed in **Table 2**. ANOVAs based on the participant (*F*1) and item (*F*2) response latencies and error percentage were conducted, following a 2 (SUBTLEX-GR count: High frequency, Low frequency) × 2 (GreekLex count: High frequency, Low frequency) design.

Analyses on the reaction time data showed a main effect of word frequency as measured by the SUBTLEX-GR frequency norms, indicating that according to this count, low frequency words were responded to 34 ms slower than high frequency words [$F1(1,29) = 55.91$, MSe = 624, $p < 0.001$; $F2(1,159) = 34.56$, MSe = 1512, $p < 0.001$]. The main effect of the GreekLex frequency count was also significant, showing that words with a low frequency according to this frequency count were responded to 15 ms slower than words characterized by GreekLex as high frequency [$F1(1,29) = 17.15$, MSe = 408, $p < 0.001$; $F2(1,159) = 7.12$, MSe = 1512, $p < 0.001$]. The interaction between the two factors was not significant (both ps > 0.16). ANOVAs on the error data did not show any significant effects (all ps > 0.27).

Moreover, separate sets of correlation analyses were performed in order to examine which of the two databases, SUBTLEX-GR or GreekLex, correlated more highly with the lexical decision data. Following the procedure used to validate the existing subtitle-based corpora (e.g., Brysbaert and New, 2009; Cai and Brysbaert, 2010; Keuleers et al., 2010), multiple regression analyses were run in order to test the predictive capacity of the GreekLex frequencies and of the alternative frequency measures that were additionally calculated for the SUBTLEX-GR entries. The additional measures were (1) the CD, and (2) the frequency of each of the words calculated based on the number of times they appeared in the corpus starting with a lowercase letter (FREQlow in the restricted version of the corpus). For each measure two sets of multiple regression analyses were performed over the lexical decision latencies and two over the error rates. In the first set of these analyses the following four predictors were included: the log10 (number of appearances + 1), $\log^2 10$ (number of appearances + 1), number of letters in the word, and number of syllables in the word (see also New et al., 2007; Brysbaert and New, 2009, for the same procedure), while in the second we only included as predictors the log10 (number of appearances + 1), $\log^2 10$ (number of appearances + 1). For the multiple regression analyses on the CD values, the log10 (number of appearances + 1), and $\log^2 10$ (number of appearances + 1) predictors were replaced by log10 (number of different contexts + 1) and $\log^2 10$ (number of different contexts + 1). To eliminate the skewness in the distribution of the reaction times, we applied logarithmic transformation (Baayen et al., 2006).

Table 1 | Mean word frequency (per million), number of orthographic neighbors (*N*), word length (number of characters) and subjective concreteness ratings of the words used in Experiment 1 (standard deviations are given within parentheses).

	Word frequency		*N*		*L*	Concret.
	SUBTLEX-GR	GreekLex	SUBTLEX-GR	GreekLex		
HF/HF	144.0(88.1)	156.7(89.3)	2.7	1.3	6.6	4.4
HF/LF	143.6(96.7)	14.4(8.2)	2.6	0.8	6.5	4.4
LF/HF	14.5(8.6)	153.7(113.6)	2.2	0.9	7.1	4.5
LF/LF	13.8(9.8)	11.8(6.8)	1.7	0.7	6.9	4.5

HF, High frequency; LF, Low frequency; L, Word length; Concret., Concreteness.

The correlation analyses showed that the word frequency per million appearances (log transformed) as given by both databases negatively correlated with the participants' lexical decision times (SUBTLEX-GR, $r = -0.49$, $p < 0.001$; GreekLex: $r = -0.22$, $p < 0.01$). Suggesting that participants responded more slowly the less frequent the words were. Critically, this correlation was significantly larger for the subtitle-based frequency counts, $t(1,169) = -3.03$, $p < 0.01$[2]. No significant correlations were found for the error rates.

The multiple regression analyses revealed that SUBTLEX-GR accounted for 28.8% while GreekLex accounted for 18.1% of the reaction time variance, in line with the results of the correlation analyses on the reaction times. Interestingly, when word length and number of syllables were excluded from the multiple regression analyses the amount of variance explained by the GreekLex frequency count dropped dramatically (8.3%), while this was not the case for SUBTLEX-GR (24.3%). Similar to what has been found for American-English, Dutch, and Chinese, the number of different contexts a word appeared in the corpus was the most effective predictor of the lexical decision times consistently outperforming the traditional frequency measure (i.e., around 4% extra variance explained independently of whether word length and number of syllables were also included in the model). In line with what Brysbaert and New (2009) reported for American-English, the number of times the words appeared in the corpus starting with a lowercase letter was also found to be a better predictor of the lexical decision latencies than the conventional frequency measure, though this advantage was not as marked as the one found for the CD measure (see **Table 3**).

Finally, in order to examine the amount of additional variance each of the two frequency measures explained when the influence of the other was partialed out, we performed two sets of multiple regression analyses over the reaction times (log transformed) and the error rates, using a three-step hierarchical approach. In the first step we included as predictors the word length and the number of syllables, since these variables tap onto the sublexical level of processing, and in the second and third steps we included the two word frequency measures [log10 (number of appearances + 1), log²10 (number of appearances + 1); see Balota et al., 2004; Yap and Balota, 2009; Cai and Brysbaert, 2010, for a similar procedure]. These analyses showed that the additional percentage of reaction time variance explained by SUBTLEX-GR after partialing out the influence of the GreekLex frequencies was substantially larger than that explained by GreekLex when the influence of the SUBTLEX-GR frequencies had been partialed out in the previous step (15.6 vs. 5.1% of additional variance explained, for SUBTLEX-GR and GreekLex, respectively). The results from the same analyses on the error rates showed that GreekLex frequencies explained 4.8% more variance when the influence of the SUBTLEX-GR frequencies had been previously partialed out, while this advantage was only of 2.8% when the two frequency measures were introduced in the model in the opposite order (see **Table 4**).

[2]The formula used to compare correlation coefficients (Pearson *r*) was

$$t = \frac{(r_{yx_1} - r_{yx_2})\sqrt{(n-3)(1+r_{x_1x_2})}}{\sqrt{2(1-r_{yx_1}^2 - r_{yx_1}^2 - r_{x_2x_1}^2 + 2r_{yx_1}r_{yx_2}r_{x_2x_1}}}$$

Table 2 | Average response times in ms (RT) and error rates (%*E*) associated with each experimental condition in Experiment 1.

GreekLex	SUBTLEX-GR					
	High frequency		Low frequency		Mean	
	RT	%*E*	RT	%*E*	RT	%*E*
High frequency	599	2.2	627	2.1	613	2.1
Low frequency	608	1.5	648	2.2	628	1.9
Mean	604	1.8	638	2.2		

Reaction times and error rates associated with responses to non-words averaged 733 ms and 5.0, respectively.

Table 3 | Percentages of reaction times (RT) and error rate (*E*%) variance (adjusted R^2) corresponding to Experiment 1 explained by the different frequency measures with and without word length (*L*) and number of syllables (NSyl) as additional variables.

Measure	R^2_{RT}	$R^2_{E\%}$
FREQ. MEASURE + *L* + NSyl		
SUBTLEX-GR FREQ.	28.8	3.0 n.s.
SUBTLEX-GR CD	32.8	4.4
SUBTLEX-GR FREQlow	30.3	2.0 n.s.
GreekLex	18.1	5.0
FREQ. MEASURE		
SUBTLEX-GR Freq.	24.3	3.7
SUBTLEX-GR CD	27.9	5.0
SUBTLEX-GR FREQlow	25.1	2.7
GreekLex	8.3	5.8

SUBTLEX-GR FREQ., number of occurrences; SUBTLEX-GR CD, number of different contexts; SUBTLEX-GR FREQlow, number of occurrences starting with a lowercase letter; n.s., non-significant at the alpha level of 0.05.

Table 4 | Adjusted R^2 values from three-step reaction time (RT) and error rate (*E*%) multiple regression analyses for Experiment 1.

Order of entry into regression model	R^2_{RT}	$R^2_{E\%}$
Step 1: L + NSyl.	10.0	–1.1 n.s.
Step 2: GreekLex	18.1	5.0
Step 3: SUBTLEX-GR	33.7	7.8
R^2 difference (Step 3 – 2)	*15.6*	*2.8*
Step 1: L + NSyl.	10.0	–1.1 n.s.
Step 2: SUBTLEX-GR	28.8	3.0 n.s.
Step 3: GreekLex	33.7	7.8
R^2 difference (Step 3 – 2)	*5.1*	*4.8*

L, Word length; NSyl., Number of syllables; n.s., non-significant at the alpha level of 0.05; R^2 difference, Additional percentage of variance explained by one frequency database when the influence of the other is partialed out.

DISCUSSION

The results from this experiment showed that similar to what was previously found with the existing subtitle-based frequency norms, the SUBTLEX-GR frequency estimates were found to successfully account for the reading performance of the participants, showing a highly significant main effect. Even more critically, the subsequently performed correlation and multiple regression analyses showed that the new SUBTLEX-GR frequencies outperformed the predictive capacity of the written frequency counts, explaining more than 10% of additional variance (see also Keuleers et al., 2010). This advantage was even larger, for the SUBTLEX-GR CD measure (an additional 4% of reaction time variance explained). These findings indicate that in the case of Modern Greek, subtitles provide a better source of word frequency measures than written texts do. However, considering the importance of this claim, a second experiment was conducted in order to test whether this result would be replicated.

In Experiment 2 we followed a procedure comparable to that previously used to validate the existing subtitle frequency databases: Lexical decision data for the Greek names of the Snodgrass and Vanderwart (1980) picture set (see Dimitropoulou et al., 2009) were collected and correlation and multiple regression analyses were performed to examine whether SUBTLEX-GR would predict participants' reading performance more effectively than GreekLex.

EXPERIMENT 2

MATERIALS AND METHODS

Participants

A group of 30 native Greek university students (mean age 24.1 ± 2.9) took part in Experiment 2.

Materials

Two hundred twenty-six words were selected from the Modern Greek picture naming norms, which corresponded to the most frequent Modern Greek name given to the pictures of the Rossion and Pourtois (2004) color picture set (Dimitropoulou et al., 2009). From the original set of 260 picture names, 34 were excluded because they met one of the following exclusion criteria: (1) they appeared as the most frequent name of more than one picture, (2) they were plurals, (3) they were multiword names or (4) they did not appear in either of the two databases. The word frequency and *N* values of the remaining picture names were then taken from GreekLex and SUBTLEX-GR. Average frequency was of 7.3 (range 0.1–133.1) and of 22.8 (range 0.1–337.7) appearances per million, while average *N* was of 1.0 and 2.2, in GreekLex and SUBTLEX-GR, respectively. Mean word length was 6.7 letters. An additional set of 226 orthophonologically legal non-words (e.g., πέθα) was also created for the purposes of the lexical decision.

Procedure

The procedure followed was exactly the same as in Experiment 1.

RESULTS

Incorrect responses (approximately 3.6% of the word data) and reaction times shorter than 250 ms and longer than 1500 ms (less than 0.5% of the word data) were excluded from the latency data analyses. Correlations were performed between reaction times and error rates and the log transformed word frequencies per million as given by GreekLex and by SUBTLEX-GR, while in the multiple regression analyses performed the same four predictors as in Experiment 1 were included: log10 (number of appearances + 1), $\log^2 10$ (number of appearances + 1), number of letters, and number of syllables.

Both the GreekLex and the SUBTLEX-GR log transformed frequencies negatively correlated with the lexical decision times, $r = -0.59$, $p < 0.001$ and $r = -0.65$, $p < 0.001$, respectively, showing that participants produced the slowest responses to the less frequent words. Importantly, this negative correlation was significantly higher for the SUBTLEX-GR frequency estimates than for the GreekLex frequencies, $t(1,223) = -2.91$, $p < 0.05$. The correlation analyses performed on the error rates showed that the less frequent the words were, the more errors were made, according to both GreekLex and SUBTLEX-GR, $r = -0.38$, $p < 0.001$ and $r = -0.46$, $p < 0.001$, respectively. However, this correlation was again larger for the new SUBTLEX-GR corpus than for GreekLex, $t(1,223) = 2.59$, $p < 0.05$.

The results of the multiple regression analyses including the same four variables as in Experiment 1 [i.e., log10 (number of appearances + 1), $\log^2 10$ (number of appearances + 1), number of letters in the word, and number of syllables in the word] followed the same pattern: the GreekLex frequency estimates explained 39.8% of the variance of the reaction times, while the SUBTLEX-GR frequency estimates explained 46.9% of the variance. Similar to what was found in Experiment 1, lowercase frequency and, even more so, CD were shown to be better predictors of the reading performance of the participants of the present experiment, further corroborating the predictive capacity of these variables (e.g., Brysbaert and New, 2009). However, in this case this advantage over the conventional frequency count was smaller than in Experiment 1 (around 2 and 1% additional variance explained by the SUBTLEX-GR CD and the SUBTLEX-GRLow measures, respectively, as compared to the SUBTLEX-GR Freq. measure). The multiple regression analyses conducted on the error rates revealed that GreekLex accounted for 17.6% of the error rate variance while SUBTLEX-GR accounted for 29.6% of the variance (see **Table 5**).

Table 5 | Percentages of reaction times (RT) and error rate (*E*%) variance (adjusted R^2) corresponding to Experiment 2 explained by the different frequency measures with and without word length (*L*) and number of syllables (NSyl) as additional variables.

Measure	R^2_{RT}	$R^2_{E\%}$
FREQ. MEASURE + *L* + NSyl		
SUBTLEX-GR FREQ.	46.9	29.6
SUBTLEX-GR CD	48.6	30.5
SUBTLEX-GR FREQlow	47.7	30.4
GreekLex	36.5	17.6
FREQ. MEASURE		
SUBTLEX-GR FREQ.	43.7	25.9
SUBTLEX-GR CD	45.6	26.3
SUBTLEX-GR FREQlow	44.4	26.9
GreekLex	36.5	14.8

SUBTLEX-GR FREQ., number of occurrences; SUBTLEX-GR CD, number of different contexts; SUBTLEX-GR FREQlow, number of occurrences starting with a lowercase letter; n.s., non-significant at the alpha level of 0.05.

Finally, using the same variables and following the same order of variable entry as in Experiment 1, we performed three-step hierarchical multiple regression analyses over the reaction times and the error rates collected in the present experiment. These analyses showed that the SUBTLEX-GR frequencies explained an additional 6.1 and 11.8% of reaction time and error rate variance, as compared to the variance explained by GreekLex. Critically, there was not such an advantage for the GreekLex frequency counts, when the influence of the SUBTLEX-GR frequencies was partialed out (see **Table 6**).

DISCUSSION

The findings of Experiment 2 showed that, similar to what had been previously reported for the existing subtitle-based corpora, word frequencies taken from Greek subtitles offer a valid estimate of Modern Greek frequency. Importantly, in line with the findings of Experiment 1, these results showed that SUBTLEX-GR outperformed the written based frequency norms provided by GreekLex (Ktori et al., 2008) in accounting for the reading performance of young Greek adults making lexical decisions on the Greek names of the Snodgrass and Vanderwart (1980) picture set.

GENERAL DISCUSSION

This study presents SUBTLEX-GR a new frequency database exclusively based on subtitles, which provides valid word frequency estimates for Modern Greek. The validity of this new tool was confirmed in two lexical decision experiments with young Greek adult readers. Importantly, we also showed that although the frequency norms based on written texts provide an effective account of reading performance, the measures of word frequency provided by SUBTLEX-GR were consistently found to be superior.

As described, the quality of SUBTLEX-GR was first examined by comparing it to GreekLex, a word frequency database taken from a written corpus of Modern Greek. The comparison with a written corpus could seem somewhat contradictory given that subtitles are transcripts of spoken language and thus comparing them with a corpus of spoken language could have seemed more appropriate. However, previous evidence has shown that subtitle-based frequency counts predict visual word recognition performance significantly better than spoken corpora (New et al., 2007; Brysbaert and New, 2009). At a descriptive level, the comparison between SUBTLEX-GR and GreekLex showed that the frequency and length distributions of the words listed in the two databases were largely comparable. This suggests that subtitle files provide a valid source to create a Modern Greek word frequency database that is not limited in terms of the characteristics word material it provides (i.e., does not include mostly short words that represent a limited range of semantic categories).

The empirical data gathered from the two experiments further confirmed that the new frequency counts based on subtitles were valid, since they explained a significant proportion of the variance in the lexical decision latencies and in the error rates obtained from two groups of young Greek adults. In further detail, in Experiment 1, where the frequency of the words as given by SUBTLEX-GR and GreekLex was orthogonally manipulated, SUBTLEX-GR explained more that 28% (up to 32.8% depending on the frequency measure used) of the reaction time variance. In Experiment 2, where the Greek names of the Snodgrass and Vanderwart (1980) picture set were presented, SUBTLEX-GR explained up to 48.6 and 30.5% of reaction time and error rate variance, respectively. These results are in line with what has recently been found for the French, the American-English, the Dutch, and the Chinese subtitle-based corpora, which were also shown to effectively predict reading behavior (New et al., 2007; Brysbaert and New, 2009; Cai and Brysbaert, 2010; Keuleers et al., 2010). Hence, our findings provide further experimental support to the proposal that subtitles provide a very representative source of linguistic material and that frequency norms based on subtitles are a valuable tool for psycholinguistic research. Such a conclusion is based on the observation that young adults, who participate in the vast majority of psycholinguistic studies, are increasingly exposed to visual media. In the case of Greeks, a recent survey revealed that 87% of the population watches television for a mean of three hours and a half per day (VPRC, 2005). Moreover, given that subtitles are transcripts of oral mostly informal communication, they tap onto a linguistic style closely related to that used by the younger part of the population, which makes them a very appropriate source of material selection, especially for studies testing this part of the population.

The importance of the type of linguistic material readers are commonly exposed to was further underlined by the fact that the SUBTLEX-GR frequency values consistently outperformed those given by the only existing written text-based frequency database for Modern Greek (GreekLex; Ktori et al., 2008). This advantage was reflected in the extra proportion of reaction time and error rate variance explained by SUBTLEX-GR as compared to GreekLex. For the reaction time data from Experiment 1, this advantage ranged between 10 and 19% depending on the frequency measure used and the variables introduced in the regression model (see **Table 3**), while for Experiment 2 it ranged from 6.7 to 12% for the reaction times and was around 12% for the error rates. In fact, the stability of this outcome was confirmed by additional hierarchical multiple regression analyses which showed that the magnitude of the observed advantage for the SUBTLEX-GR frequencies remained constant when the influence of the GreekLex frequencies had been previously partialed out, while this was not the case when GreekLex

Table 6 | Adjusted R^2 values from three-step reaction time (RT) and error rate (E%) multiple regression analyses for Experiment 2.

Order of entry into regression model	R^2_{RT}	$R^2_{E\%}$
Step 1: *L* + NSyl.	17.6	4.7
Step 2: GreekLex	39.8	17.6
Step 3: SUBTLEX-GR	46.9	29.4
R^2 difference (Step 3 – 2)	*7.1*	*11.8*
Step 1: *L* + NSyl.	17.6	4.7
Step 2: SUBTLEX-GR	46.9	29.6
Step 3: GreekLex	46.9	29.4
R^2 difference (Step 3 – 2)	*0*	*–0.2*

L, Word length; NSyl., Number of syllables; n.s., non-significant at the alpha level of 0.05; R^2 difference, Additional percentage of variance explained by one frequency database when the influence of the other is partialed out.

was included in the regression model after SUBTLEX-GR. The outcome of this comparison between subtitle and written text-based frequency estimates for Modern Greek is not such a striking result if one considers that the vocabulary used to subtitle movies or television series is very similar to that used by young Greek adults. In contrast, the vocabulary on which written text-based frequency counts are based is more sophisticated and somewhat influenced by a variation of the Greek language which is no longer in use. Aside from this, the advantage observed for SUBTLEX-GR over GreekLex seems to also reflect the fact that young Greek people are intensely exposed to subtitle reading (unlike French, American, or Chinese young adults). In fact, the resemblance of our findings to those obtained in Dutch, a language whose speakers are also used to subtitle reading, suggests that this could indeed be the case. In the multiple regression analyses performed over the lexical decision latencies of a group of Dutch university students, Keuleers et al. (2010) found that SUBTLEX-NL frequencies offered an advantage of almost 10% in explained variance over the written frequencies for Dutch given by CELEX. Thus, our results suggest that this extensive subtitle reading experience rendered the subtitle-based frequency counts even more effective in accounting for the reading performance of the participants than in linguistic societies in which subtitle reading is less common.

Another important outcome of the experimental validation of SUBTLEX-GR was the consistently advantage found for the frequency of occurrence of the words in the corpus starting with a lowercase letter and for the CD of the words (in our analyses SUBTLEX-GR Low and SUBTLEX-GR CD, respectively) over the conventionally used total frequency of occurrence of the word. The distinction of the words based on whether they were encountered within the subtitle files starting with an upper or a lowercase letter is useful in order to identify words that are mostly used as proper names and to avoid selecting them as experimental materials. Similar to what we found in both experiments, Brysbaert and New (2009) found that the lowercase frequency was a better measure than the total frequency, suggesting that whether or not words appear mostly starting with an upper or a lowercase letter influences the reading behavior of the participants. Nevertheless, the CD proved to be an even better frequency measure, providing an additional advantage over the commonly used total frequency of occurrence that reached 4%, in line with what has been reported for English, Dutch, and Chinese (see Adelman et al., 2006; Brysbaert and New, 2009; for a discussion of the potential explanation of this benefit). Hence, our findings further corroborate the value of CD and suggest that in addition to the typically used frequency measure this alternative measure of word frequency should be also taken into account when searching for experimental word materials.

As a final remark, we opted for presenting the results of an additional cross-validation of SUBTLEX-GR: its comparison to $SUBTLEX_{US}$. The reasoning that led us to perform this comparison was that since both databases were mostly based on subtitle files that corresponded to American film and television productions, the frequency values given for similar Greek and English words could be expected to highly correlate. As a way to test whether this is indeed the case we compared the two frequency counts for the same words in their Greek and their English version using translation equivalents. We performed such a comparison on the frequency values given for the words used in Experiments 1 and 2. After translating to English the words from Experiment 1, we eliminated the words with multiple-word English translations and those with the same English translation. The word frequency per million of the remaining 166 words (out of the initial 172) as given by $SUBTLEX_{US}$ was identified and correlation analyses were performed. For the Greek picture names taken from Dimitropoulou et al. (2009) that were used in Experiment 2, the subsequent correlation analyses were performed on the frequency values of the unique single-word Greek–English translation pairs (199 out of the initial 226). For the two sets of experimental items put under test, the frequency values given by $SUBTLEX_{US}$ for the English translations of the Greek words correlated significantly with the SUBTLEX-GR frequencies given for the Greek words, $r = 0.50$, $p < 0.001$ and $r = 0.49$, $p < 0.001$, for the word pairs of Experiment 1 and Experiment 2, respectively. Interestingly, the same correlation analyses performed between the $SUBTLEX_{US}$ and GreekLex frequencies revealed that a significant correlation existed only for the words of Experiment 2 ($r = 0.21$, $p < 0.01$), though to a lesser extent than the one obtained between the two subtitle-based corpora (as shown by the statistical comparison of the two correlations; $p < 0.001$). The absence of such a correlation for the words of Experiment 1 in the presence of a significant correlation between SUBTLEX-GR and $SUBTLEX_{US}$ for the same set of words indicates that the two subtitle corpora (American-English and Greek) are highly comparable and that the common origin of the raw subtitle files of the two corpora is clearly reflected in the two frequency counts.

CONCLUSION

Starting with French (New et al., 2007), it has been by now shown in a number of languages that subtitles represent a valuable source of up-to-date linguistic material providing a very good approximation of the real life exposure to language of young adults, the age group most commonly tested in psycholinguistic studies. The results of our lexical decision experiments showed that the relevance of subtitle-based corpora is even more marked for populations systematically exposed to subtitle reading, as is the case of young Greek adults, thus providing further experimental support to the idea that "*the more natural the language use is, the better the frequency norms account for lexical decision times*" (Brysbaert and New, 2009, p. 986). In the light of the present findings, we strongly encourage psycholinguists investigating Modern Greek to take into consideration the subtitle-based frequency counts presented in SUBTLEX-GR when creating their experimental materials.

ACKNOWLEDGMENTS

This research has been partially supported by Grants PSI2009-08889 and CONSOLIDER-INGENIO 2010 (CSD2008-00048) from the Spanish Ministry of Science and Innovation. Maria Dimitropoulou was the recipient of a post-graduate grant from the Government of the Canary Islands (BOC 241, 02/12/2008). Many thanks are due to N. Glaros and the Institute of Language and Speech Processing (Athens, Greece) for their collaboration and to Manuel Perea for his insightful suggestions.

REFERENCES

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

Alija, M., and Cuetos, F. (2006). Effects of the lexical-semantic variables in visual word recognition. *Psicothema* 18, 485–491.

Baayen, R. H., Feldman, L. B., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *J. Mem. Lang.* 55, 290–313.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database (Release 2) [CDROM]*. Philadelphia, PA: Linguistic Data Consortium.

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

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition for single syllable words. *J. Exp. Psychol. Gen.* 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., and Treiman, R. (2007). The English lexicon project. *Behav. Res. Methods* 39, 445–459.

Bonin, P., Chalard, M., Méot, A., and Fayol, M. (2001). Age-of-acquisition and word frequency in the lexical decision task: further evidence from the French language. *Curr. Psychol. Cogn.* 20, 401–443.

Brysbaert, M., and 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. *Behav. Res. Methods* 41, 977–990.

Burgess, C., and Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: moving on from Kučera and Francis. *Behav. Res. Methods Instrum. Comput.* 30, 272–277.

Cai, Q., and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE* 5, e10729. doi: 10.1371/journal.pone.0010729.

Carreiras, M., Perea, M., and Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: cross-task comparisons. *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 857–871.

Coltheart, M., Davelaar, E., Jonasson, J. T., and Besner, D. (1977). "Access to the internal lexicon," in *Attention and Performance VI*, ed. S. Dornic (Hillsdale, NJ: Erlbaum), 535–555.

Davis, C. J., Perea, M., and Acha, J. (2009). Re(de)fining the orthographic neighbourhood: the role of addition and deletion neighbours in lexical decision and reading. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1550–1570.

Dimitropoulou, M., Duñabeitia, J. A., Blitsas, P., and Carreiras, M. (2009). A standardized set of 260 pictures for modern Greek: norms for name agreement, age of acquisition and visual complexity. *Behav. Res. Methods* 41, 584–589.

Dimitropoulou, M., Duñabeitia, J. A., and Carreiras, M. (in press-a). Phonology by itself: masked phonological priming effects with and without orthographic overlap. *Eur. J. Cogn. Psychol.*

Dimitropoulou, M., Duñabeitia, J. A., and Carreiras, M. (in press-b). Masked translation priming effects with low proficient bilinguals. *Mem Cogn.*

Duñabeitia, J. A., Marín, A., and Carreiras, M. (2009a). Associative and orthographic neighborhood density effects in normal aging and Alzheimer's disease. *Neuropsychology* 23, 759–764.

Duñabeitia, J. A., Perea, M., and Carreiras, M. (2009b). There is no clam with coats in the calm coast: delimiting the transposed-letter priming effect. *Q. J. Exp. Psychol.* 62, 1930–1947.

Duñabeitia, J. A., Molinaro, N., Laka, I., Estévez, A., and Carreiras, M. (2009c). N250 effects for letter transpositions depend on lexicality: casual or causal? *Neuroreport* 20, 381–387.

Duñabeitia, J. A., and Vidal-Abarca, E. (2008). Children like dense neighborhoods: orthographic neighborhood density effects in novel readers. *Span. J. Psychol.* 11, 26–35.

Forster, K. I., and Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behav. Res. Methods Instrum. Comp.* 35, 116–124.

Georgiou, G. K., Protopapas, A., Papadopoulos, T. C., Skaloumbakas, C., and Parrila, R. (2010). Auditory temporal processing and dyslexia in an orthographically consistent language. *Cortex* 46, 1330–1344.

Hatzigeorgiu, N., Mikros, G., and Carayannis, G. (2005). Word length, word frequencies and Zipf's law in the Greek language. *J. Quant. Linguist.* 12, 167–184.

Holcomb, P. J., Grainger, J., and O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *J. Cogn. Neurosci.* 14, 938–950.

James, C. T. (1975). The role of semantic information in lexical decisions. *J. Exp. Psychol. Hum. Percept. Perform.* 1, 130–136.

Kazazis, K. (2001). Dismantling modern Greek diglossia: the aftermath. *Lingua e stile* 36, 291–298.

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

Koliopoulou, M. (2008). "The loose multiword compounds of modern Greek under the prism of construction morphology," in *New Perspectives in Greek Linguistics*, eds N. Lavidas, E. Nouchoutidou, and M. Sionti (Cambridge: Cambridge Scholars Publishing), 213–224.

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.

Ktori, M., and Pitchford, N. J. (2009). Development of letter position processing: effects of age and orthographic transparency. *J. Res. Read.* 32, 180–198.

Ktori, M., van Heuven, W. J. B., and Pitchford, N. J. (2008). GreekLex: a lexical database of modern Greek. *Behav. Res. Methods* 40, 773–783.

Kučera, H., and Francis, W. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* 10, 707.

Mackridge, P. (2009). *Language and National Identity in Greece 1766-1976*. New York: Oxford University Press.

Mikros, G., Hatzigeorgiu, N., and Carayannis, G. (2001). Basic quantitative characteristics of the modern Greek language using the hellenic national corpus. *J. Quant. Linguist.* 8, 175–185.

Morrison, C. M., and Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *J. Exp. Psychol. Learn Mem. Cogn.* 21, 116–113.

Müller, O., Duñabeitia, J. A,, and Carreiras, M. (2010). Orthographic and associative neighborhood density effects: what is shared, what is different? *Psychophysiology* 47, 455–466.

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

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: a new French lexical database. *Behav. Res. Methods* 36, 516–524.

New, B., Pallier C., Ferrand L., and Matos, R. (2001). Une base de données lexicales du Français contemporain sur Internet: LEXIQUE [a lexical database on the internet about contemporary French: LEXIQUE]. *Annee. Psychol.* 101, 447–462.

Orfanidou, E., and Sumner, P. (2005). Language switching and the effects of orthographic specificity and response repetition. *Mem. Cogn.* 33, 355–369.

Protopapas, A., and Vlahou, E. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behav. Res. Methods* 41, 991–1008.

Ralli, A. (2007). *The Composition of Words. A Morphological Cross-linguistic Approach [in Greek]*. Athens: Patakis.

Rossion, B., and Pourtois, G. (2004). Revisiting snodgrass and vanderwart's object set: the role of surface detail in basic-level object recognition. *Perception* 33, 217–236.

Scarborough, D. L., Cortese, C., and Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *J. Exp. Psychol. Hum. Percept. Perform.* 3, 1–17.

Schwanenflugel, P. J., and Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *J. Exp. Psychol. Learn. Mem. Cogn.* 9, 82–102.

Sebastián-Gallés, N., Martí, M. A., Carreiras, M., and Cuetos, F. (2000). *Lexesp: Léxico Informatizado Del Español*. Barcelona: Edicions Universitat de 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.

VPRC. (2005). *Second Panhellenic Survey on Reading Behavior and Cultural Practices*. Greece: On behalf of the "National Center of Book".

Whaley, C. P. (1978). Word-non-word classification time. *J. Verbal Learn. Verbal Behav.* 17, 143–154.

Yap, M., J., and Balota, D. A. (2009). Visual word recognition of multisyllabic words. *J. Mem. Lang.* 60, 502–529.

Yarkoni, T., Balota, D. A., and Yap, M. J. (2008). Moving beyond Coltheart's N: a new measure of orthographic similarity. *Psychon. Bull. Rev.* 15, 971–979.

Zevin, J. D., and Seidenberg, M. S. (2002). Age of acquisition effects in reading and other tasks. *J. Mem. Lang.* 47, 1–29.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 May 2010; accepted: 16 November 2010; published online: 16 December 2010.

Citation: Dimitropoulou M, Duñabeitia JA, Avilés A, Corral J and Carreiras M (2010) Subtitle-based word frequencies as the best estimate of reading behavior: the case of Greek. Front. Psychology ***1****:218. doi: 10.3389/fpsyg.2010.00218*

This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.

APPENDIX

INFORMATION REGARDING THE SUBTLEX-GR CORPUS AND THE CORRESPONDING FILES AVAILABLE TO RESEARCHERS

Out of the different 5,508 contexts that were identified within the approximately 6,100 unique subtitle files constituting our raw material, 4,001 corresponded to films and 1,507 to television series episodes (an average of 7.9 episodes per series). Accordingly, out of the more than 27 million space-separated tokens of SUBTLEX-GR, 84.8% were taken from movies and only 15.2% from television shows. The films and the television series to which the subtitles corresponded were mostly USA productions (71.7%), in line with the American dominance of the film and TV industry. Out of the remaining 28.3% subtitle files, nearly one-third (9.5%) corresponded to UK productions and around two-thirds (18.8%) to non-English productions, mostly French, German, and Spanish. Hence, most of the subtitles used were transcripts of the English language.

Researchers can freely access online the SUBTLEX-GR frequency norms at www.bcbl.eu/databases/subtlexgr. Once reaching this website, visitors will find four text files (.txt): (i) a file called SUBTLEX-GR_full.txt, corresponding to the entire corpus, including non-Greek characters and strings, (ii) a file called SUBTLEX-GR_CD.txt, which lists only the entries encountered in more than two different contexts (CD > 2), (iii) a file called SUBTLEX-GR_restricted.txt, which lists only correctly spelled existing Modern Greek words, and (iv) a file called SUBTLEX-GR_letters.txt, listing the letters of the Greek alphabet and their number of appearances within the subtitle files.

The SUBTLEX-GR_CD.txt file provides a "cleaner" version of the corpus than the one found in the SUBTLEX-GR_full.txt file, with 187,021 types and 26,981,066 tokens. By using a contextual diversity cut-off of two, most of the optical character recognition (OCR) mistakes and illegal strings were eliminated, while very low frequency entries corresponding to neologisms and supposedly illegal constructions were maintained (see also Keuleers et al., 2010, for a similar procedure). Even though we consider this to be a useful additional tool for specific word material selection, it should be noted that this version of the corpus is not error-free.

A completely error-free version of the corpus is given in the SUBTLEX-GR_restricted.txt file. This version resulted from cross-checking the corpus with a Modern Greek spell-checker which includes more than 1,600,000 inflected word forms (Symfonia software, ILSP). Through this process, spelling errors due to OCR mistakes and word types not found in the Symfonia software were removed. The rejected strings made up a total of 75.6% of the types, but importantly, only 16.5% of the tokens (see also Brysbaert and New, 2009 for a similar procedure). For the remaining 145,631 different word types, accounting for a total of 23,152,956 tokens, a set of additional frequency and lexical measures was also calculated. We strongly encourage researchers to use this version of the corpus when looking for word material, since it only contains legal and correctly spelled Modern Greek lexical entries and is thus much more manageable.

The information contained in each of the columns (from left to right) of the different versions of the word corpus is the following (se also $SUBTLEX_{US}$, Brysbaert and New, 2009; SUBTLEX-NL, Keuleers et al., 2010; for a similar lay-out):

1. ID: The number of each entry.
2. Word: The entries ordered alphabetically. Please note that in the SUBTLEX-GR_restricted.txt file the entries are presented starting with either an upper or a lower case letter depending on how they were encountered more times in the corpus. This was done so that researchers would be able to identify words that correspond mostly to proper names (or to words that are also used as names, e.g., *Ε*ιρηνικός [pacific], appears with an uppercase because it is mostly used as the ocean's name). We would recommend researchers to avoid using these words as experimental items since they have a rather special representational status.
3. FREQcount: The number of occurrences of each entry (raw frequency) in the subtitle files.
4. CD: The number of different contexts (films and television series) a word appeared in.
5. SUBTLEX_WF: The word frequency per million words with four digit precision. This value was calculated by multiplying the number of occurrences of an entry within the corpus (i.e., FREQcount) by a million and then dividing it by the number of tokens the database (in each of its different versions). This measure allows matching frequency values across different databases since it does not take into account the corpus size.
6. Lg10WF: This value corresponds to the log10(FREQcount + 1).
7. SUBTLEX_CD: The percentage of different contexts (films and television episodes) a word appeared in, with four digit precision.
8. Lg10CD: This is the log10(CDcount + 1) with four digit precision. According to our analyses, this is the most valid frequency measure for material selection (see also Keuleers et al., 2010).

The following additional columns are also included in the SUBTLEX-GR_restricted.txt file:

9. FREQlow: The number of times a word appears in the corpus starting with a lowercase letter. Brysbaert and New (2009) showed that for proper names this measure is more representative than the total number of appearances of the name (starting with an upper or a lowercase letter).
10. FREQupper: The number of times a word appears in the corpus starting with an uppercase letter. This measure does not take into account the number of times the entire word appeared written in uppercase letters.
11. *N*: Number of orthographic neighbors of each entry found within the restricted corpus. The influence of *N* on the reading performance has been repeatedly shown in different tasks (e.g., Carreiras et al., 1997), different populations (e.g., Duñabeitia and Vidal-Abarca, 2008; Duñabeitia et al., 2009a) and with different experimental techniques (e.g., Holcomb et al., 2002; Müller et al., 2010) and should thus be taken under consideration when selecting experimental materials.
12. OLD20: The Orthographic Levenshtein Distance 20 score (Levenshtein, 1966), is a measure of orthographic similarity between two words that stands for the minimum number of

substitutions, insertions, or deletions required to turn one word into the other. We opted for calculating the OLD20 values of the words included in the restricted version of the corpus, due to recent findings showing that OLD20 score is a better predictor of reading performance than the standard *N* measure (e.g., Yarkoni et al., 2008).

13. Length: The length of each entry counted in number of characters.
14. SUBTLEX_WF_full: The word frequency per million words with four digit precision, using as reference value the total number of tokens included in the full version of the corpus. It could be argued that (i) frequency values calculated on the entire "unclean" corpus would be more exact since they will be divided by the true total number of tokens, and (ii) that the relative frequencies would be also more representative since words found in the left-most part of the frequency distribution will be also included.