Quarterly Journal of Experimental Psychology (in press).

On the advantages of word-frequency and contextual diversity measures extracted from subtitles: The case of Portuguese

Ana Paula Soares[1], João Machado[1], Ana Costa[1], Álvaro Iriarte[2], Alberto Simões[2], José João de Almeida[3], Montserrat Comesaña[1] & Manuel Perea[4]

[1]Human Cognition Lab, CIPsi, School of Psychology, University of Minho, Portugal.

[2]Centre for Humanistic Studies, University of Minho, Portugal.

[3]Computer Science and Technology Center, University of Minho, Portugal.

[4]ERI-Lectura and Departamento de Metodología, Universitat de València, Valencia, Spain

Corresponding author:

Ana Paula Soares

Human Cognition Lab, CIPsi, School of Psychology,
University of Minho
Campus de Gualtar
4710-057 Braga, Portugal
E-mail: asoares@psi.uminho.pt
Phone: + 351 253604236

Abstract

We examined the potential advantage of the lexical databases using subtitles and present SUBTLEX-PT, a new lexical database for 132,710 Portuguese words obtained from a 78 million corpus based on film and television series subtitles, offering word-frequency and contextual diversity measures. Additionally we validated SUBTLEX-PT with a lexical decision study involving 1,920 Portuguese words (and 1,920 non-words) with different lengths in letters (M = 6.89, SD = 2.10) and syllables (M = 2.99, SD = 0.94). Multiple regression analyses on latency and accuracy data were conducted to compare the proportion of variance explained by the Portuguese subtitle-word frequency measures with that accounted by the recent written-word frequency database (P-PAL; Soares et al., 2014a). As its international counterparts, SUBTLEX-PT explains approximately 15% more of the variance in the lexical decision performance of young adults than P-PAL database. Moreover, in line with recent studies, contextual diversity accounted for approximately 2% more of the variance in participant´s reading performance than the raw frequency counts obtained from subtitles. SUBTLEX-PT is freely available for research purposes at http://p-pal.di.uminho.pt/about/database.

Keywords: word frequency; subtitles; Portuguese.

Running head: SUBTLEX-PT

Introduction

The number of times that a word occurs in a language (i.e., its frequency of use) is one of the most important variables in language processing, explaining more than 30% of the variance in word recognition and naming latencies (see, for example, Baayen, Feldman, & Schreuder, 2006; Balota et al., 2004; Brysbaert, & Cortese, 2011; Brysbaert et al., 2011a; Cortese & Khanna, 2007; Howes & Solomon, 1951; Keuleers, Diependaele, & Brysbaert, 2010a; Murray & Forster, 2004; Yap & Balota, 2009). Subsequently, word-frequency plays a central role in all current models of visual-word recognition and reading (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Davis, 2010; Engbert, Nuthmann, Richter, & Kliegl, 2005; Grainger & Jacobs, 1996; McClelland & Rumelhart, 1985; Plaut, McClelland, Seidenberg, & Patterson, 1996; Reichle, Pollatsek, Fisher, & Rayner, 1998).

However, the predictive validity of word frequency measures in psycholinguistic experiments has been recently questioned. In particular, recent studies have shown that the extent to which word frequency predicts the linguistic performance of individuals depends on the type of language register from which it is obtained (e.g., see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2007; Brysbaert & Cortese, 2011; Brysbaert & New, 2009; Brysbaert et al., 2011a; Brysbaert, Keuleers, & New, 2011b; Burgess & Livesay, 1998; New, Brysbaert, Veronis, & Pallier, 2007; Zevin & Seidenberg, 2002). Traditionally, word frequency has been obtained from the assembling of large amounts of written texts (e.g., books, periodicals) and by counting the number of times a word appears in these corpora. This approach was first adopted by Thorndike (1921) in the work entitled

"Teacher's Word Book", which provides a frequency list for 10,000 English words extracted from the manual compilation of English texts totaling the impressive number of 4,5 million words - updated in 1944 to 30,000 words with the collaboration of Lorge (Thorndike & Lorge, 1944). This approach was also followed by Kučera and Francis (1967), who compiled the first electronic corpus (the Brown corpus), which yielded the most widely used word frequency norms in English: the Kučera and Francis norms (1967; hereafter KF). In spite of its extensive use in psycholinguistic experiments since the 1970s, the predictive validity of the KF norms has been questioned in large-scale studies that collected lexical decision and/or word naming times for a vast number of words and then tested which of the word-frequency measures better predicts the speed and/or accuracy of the linguistic performance of individuals (e.g., Balota et al., 2004; Brysbaert & New 2009; Brysbaert et al., 2011a; Burgess & Livesay, 1998; Zevin & Seidenberg, 2002).

For instance, Brysbaert and New (2009) in a seminal study that originated the "SUBTLEX" movement (see also New et al., 2007), showed that the KF norms explained a significantly lower percentage of variance in word recognition times (6% less) and accuracy (10% less) than the frequency norms obtained from a corpus of approximately 51 million words obtained from American English films and television (TV) series subtitles (the SUBTLEX-US). The small size of the KF corpus (approximately 1 million words) as well as the fact that it is based on a limited number of dated samples of American English publications (about 500 samples of texts published in 1961) could explain these results. Indeed, Brysbaert and New (2009) for example recommend that the corpus should contain between 3,000 to 10,000 different text samples in order to be representative of the lexicon

of a language. Moreover, from a statistical point of view, extracting word frequency from a large corpus is also better because, as Lee (2003) pointed out, the standard error of the word counts varies as a function of the square root of the sample size (i.e., it gets smaller as the sample gets larger). Thus, extracting word frequency from a larger corpus allows for a more accurate measure of word frequency. Furthermore, larger corpus also allows for low-frequency words to be represented in the corpus and to establish finer and subtle distinctions between them (Burgess & Livesay, 1998). This is an increasingly important point since the recent works developed under the English Lexicon Project (Balota et al., 2007), the French Lexicon Project (Ferrand et al., 2010), the Dutch Lexicon Project (Keuleers et al., 2010a), and the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012) showed that almost the entire word frequency effect in the explanation of lexical decision times lies in word frequency intervals below 10 occurrences per million words ($Log_{10} = 1$), with the most significant effect being observed for words with a frequency between 0.1 ($Log_{10} = -1$) and 1 ($Log_{10} = 0$) per million words (see Brysbaert et al., 2011b and also Keuleers et al., 2012 for details).

Nonetheless, the outperformance of word frequency norms obtained from film and TV series subtitles over the traditional written-word frequency norms obtained from texts is also observed when larger and more recent corpora are considered, such as the British National Corpus (88 million words; Leech, Rayson, & Wilson, 2001), the Zeno corpus (17 million words; Zeno, Ivens, Millard, & Duvvuri, 1995), the Hyperspace Analogue to Language (HAL) (more than 130 million words; Lund & Burgess, 1996), the CELEX database (17.9 million words; Baayen, Piepenbrock, & van Rijn, 1993), or even Google's

Ngram Viewer Books measure, which is based on an impressive 131 billion word corpus from digitized American English books published since 1800 (Michel et al. 2011) (see Brysbaert & New, 2009 and Brysbaert et al., 2011b for details). Therefore, the advantage of subtitle-word frequency measures cannot be explained based only on the size of the corpus. Corpus representativeness, i.e., the extent to which the corpus includes a full range of linguistic samples that represent a language as a whole (see Sinclair, 2005), should also be considered.

Since most of word frequency norms are obtained from edited texts (e.g., novels, poetry, newspapers, technical writing) usually produced by professionals and skilled writers, one may wonder whether these texts provide "good" samples of the ordinary use that natives make of his/her language. Skilled writers have a special care with the way they write. Typically they use a more eloquent language and refrain from using repeated words which yield greater lexical diversity, and hence to a tendency to overestimate the frequency of rare words and to underestimate the frequency of more common words (Baayen 2001; Brysbaert & New, 2009; New et al., 2007). This affects language representativeness in these corpora, and consequently may generate a significant bias in the way frequency counts are obtained (see Baayen, 2001; Breland, 1996; Brysbaert et al., 2011a). Additionally, the fact that the texts included in these corpora are usually extensively revised, which is a highly time-consuming task, can also lead to underestimate words that have been recently introduced into the language and to overestimate words that are no longer in common usage. Thus, since the language used in film and TV subtitles approximates more closely to everyday language, is easily obtained from various Internet

sites (for a discussion about the legal issues involved in the use of subtitles for research purposes see Keuleers, Brysbaert, & New, 2010b), and captures language samples from "real" social situations and human interactions (which is more difficult in written traditional corpora), it is not surprising that frequency norms obtained from film and TV subtitles constitute an interesting alternative to the *de facto* language used by native speakers.

New, Brysbaert, Veronis, and Pallier (2007) were the first authors to empirically explore this idea. They compiled a 52 million-word corpus from 9,474 French films and TV series which was then validated by testing how well this new frequency measure predicted word processing times when compared to pre-existing word frequency measures for French, namely those from the spoken *Corpus du Référence du Français Parlé* [CRFP] (Equipe DELIC, 2004) and from the written corpus developed by New, Pallier, Brysbaert, & Ferrand (2004) for the same language. Results showed that frequency measures from film and TV subtitles explained approximately 10% more of the variance of the lexical decision times for the 240 French words collected by the authors with 17 native French participants and for the 234 words previously collected by Bonin, Charlard, Méot, & Fayol (2001).

The advantage of film and TV series subtitles over other frequency norms was immediately confirmed by Brysbaert and New (2009) for American English (SUBTLEX-US; see Brysbaert, New, & Keuleers, 2012 for an extension) using a broader pool of words (more than 30,000 words) from the English Lexicon Project (Balota et al., 2007). Since then similar databases were developed in other languages such as Chinese (SUBTLEX-

CH: Cai & Brysbaert, 2010), Dutch (SUBTLEX-NL: Keuleers et al., 2010b), Greek (SUBTLEX-GR: Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010), German (SUBTLEX-DE: Brysbaert et al., 2011a), Spanish (SUBTLEX-ESP: Cuetos, Glez-Nosti, Barbon, & Brysbaert, 2011; EsPal: Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) and British English (SUBTLEX-UK: van Heuven, Mandera, Keuleers, & Brysbaert, 2014). In all these languages subtitle word-frequency measures have proven to outperform written-word frequency norms, hence suggesting that the linguistic style of subtitled films and TV series is highly representative (at least more representative than the language used in written corpora) of the linguistic experience of young adults, particularly of the university populations traditionally recruited in psycholinguistic studies.

Recent reports show that the reading habits of young adults are declining. A report conducted in 2004 by the National Endowment for the Arts showed that young adults are more engaged in activities such as watching television, surfing the Web, listening to their iPods, talking on mobile phones, and messaging their friends than in reading activities. The number of 17-year-olds who never read for pleasure increased from 9% in 1984 to 19% in 2004. Almost a half of the Americans between ages 18 and 24 never read books for pleasure. Between ages 15 and 24 young adults spend between two and two and a half hours a day watching TV and seven minutes reading. Therefore, it is not surprising that subtitle word frequency norms account more significantly for the performance of young adults in word recognition and naming than the frequency measures obtained from written corpora (e.g., novels, poetry, newspapers, technical writing) to which young adults seem to be increasingly less exposed. Interestingly, the advantage of subtitle word frequency norms

is also observed when oral corpora are considered (see for example the CRFP corpus in New et al., 2007 work), and also in languages in which most films and TV series are not subtitled but dubbed (i.e. there are voices from native speakers overlapping the original voices), such as French, Spanish or German. In line with the proposals of recent models of visual-word recognition (see, for example, Ziegler, Petrova, & Ferrand, 2008), this seems to suggest that the language conveyed by audiovisual media affects word recognition and the reading performance of young adults irrespectively of the discursive modality (oral or written) used. Thus, extracting word frequencies from subtitles of film and TV series seems to be a highly valuable alternative to written-word counts, offering more reliable word frequency measures that are particularly suited for studies based on word latencies.

However, in spite of their relevance and availability for a growing number of languages (e.g., Chinese, Dutch, English-American, English-British, French, German, Spanish), word frequency norms from film and TV series subtitles are still nonexistent for Portuguese. Portuguese is a Romance language spoken by approximately 220 million people mainly in Portugal and Brazil, and also in Angola, Mozambique, Cape Verde, Guinea-Bissau, São Tomé and Principe, East Timor, Equatorial Guinea and Macau. There are also communities of Portuguese speakers in Goa, Daman and Diu in India, and in Malacca in Malaysia. Besides being considered one of the most spoken languages in the world (occupying the seventeenth position in the world language statistics, the third in the spoken European languages and the first in the languages spoken in the Southern Hemisphere), Portuguese presents several distinctive features from other alphabetic languages. For instance, in contrast to other Romance languages, Portuguese has a more

opaque writing system than European languages like Spanish or German, but it is less opaque than English or French. It is therefore considered a language of intermediate orthographic depth as far as the mapping between spelling and sound is concerned, which has a clear impact on reading and spelling acquisition (see for example Alegria et al., 2003; Goswami et al., 1998; Seymour et al., 2003). Furthermore, Portuguese is also distinguishable from other Languages such as Chinese, French, Italian or Spanish, since it is considered a stress-timed language with well defined syllable boundaries, highly diverse and complex syllable structures with vowel reduction processes (see Frota, Vigário, & Martins, 2002, for details). Hence, Portuguese is an interesting language for studying language representation and processing, and not only for studies on reading and spelling acquisition as so far developed. Therefore, having reliable frequency norms available for Portuguese such as the ones obtained from subtitles will constitute an important resource for the development of cross-linguistic studies that take advantage the characteristics of the Portuguese language. This will complement a series of lexical databases available in Portuguese for the conduction of cognitive and psycholinguistic research with adults (e.g., Soares et al., 2012, 2013a, 2014a) and children as well (Soares et al., 2014b; Comesaña et al., 2014).

Thus in this paper, we introduce SUBTLEX-PT, a new word frequency measure for 132,710 Portuguese words (wordforms) obtained from a 78,019,765 million word corpus based on 17,496 European Portuguese (EP) films and TV series subtitles screened between 1990 and 2011. In Portugal, like many other European countries, national film production is reduced, and all foreign films (mostly of American origin) are subtitled. Thus, compiling

a subtitle corpus for Portuguese is a relatively easy task to accomplish and can offer, as its counterparts, a valuable research tool for the Portuguese scientific community who uses verbal stimuli in their experiments, especially for those who works with word latencies.

In the following sections, we will detail the procedures adopted in the development of the SUBTLEX-PT corpus and in the computation of its frequency (raw counts) and contextual diversity (CD) measures. CD was firstly defined by Adelman, Brown, and Quesada (2006) as the number of documents a word appears in. It was then adopted to subtitle databases by Brysbaert and New (2009), who operationalized CD as the number of films or TV series in which a word appears. Brysbaert and New (2009) observed that CD accounts for approximately 4% more variance in word recognition times than the standard frequency measure per million words. This result has been systematically confirmed in different studies (e.g., Cai & Brysbaert, 2010; Dimitropoulou et al., 2010; Keuleers et al., 2010; Plummer, Perea, & Rayner, 2014; van Heuven et al., 2014), which suggests that the diversity of contexts in which a word appears, and not only the pure computation of the number of word occurrences *per se* (i.e., independently of the number of contexts), is the best predictor of the reading performance of young adults (see Perea, Soares, & Comesaña, 2013, for recent evidence with developing readers). It also includes the new Zipf scale frequency measure, recently proposed by van Heuven et al. (2014). This new measure accounts for the number of times a word appears in the corpus in a logarithm 7-point Likert scale and is assumed as a much easier way to understand word frequency. Indeed in the Zipf scale words frequency ranges from 1 to 7 points, with the values 1-3 indicating low-frequency words (with frequencies of 1 per million words and lower) and the values 4-7

indicating high-frequency words (with frequencies of 10 per million words and higher) (see van Heuven et al., 2014 for details).

After presenting SUBTLEX-PT, we validate this new word frequency database by testing to what extent SUBTLEX-PT will predict the lexical decision performance of Portuguese young adults (college students) better than the Portuguese written-word frequency norms obtained recently for Portuguese from the Procura-PALavras (P-PAL) database (see Soares et al., 2014a; available at http://p-pal.di.uminho.pt/tools). Until the beginning of the year 2000, frequency lexicons for Portuguese were based on small corpora such as the *Português Fundamental* corpus (Universidade de Lisboa, 1987), or the *Léxico Multifuncional Computorizado do Português Contemporâneo* (Bacelar do Nascimento, Pereira, & Saramago, 2000) for which word statistics were very limited. Recently, within the scope of the P-PAL project, Soares and colleagues extracted a new word frequency measure for contemporary EP based on a large corpus (more than 227 million words) composed essentially of written newspaper texts (see Soares et al., 2014a). The P-PAL written-word frequency measure will thus be used in this work to validate the new Portuguese subtitle word-frequency database presented in this paper. In line with the results obtained for French (New et al., 2007), and in the SUBTLEXs databases (-US: Brysbaert & New, 2009; -CH: Cai & Brysbaert, 2010; -NL: Keuleers et al., 2010; -GR: Dimitropoulou et al., 2010; -DE: Brysbaert et al., 2011a; -ESP: Cuetos et al., 2011; -UK: van Heuven et al., 2014), we expect that SUBTLEX-PT will be a better predictor of the word recognition performance of young adults, and thus will constitute a reliable resource for cognitive research.

Material and methods

Corpus sampling

A total of 17,496 European Portuguese subtitle files obtained from 8,506 films (49%) and 8,990 (51%) TV series provided by the Open Subtitles (OS) website (available at http://opus.lingfil.uu.se/OpenSubtitles_v2.php, see Tiedemann, 2009) constitute the raw material for SUBTLEX-PT. These subtitles were screened between 1990 and 2011, although the vast majority of films and TV subtitles were screened between 2000 and 2011 (72%).

In the SUBTLEX-PT corpus sampling different criteria were established. Firstly, only subtitles catalogued as film or TV series were included in the corpus (for instance subtitles from videogames were excluded). Films and TV series marked as damaged at the OS website were also excluded. Secondly, only film and TV subtitles unequivocally catalogued by the identification number (ID) in the Internet Movie Database (IMDb) according to genre, year and subtitle type were admitted. Thirdly, in order to avoid potential duplications in the corpus, only TV series subtitles containing information on the number and season of each episode were included. Fourth, because OS provides subtitles developed by individual users (which are therefore not revised), we analyzed potential orthographic errors by using the Portuguese spelling analyzer JSpell (Simões & Almeida, 2001) and by crosschecking these errors with the lexical entries from the P-PAL database (which contains approximately 208,000 wordforms). Film and TV series subtitles with a number of orthographic errors equal to or higher than 20% were excluded. Lastly, in order to obtain a diversified corpus we included subtitles from all the 26 film genres according to

the IMBD's classification (see http://www.imdb.com/genre/). Figure 1 shows the distribution of film and TV subtitles in SUBTLEX-PT.

<INSERT FIGURE 1>

As depicted in Figure 1, SUBTLEX-PT includes films and television series pertaining to all IMBD genres, although most of them are drama (21.2%). Subsequently, the most representative genres are: comedy (11%), thriller (10.2%), action (8%), crime (7.7%), mystery (7.2%), science-fiction (6%), adventure (5.3%), romance (5.3%) and fantasy (3.5%). The remaining genres are also featured in SUBTLEX-PT although less significantly as shown in Figure 2.

From the 17,496 Portuguese subtitle files incorporated in SUBTLEX-PT we identified a total of 77,981,300 space-separated tokens. In order to identify words (types), we implemented a similar strategy to the one used for P-PAL (see Soares et al, 2014a). Specifically, we eliminated proper nouns, isolated syllables, abbreviations (e.g., vol. [for the English word 'volume'] or art. [for the English word 'article']), symbols and unconventional orthographic forms (e.g., @ or €). Numerals, loanwords, and proper nouns with the same orthography as common nouns (e.g. the adjective *clara* [light] is also a proper noun) were maintained. Hyphenated words were also maintained, except verbs with clitic pronouns. These inflected forms are in fact a combination of two or more words in a compound verb form, and similar to P-PAL they were split into their constituents. For example, the verb form *preparavam-se* [they prepared themselves] was split into the verb

form *preparavam* [they prepared] and the clitic pronoun *se* [themselves]. The original frequency of the compound verb form was added to the final verb form and clitic pronoun. Occasionally, verb forms had to be fixed. For instance, verb forms ending with *á* and followed by a clitic pronoun (e.g., *encestá-la* [to score it]) were fixed by replacing *á* with *ar*, thus forming the original natural unhyphenated verb form *encestar* [to score], and subtracting the personal pronoun and the hyphen. Word frequencies were then added to the fixed verb form and to the clitic pronoun, which is also a lexical entry in SUBTLEX-PT. Contractions were split into their lexical constituents (e.g. *dele* [his] is a contraction of the preposition *de* [of] and the personal pronoun *ele* [him] and was split into *de* and *ele*) and their original frequencies were assigned to each lexical item. Multiword items, i.e., unhyphenated words such as phrases, idioms and collocations, were also split into their lexical constituents and the original frequency value was added to each item. There is one single entry for nonhomophonic homographs (e.g., *sede* ['sedə], Portuguese word for thirst, and *sede* ['sɛdə], Portuguese word for headquarters) and homonyms (e.g., *castanha* [noun], Portuguese word for chestnut, and *castanha* [adjective], the feminine for brown). Based on this procedure, SUBTLEX-PT comprises a total of 132,710 different wordforms (types) from a corpus of approximately 78 million words.

The size of the SUBTLEX-PT corpus is comparable to other subtitle corpus with the exception of the EsPal, which comprises 244,933 words (types) obtained from a 460 million corpus based on 138,783 films and TV series (see Duchon et al., 2013) and the SUBTLEX-UK, which comprises 201,700 million words obtained from 49,099 BBC broadcasts (see van Heuven et al., 2014). The remaining subtitle databases contain a

substantially smaller corpus. The first subtitle database developed for French was based on a 52 million corpus from 9,479 films and TV series (New et al., 2007), the SUBTLEX-US (Brysbaert & New, 2009) was based on a 51 million corpus from 8,388 films and TV series, the SUBTLEX- GR (Dimitropoulou et al., 2010) was based on a 27 million corpus from 6,032 films and TV series, the SUBTLEX-NL (Keuleers et al., 2010) was based on a 44 million corpus from 8,443 films and TV series, the SUBTLEX-CH (Cai & Brysbaert, 2010) was based on a 33,5 million corpus from 7,148 films and TV series, the SUBTLEX-ESP (Cuetos et al., 2011), was based on a 41 million corpus from 3,780 films and TV series, and lastly the SUBTLEX-DE (Brysbaert et al., 2011a) was based on a 25 million corpus from 4,610 films and TV series. Nevertheless, as shown by Brysbaert & New (2009), for corpora larger than 30 million words the advantage in the explanation of linguistic performance is not significant.

SUBTLEX-PT database

The SUBTLEX-PT database can be downloaded at http://p-pal.di.uminho.pt/about/databases as an excel file. Following each of its 132,710 lexical entries (wordforms) the SUBTLEX-PT file contains 12 columns that provide several grammatical, sublexical and frequency data taken from P-PAL (available at http://p-pal.di.uminho.pt/tools). Specifically, SUBTLEX-PT provides four raw frequency measures for each of its wordforms (type): number of occurrences in the corpus ($SUBTLEX\_FREQ_{count}$), number of occurrences per million words ($SUBTLEX\_FREQ_{mil}$), and Base 10 logarithm (LOG10), computed from $FREQ_{count}+1$ ($SUBTLEX\_LOG10_{freq}$).

Adding 1 to the number of occurrences in the corpus makes it possible to match stimuli from different corpora when a stimulus is not present in a corpus, as recommended by Brysbaert and Diependaele (2013). We also computed the new standardized Zipf scale frequency measure [SUBTLEXZipf] proposed by van Heuven et al. (2014). This measure is very similar to the SUBTLEX_LOG10$_{freq}$ but as mentioned above is an easier way to understand word frequency since Zipf values range from 1 to 7 (with values 1-3 indicating low-frequency words and values 4-7 indicating high-frequency words; see van Heuven et al., 2014 for details).

Similar to SUBTLEX-US (Brysbaert & New, 2009; Brysbaert et al., 2012), SUBTLEX-NL (Keuleers et al., 2010), SUBTLEX-GR (Dimitropoulou et al., 2010), SUBTLEX-CH (Cai & Brysbaert, 2010) and SUBTLEX-UK (van Heuven et al., 2014), SUBTLEX-PT provides three Contextual Diversity (CD) measures: number of different films and TV series in which the word appears (SUBTLEX_CD$_{count}$), the percentage of films and TV series in which the word appears (SUBTLEX_CD$_{\%}$) and LOG$_{10}$ of number of different films and TV series in which the word appears + 1 (SUBTLEX_LOG10$_{CD}$). SUBTLEX-PT also provides the following information from the P-PAL database: number of letters in the word (N$_{lett}$), number of syllables in the word (N$_{syll}$), written-word frequency (per million words) (P-PAL$_{freq}$), LOG$_{10}$ P-PAL frequency counts computed from FREQ$_{count}$+1 (P-PAL_LOG10$_{freq}$) and Part-of-Speech (PoS) information. Similar to P-PAL, content and function words in SUBTLEX-PT cover the following PoS categories: nouns (N), adjectives (ADJ), verbs (V), adverbs (ADV), conjunctions (CONJ), determiners (DET), interjections (INT), quantifiers (QUANT), prepositions (PREP), and pronouns

(PRON) (see Soares et al., 2014a, for details about PoS categorization). Because syntactic ambiguity is very common in Portuguese, where words like *ilustrado* [illustrated] can be used both as a verb form and an adjective, SUBTLEX-PT includes all grammatical classes the word has been assigned to in P-PAL according to their frequency of occurrence. PoS tags are comma separated.

Testing SUBTLEX-PT frequencies

To empirically validate SUBTLEX-PT, we conducted a lexical decision study involving 1,920 Portuguese words (and 1,920 non-words). Then we performed multiple regression analyses on latency and accuracy data to compare the proportion of variance accounted by the Portuguese subtitle-word frequency presented in this paper (SUBTLEX-PT) with the proportion of variance accounted by the Portuguese written-word frequency provided by P-PAL database (see Soares et al., 2014a; available at http://p-pal.di.uminho.pt/tools). We chose the lexical decision task because it is the most common task used in studies aiming to test the quality of subtitle word frequency measures (e.g., Brysbaert & New, 2009; Brysbaert et al., 2011a; Cuetos et al., 2011; Dimitropoulou et al., 2010; Keuleers et al., 2010; New et al., 2007; van Heuven et al., 2014) and also because it is highly sensitive to word-frequency effects, as shown by Balota et al. (2004).

The 1,920 Portuguese words selected to integrate the lexical decision study to validate SUBTLEX-PT were obtained from a pool of 3,800 words for which we are collecting subjective norms of imageability, concreteness, and subjective frequency (Soares et al., 2013b). The 1,920 words selected occur simultaneously in SUBTLEX-PT

and in P-PAL databases and present different lengths in number of letters (from 2 to 12 letters), number of syllables (from 1 to 6 syllables) and per million word frequency (from low, medium and high frequency intervals). Though most of SUBTLEX studies were conducted with one- or two-syllable long words (see, for example, Brysbaert & New, 2009; Keuleers et al., 2010; van Heuven et al., 2014) there are growing demands in the literature (Yap & Balota, 2009, for example) towards the use of multisyllable words in psycholinguistics studies. Moreover, considering the characteristics of Portuguese, the vast majority of EP words extend beyond one-syllable long. For example, in the P-PAL wordform database (which integrates approximately 208,000 words) only 641 words present one-syllable (0.3% of the total lexicon), while 14,359 words present two-syllables (7% of the total lexicon), 47,162 present words three-syllables (22.7% of the total lexicon) and 145,452 words present more than three-syllables (70% of the total lexicon). Therefore, the pool selected for the empirical validation of SUBTLEX-PT included a more lexically diverse data set of words than the previously SUBTLEX studies in order to represent more closely the lexical diversity of the Portuguese language. From the total set of 1,920 words selected, 553 words (28%) pertain to the short word group (i.e., length varies between 2 and 5 letters and 1 and 2 syllables), 948 (49.4%) are medium words (length varies between 6 and 8 letters and 3 to 4 syllables) and 419 (21.8%) are long words (length varies between 9 and 12 letters with more than 4 syllables length). Moreover, for each of these word lengths we assured the existence of low-frequency words (<10 occurrences per million words), medium-frequency words (11-74 occurrences per million words) and high-frequency words (≥ 75 occurrences per million words) in each corpus. Specifically, as far

as the P-PAL corpus is concerned, in the short words group, 117 words were low-frequency words (21.2%), 272 were medium-frequency words (49.2%) and 164 were high-frequency words (29.7%). In the medium words group, 219 were low-frequency words (23.1%), 497 were medium-frequency words (52.4%) and 232 were high-frequency words (24.5%). In the long words group, 94 were low-frequency words (22.4%), 236 were medium-frequency words (56.3%) and 89 were high-frequency words (21.2%). Regarding the SUBTLEX-PT corpus, in the short words group, 96 words were low-frequency words (17.4%), 275 were medium-frequency words (49.7%) and 182 were high-frequency words (32.9%). In the medium words group, 318 were low-frequency words (33.5%), 485 were medium-frequency words (51.2%) and 145 were high-frequency words (15.3%). Lastly, in the long words group, 225 were low-frequency words (53.7%), 174 were medium-frequency words (41.5%) and 20 were high-frequency words (4.8%).

It is also worth noting that although Brysbaert & New (2009), Keuleers et al. (2010) and van Heuven et al. (2014) tested their subtitle-word frequency measures considering a sample of thousands of words (obtained from the English Lexicon Project - Balota et al., 2007; the Dutch Lexicon Project - Keuleers et al., 2010; and the British Lexicon Project - Keuleers et al., 2012, respectively), other authors test the predictive validity of their subtitle norms using a much more restricted set of items. For example New et al. (2007) validated the French lexical database using lexical decision times for 240 words, Dimitropoulou et al. (2010) validated the Greek lexical database using lexical decision times for 172 words, Brysbaert et al. (2011a) validated the German lexical database using lexical decision times for 455 words, and Duchon et al. (2013) validated the

subtitle norms in EsPal based on the word naming times of 240 words provided by Cuetos & Barbón (2006) and on the picture naming times of 139 words from Cuetos, Ellis, and Álvarez (1999). Therefore, the pool of words selected for the lexical decision study to validate SUBTLEX-PT seems to be suitable, not only because of the number of words it contains, but especially because of its lexical diversity.

Participants

A total of 58 Portuguese college students (52 women and 6 men) from the University of Minho, with ages between 18 and 23 years (M = 21.3; SD = 3.06) took part in the experiment. All participants had normal (or corrected-to-normal) vision and were native speakers of European Portuguese. The majority of participants were right-handed (92%). Participants received course credit for their participation in the experiment. The experiment was conducted with the approval of the Ethics Committee for Human Research of the University of Minho (Braga, Portugal). Written consent was obtained from the all the participants.

Materials

Stimuli consisted of 1,920 Portuguese words that vary in number of letters (M = 6.9, SD = 2.10, *range*: 2 to 12 letters), number of syllables (M = 2.99, SD = 0.94, *range*: 1 to 6) and per million frequency both in P-PAL (M = 67.33, SD = 110.83, *range*: 0.44 to 1,214.45) and in SUBTLEX-PT (M = 61.41, SD = 142.33, *range*: 0.09 to 1,907.57) databases. Although word frequency values for these 1,920 words in P-PAL and

SUBTLEX-PT databases are very similar, $t(3,838) = 1.44$, $p = 0.15$, the *Pearson* correlation between them is only moderately significant, $r = 0.51$, $p < .001$. It should be further noted that these 1,920 words present CD values varying between 7 (0.04%) and 8,960 (51%) (M = 1,372 films and TV series) in SUBTLEX-PT. CD values were not presented for P-PAL since these values were not provided in that database.

For the purpose of the lexical decision task a set of 1,920 Portuguese legal nonwords matched in number of letters ($M = 6.9$, $SD = 2.09$ *range*: 2 to 12) and syllables ($M = 2.99$, $SD = 0.94$, *range*: 1 to 6) with the experimental words were also created. Nonwords were created by replacing one or two letters in the medial and final positions of the base-words (e.g., the nonword *utobia* was created from the base-word *utopia* [utopia] originates), respecting the phonotactic restrictions of Portuguese. Both words (1,920) and nonwords (1,920) were distributed across four experimental lists with 960 stimuli each (480 words and 480 nonwords). Each list had a similar number of words from different lengths (short, medium, long) and frequency intervals (low, medium, high) in order to ensure that all word types were included in each experimental list (list 1: $\chi^2(4) = 3.24$, $p = .52$; list 2: $\chi^2(4) = 4.60$, $p = .33$; list 3: $\chi^2(4) = 3.76$, $p = .44$; and list 4: $\chi2(4) = .38$, p = .98). Nonwords were also balanced across lists. Forty-eight practice items were created (24 words and 24 nonwords) to familiarize participants with the task. They were evenly distributed across the four experimental lists (6 words and 6 nonwords per list). The four lists of stimuli were assessed by each participant in four experimental sessions separated by a one-week interval. List presentation was counterbalanced across participants (24

different orders). Participants were randomly assigned to each order of list presentation (approximately two participants per list).

Procedure

The experiment was run individually in a soundproof booth. Presentation of stimuli and recording of responses were controlled by DMDX software (Forster & Forster, 2003). Participants were asked to decide as quickly and accurately as possible if the string of letters presented at the center of the computer screen was or was not a real word in Portuguese (i.e., a lexical decision task – LDT). If participants considered that the string of letters was a real word in Portuguese they should press the "Z" key on the keyboard ("sim"[yes] response). Conversely, if they considered that it was not a real word in Portuguese they should press the "M" key on the keyboard ("não"[no] response). Both speed and accuracy were stressed in the instructions.

The task comprises responses to 960 trials. Each trial consisted of a sequence of two visual events. The first was a fixation point (+) presented at the center of the computer screen for 500 ms. The fixation point was immediately replaced by the stimulus (word or nonword) at the center of the computer screen and remained on screen until participants responded or until 2,500 ms had elapsed. The order of the stimuli was randomized per participant. Participants were informed of the existence of several pauses (12) during the experiment (every 80 trials). Prior to the 960 experimental trials, participants received 12 practice trials (6 words and 6 nonwords). The task was performed by participants four times (each time a different experimental list), separated by a week interval. Each

experimental session lasted approximately 45 minutes. The entire procedure lasted about 3 hours per participant.

RESULTS AND DISCUSSION

Multiple regression analyses on latency and accuracy data were conducted in order to compare the proportion of variance accounted by the Portuguese subtitle-word frequency measures (SUBTLEX-PT) with that accounted for by the written-word frequency provided by P-PAL. LOG10 and $LOG10^2$ were considered as predictors from the P-PAL and the SUBTLEX-PT databases. For SUBTLEX-PT we also considered the LOG10 and $LOG10^2$ from CD and Zipf measures. CD and Zipf measures were not introduced in the analysis for P-PAL since they were not provided in the database. LOG10 and $LOG10^2$ were computed and introduced in the regression analysis as predictors because previous studies on subtitle-word frequencies had also used these measures (e.g., see Brysbaert & New, 2009; Dimitropoulou et al., 2010; Keuleers et al., 2010) since Balota et al. (2004; see also Baayen et al., 2006) found that the relationship between LOG frequency and word latencies is not completely linear and is better captured by the LOG square value.

In the dataset, the mean accuracy (percentage of correct answers) of the participants was 97% (*SD* = 5.37) for words and 96% (*SD* = 5.11) for nonwords. The lexical decision times for correct responses were 569.8 msec (*SD* = 43.63) for words and 650.8 msec (*SD* = 53.25) for nonwords. To ensure that extreme response latencies did not disproportionately influence the item's reaction time (RT) for the correct responses, we first eliminated any

response latencies faster or slower than 2,500 msec. Secondly, any RTs below or above 3 *SD*s from the mean raw RT of each participant were also excluded. Furthermore, eleven words were eliminated from the analyses since they were assessed by more than a third of the participants as nonwords (e.g., words such as *cárcere*[prison cell], *asilo*[asylum], credor[creditor], *esófago*[esophagus], *lodo*[sludge] and *zelo*[zeal] which were low-frequency words both in P-PAL and SUBTLEX-PT). Thus, a total of 1,909 words were considered in the RT and accuracy data. Table 1 shows the percentages of variance (R^2) in lexical decision times (RT) and accuracy (Acc) accounted for by the P-PAL word frequency [wf] and the SUBTLEX-PT word frequency [wf], contextual diversity [cd] and Zipf [zipf] measures.

<INSERT TABLE 1>

Concerning P-PAL, Table 1 shows the % of variance explained when the LOG10 entered as a predictor in the regression analysis both for RT, $F(1, 1907) = 428.44$, $p < .001$, and accuracy data, $F(1, 1907) = 163.78$, $p < .001$; and then when both LOG10 and $LOG10^2$ entered in the RT, $F(2, 1906) = 229.83$, $p < .001$, and accuracy, $F(2, 1906) = 105.85$, $p < .001$ analyses. The LOG10 measure explained 19.6% of the variance in RT and 7.9% of the variance in the accuracy data. Adding $LOG10^2$ in the regression equation increases significantly the percentage of variance explained both in RT (21.2%) and accuracy (10%) (*F* change, $p < .001$), in line with previous findings in other languages (e.g., Brysbaert & New, 2009; Dimitropoulou et al., 2010; Keuleers et al., 2010).

Concerning SUBTLEX-PT word frequency with LOG10 as predictor for RT, $F(1, 1907) = 999.37$, $p < .001$ and accuracy data, $F(1, 1907) = 163.53$, $p < .001$, table 1 shows that the LOG10 subtitle-word frequency explains 34.7% of the variance in RTs and 7.9% of the variance in the accuracy data. As for P-PAL measures, adding $LOG10^2$ in the regression significantly increases the percentage of variance explained both in RT (36.1%), $F(2, 1906) = 525.45$, $p < .001$, and accuracy (9.3%), $F(2, 1906) = 98.17$, $p < .001$ data (F change, $p < .001$). Compared to P-PAL, SUBTLEX-PT raw frequency explains a significantly higher percentage of variance in RTs (14.9% more), thus showing that in Portuguese, like in other languages (e.g., Chinese, Dutch, English-American, English-British, French, German, Greek, Spanish), word frequencies extracted from subtitles are indeed a better predictor of the reading times of young adults (college students) than written-word frequencies obtained from texts even when a more lexically diverse set of words was considered. In the accuracy data P-PAL captures a slightly 0.7% more of variance than SUBTLEX-PT when considering the two word frequency measures – although for the LOG10 measure P-PAL and SUBTLEX-PT accounted for exactly the same percentage of variance (7.9%).

It should be noted that when we take three word-frequency intervals into account (i.e., low-frequency words [<10 occurrences per million words], medium-frequency words [11-74 occurrences per million words] and high-frequency words [≥ 75 occurrences per million words]), the linear correlation coefficients between LOG10 frequency and RT (or Acc) showed that SUBTLEX-PT outperformed P-PAL in all frequency intervals (all $p_s < .001$), regardless of whether these intervals were based on P-PAL or on the SUBTLEX-PT

corpus. In particular, when classifying words in terms of low, medium, and high-frequency intervals in the P-PAL corpus, the *Pearson* coefficients between LOG10_SUBTLEX-PT and RTs were -.50, -.52, and -.48, respectively (all $p_s < .001$), whereas the *Pearson* coefficients between LOG10_P-PAL and the RTs were -.32, -.19, and -.17, respectively (all $p_s < .001$). Similarly, the *Pearson* coefficients between LOG10_SUBTLEX-PT and the Acc data were -.29, and -.20 for the low and medium-frequency intervals, respectively (all $p_s < .001$), whereas the *Pearson* coefficients between LOG10_P-PAL and the Acc data were -.23 and -.14 for the low and the medium-frequency intervals, respectively (all $p_s < .001$) – note that the accuracy data for high-frequency words did not reach statistical significance, thus reflecting a ceiling effect. The pattern of results in the RT data was very similar when word-frequency intervals were based on the SUBTLEX-PT corpus (the *Pearson* coefficients between LOG10_SUBTLEX-PT and RTs for low, medium, and high-frequency intervals were -.34, -.31, and -.23 respectively – all $p_s < .001$; whereas for the LOG10_P-PAL measure were -.26, -.21, and -.14, respectively – all $p_s < .001$), although in the Acc data P-PAL showed a slight advantage in line with the abovementioned results (*Pearson* coefficients between LOG10_P-PAL and the Acc for low- and medium-frequency intervals were -.28, and -.16, respectively - all $p_s < .001$; while for the LOG10_SUBTLEX-PT were -.19 and -.10, respectively – all $p_s < .001$ - again in Acc the results for the high-frequency words did not reach statistical significance). In sum, when word frequency is considered in separate categories, SUBTLEX-PT outperforms P-PAL. Remarkably, this was the case even in low-frequency words, where P-PAL might be

expected to have some advantage due to its larger corpus size (Burgess & Livesay, 1998; Lee, 2003).

Furthermore, it is worth noting that the results of the regression analyses conducted showed that the Portuguese subtitle word-frequency accounted for the highest difference in young adults' reading performance (approximately 15%) when compared to those obtained from other written-word frequency measures in other languages. For example, for Greek SUBTLEX-GR accounted for 10.5% more variance in reading latencies than GreekLex (see Dimitropoulou et al., 2010) and for Dutch SUBTLEX-NL accounted for 8% more variance than CELEX (see Keuleers et al., 2010), countries that also use subtitles extensively. This result may indicate that in Portugal young adults were more exposed to the language register conveyed by audiovisual media (films and TV series) than young adults from other countries. Although the reading habits of the Portuguese population has increased over the past decades as evidenced by comparing the data from the 2007 study "*A leitura em Portugal*"[Reading in Portugal] (Santos, Neves , Lima , & Carvalho , 2007) with the data from the 1997 study "*Hábitos de leitura: um inquérito à população portuguesa*"[Reading habits: a Portuguese population survey] (Freitas, Casanova, & Alves, 1997), at European level, the Portuguese levels of reading habits are still low. For instance in the study conducted by the European Social Survey (2002-2008) the total number of Portuguese who claim "not spending any time reading newspapers per day" is 12.4% above the average of the European Union (which in 2008 was situated in 36.3%). Importantly, more than 86% of the Portuguese population stated watching television for

more than one hour per day while only 12% report spending more than one hour per day reading (see Santos et al., 2007 for details).

Table 1 also shows results for the CD measure in SUBTLEX-PT when LOG10 entered as a predictor in the RT, $F(1, 1907) = 1119.92$, $p < .001$ and accuracy analysis $F(1, 1907) = 195.21$, $p < .001$, and then when LOG10) + $LOG10^2$ entered as predictors again in the RT $F(2, 1906) = 566.13$, $p < .001$ and in the accuracy, $F(2, 1906) = 108.27$, $p < .001$ analyses. LOG10 CD measure explains 37.5% of the variance in RTs and 9.3% of the variance in the accuracy data. Adding $LOG10^2$ significantly increases the percentage of variance explained both in RT (37.9%) and accuracy (10.2%) data (F change, $p < .001$). Compared to the SUBTLEX-PT word frequency measure, the SUBTLEX-PT CD measure explains 1.8% more of variance in RT and approximately 1% more of variance in accuracy which is in line with the findings obtained for English (American: Brysbaert & New, 2009; British: van Heuven et al., 2014), Greek (Dimitropoulou et al., 2010), Dutch (Keuleers et al., 2010) and Chinese (Cai & Brysbaert, 2010). The additional variance accounted for by CD measures is significant if we consider, as van Heuven et al. (2014) highlighted, that many variables explain less than 1% of the variance once other variables such as word frequency, length, and neighborhood statistics are partialled out (Brysbaert & Cortese, 2011; Brysbaert et al., 2011a).

Introducing the new standardized Zipf scale in the regression analysis revealed essentially the same results as the ones previously observed for the $SUBTLEX_{wf}$ measure. Indeed when Zipf is introduced as predictor for both RT, $F(1, 1907) = 1021.27$, $p < .001$ and accuracy data, $F(1, 1907) = 163.77$, $p < .001$, table 1 shows that the Zipf measure

explains the same 34.7% of the variance in RTs and 7.9% in the accuracy data. As expected, adding $Zipf^2$ in the regression significantly increases the percentage of variance explained both in RT (36 %), $F(2, 1906) = 539.83$, $p < .001$, and accuracy (9.3%), $F(2, 1906) = 97.98$, $p < .001$ data (F change, $p < .001$) which is fundamentally the same as the percentage captured by the LOG 10 + $LOG10^2$ of $SUBTLEX_{wf}$ (in RT, $SUBTLEX_{wf}$ measure explains 0.1% more of the variance).

Thus, the results obtained from the empirical validation of the SUBTLEX-PT showed that in Portuguese, like in other Indo-European languages (e.g., Dutch, English-American, English-British, French, German, Greek, Spanish) and in Sino-Tibetan languages (e.g., Chinese), the number of different contexts (i.e., films/TV series) in which a word appears is indeed the most effective predictor of the word recognition of young adults (college students), explaining the higher percentage of the variance in the word latencies (approximately 38%). Note that the percentage of variance captured by the SUBTLEX-PT CD measures in the accuracy data is also very close to the one captured the P-PAL raw frequency (10% vs. 10.2%, respectively), which was the highest percentage of variance captured in this dataset. Therefore, the CD measure in Portuguese –as in other languages– should be preferable over other word frequency measures (particularly from those extracted on written-word raw counts) when selecting stimuli for experiments focused on word identification latencies. Although exploring the mechanisms that underlie such effect is beyond the scope of the present study, it is important to highlight that current models of visual word recognition (e.g., Coltheart et al., 2001; Davis, 2010; Engbert, et al., 2005; Grainger & Jacobs, 1996; McClelland & Rumelhart, 1985; Plaut et al., 1996; Reichle

et al., 1998) should be modified to account for the fact that the diversity of contexts in which a word appears, and not only the frequency *per se* (i.e., independently of the number of contexts), affects word recognition (see Johns, Gruenenfelder, Pisoni, & Jones, 2012, Plumer et al., 2014, for suggestions on how these models can account for CD effects).

CONCLUSION

In the present study, we presented a new lexical frequency measure for Portuguese based on subtitles of films and TV series. As its counterparts in other languages, SUBTLEX-PT explains more variance (16.1% more) in the word recognition times of Portuguese young adults (college students) than the written-word frequency measures obtained from P-PAL, largely based on newspaper corpora (see Soares et al., 2014a for details) and 0.2% more of variance than P-PAL in the accuracy data. Overall, results showed that with a more lexically diverse set of words, SUBTLEX-PT CD measures outperforms the SUBTLEX-PT raw counts measures both in RTs (explaining 1.8% more of the variance) and accuracy data (explaining 0.9% more of the variance) of the reading performance of young adults, in line with the data previously found in English (American and British) and in other languages such as Greek, Chinese, Dutch, German or Spanish.

Compared with the P-PAL, SUBTLEX-PT frequencies represent a significant improvement in explained variance in RTs, thus constituting a valuable resource for cognitive studies based on verbal materials. Although the lexical information contained in the P-PAL lexical database remains invaluable, the SUBTLEX-PT word frequencies, particularly CD measures, should be preferred over the P-PAL written-word frequencies

when selecting stimuli for experiments based on word latencies. Since SUBTLEX-PT outperforms P-PAL it is important that future extension of the Portuguese subtitles database presented here should consider the possibility of computing other lexical and sublexical statistics (e.g., orthographic, phonological, neighborhood, syllable, trigram bigram, and biphone measures) as the ones provided by the P-PAL web application (see http://p-pal.di.uminho.pt/tools), in line with what was recently developed for Spanish with the EsPal database (Duchon et al., 2013). SUBTLEX-PT is freely available for research purposes at http://p-pal.di.uminho.pt/about/database.

Author Note:

This work is part of the research project “Procura Palavras (P-Pal): A software program for deriving objective and subjective psycholinguistic indices for European Portuguese words” (PTDC/PSI-PCO/104679/2008), funded by FCT (Fundação para a Ciência e Tecnologia), and FEDER (Fundo Europeu de Desenvolvimento Regional) through the European programs QREN (Quadro de Referência Estratégico Nacional), and COMPETE (Programa Operacional Factores de Competitividade).

References

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

Alegria, J., Marin, J., Carrillo, S., & Mousty, P. (2003). Les premiers pas dans l'acquisition de l'orthographe en function du caractère profound ou superficial du système alphabétique : comparaison entre le français et l'espagnol. In M. N. Romdhane, J.-E. Gombert & M. Belajouza (Eds), L'apprentissage de la lecture: Perspectives comparatives (pp. 51-67). Rennes: Presses Universitaires de Rennes

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. Brazilian Journal of Applied Linguistics, 11, 295-328

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. Journal of Memory and Language, 53, 496-512.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1993). The CELEX lexical database. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bacelar do Nascimento, M. F., Pereira, L. A. S., & Saramago, J. (2000). Portuguese Corpora at CLUL. In Proceedings of the Second International Conference on Language Resources and Evaluation (pp. 1603-1607). Athens, Greece.

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., …Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39, 445-459.

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2001). Age-of-acquisition and word frequency in the lexical decision task: Further evidence from the French language. Current Psychology of Cognition, 20(6), 401-443.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. Psychological Science, 7, 96-99.

Brysbaert, M. & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? Quarterly Journal of Experimental Psychology, 64, 545-559

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. Behavior Research Methods, 45, 422-430

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41, 977-990.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011a). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. Experimental Psychology, 58, 412-424.

Brysbaert, M., Keuleers, E., & New, B. (2011b). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. Frontiers in Psychology, 2:27.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding Part-of-Speech information to the SUBTLEX-US word frequencies. Behavior Research Methods, 44, 991-997.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting RT in a basic word recognition task: Moving on from Kučera and Francis. Behavior Research Methods, Instruments, & Computers, 30, 272–277.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies ased on film subtitles. PLoS ONE, 5, e10729.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. Psychological Review, 108, 204-258.

Comesaña, M., Fraga, I., Moreia, A. J., Frade, C. S., & Soares, A. P. (2014). Free associate norms for 139 European Portuguese words for children from different age groups. Behavior Research Methods. DOI:10.3758/s13428-013-0388-0.

Cuetos, F., & Barbón, A. (2006). Word naming in Spanish. European Journal of Cognitive Psychology, 18(03), 415–436.

Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. Behavior Research Methods, 31(4), 650–658.

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. Psicologica, 32, 133-143.

Davis, C. J. (2010). The spatial coding model of visual word identification. Psychological Review, 117, 713-758.

Dimitropoulou, M., Duñabeitia, J.A., Avilés, A., Corral, J., & Carreiras, M., (2010). Subtitle-based word frequencies as the best estimate of reading behaviour: The case of Greek. Frontiers in Psychology, 1:218, 1-12.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., Carreiras, M. (2013). EsPal: One-stop Shopping for Spanish Word Properties. Behavior Research Methods, 45, 1246-1258.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl. R. (2005). SWIFT: A dynamical model of saccade generation during reading. Psychological Review, 112. 777-813.

European Social Survey (2008). ESS4-European Social Survey 2002/2008. Available at www.europeansocialsurvey.org/.

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

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, 35(1), 116–124.

Freitas, E., Casanova, J. L., & Alves, N. A. (1997). Hábitos de leitura: Um inquérito à população portuguesa. Lisboa: Dom Quixote.

Frota, S., Vigário, M., & Martins, F. Frota (2002). Language discrimination and rhythm classes: Evidence from Portuguese. In B. Bel & I. Marlien (Eds.), Proceedings of

the 1st International Conference on Speech Prosody (pp. 315-318). Université de Provence: Laboratoire de Parole et Language: Aix-en-Provence, France.

Goswami, U., Gombert, J.E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French and Spanish. Applied Psycholinguistics, 19,19-52.

Grainger, J. & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. Psychological Review, 103, 518-565.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word probability. Journal of Experimental Psychology, 41, 401-410.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. Journal of the Acoustical Society of America, 132:2, EL74-EL80.

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

Keuleers, E., Diependaele, K. & Brysbaert, M. (2010a). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. Frontiers in Psychology, 1:174. doi: 10.3389/fpsyg.2010.00174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. Behavior Research Methods, 44, 287-304

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

Lee, C. J. (2003). Evidence-based selection of word frequency lists. Journal of Speech-Language Pathology and Audiology, 27(3), 172-175.

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

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instrumentation, and Computers, 28, 203-208.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of Basic Findings. Psychological Review, 88, 375-407.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K.; Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331, 176–182.

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

National Endowment for the Arts (NEA, 2004). Reading at risk: A survey of literary reading in America. National Endowment for the Arts. NW, Washington, DC.

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

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. Behavior Research Methods, Instruments, & Computers, 36(3), 516-524.

Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word-identification times in young readers. Journal of Experimental Child Psychology, 116, 37-44.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review, 103, 56-115.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 275-283.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. Psychological Review, 105, 125–157.

Santos, M. L., Neves, J., Lima, M. J., & Carvalho, M. (2007). A leitura em Portugal. Lisboa: Gabinete de Estatística e Planeamento da Educação (GEPE).

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. British Journal of Psychology, 94,143-174.

Simões, A. M., & Almeida, J. J. (2001). Jspell. In Actas do Encontro Nacional da Associação Portuguesa de Linguística. Lisboa: Associação Portuguesa de Linguística.

Sinclair, J. (2005). Corpus and text: Basic Principles. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 1-16). Oxford, UK: Oxbow Books.

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. Behavior Research Methods, 44(1), 256-269. DOI: 10.3758/s13428-011-0131-7.

Soares, A. P., Costa, A., Machado, J., Silva, A., Oliveira, J., Gonçalves, A. M., & Comesaña, M. (2013b). Subjective frequency, imageability and concreteness norms for 3,800 European Portuguese words. Poster presented at 18th Conference of the European Society for Cognitive Psychology (ESCOP), 29 August-01 September, Budapest, Hungary.

Soares, A. P., Iriarte, A., Almeida, J. J., Simões, A., Costa, A., França, P., Machado, J., & Comesaña, M. (2014a). Procura-PALavras (P-PAL): Uma nova medida de frequência lexical do Português Europeu contemporâneo [Procura-PALavras (P-PAL): A new measure of word frequency for contemporary European Portuguese]. Psicologia: Reflexão e Crítica, 27(1), 1-14.

Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, A., Almeida, J. J., Pinheiro, A. P., & Comesaña, M. (2014b). ESCOLEX: A grade-level lexical database from European Portuguese Elementary to Middle School textbooks. Behavior Research Methods, 46(1), 240-253.

Soares, A. P., Pinheiro, A. P., Costa, A., Frade, S., Comesaña, M., & Pureza, R. (2013a). Affective auditory stimuli: Adaptation of the International Affective Digitized Sounds (IADS-2) for European Portuguese. Behavior Research Methods, 45(4), 1168-1181.

Thorndike, E. L. (1921). The teacher's word book. New York: Teachers College, Columbia University.

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

Tiedemann, J. (2009). News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova &. Mitkov (Eds.), Recent advances in natural language processing (pp. 237-248). Amsterdam/Philadelphia: John Benjamins.

Universidade de Lisboa (1987). Português fundamental: Métodos e documentos. Lisboa: Instituto de Investigação Científica, 1987.

van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. The Quarterly Journal of Experimental Psychology DOI: 10.1080/17470218.2013.850521.

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

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

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

Ziegler, J. C., Petrova, A., & Ferrand, L. (2008). Feedback consistency effects in visual and auditory word recognition: where do we stand after more than a decade? Journal of Experimental Psychology: Learning Memory & Cognition, 34, 643-661.

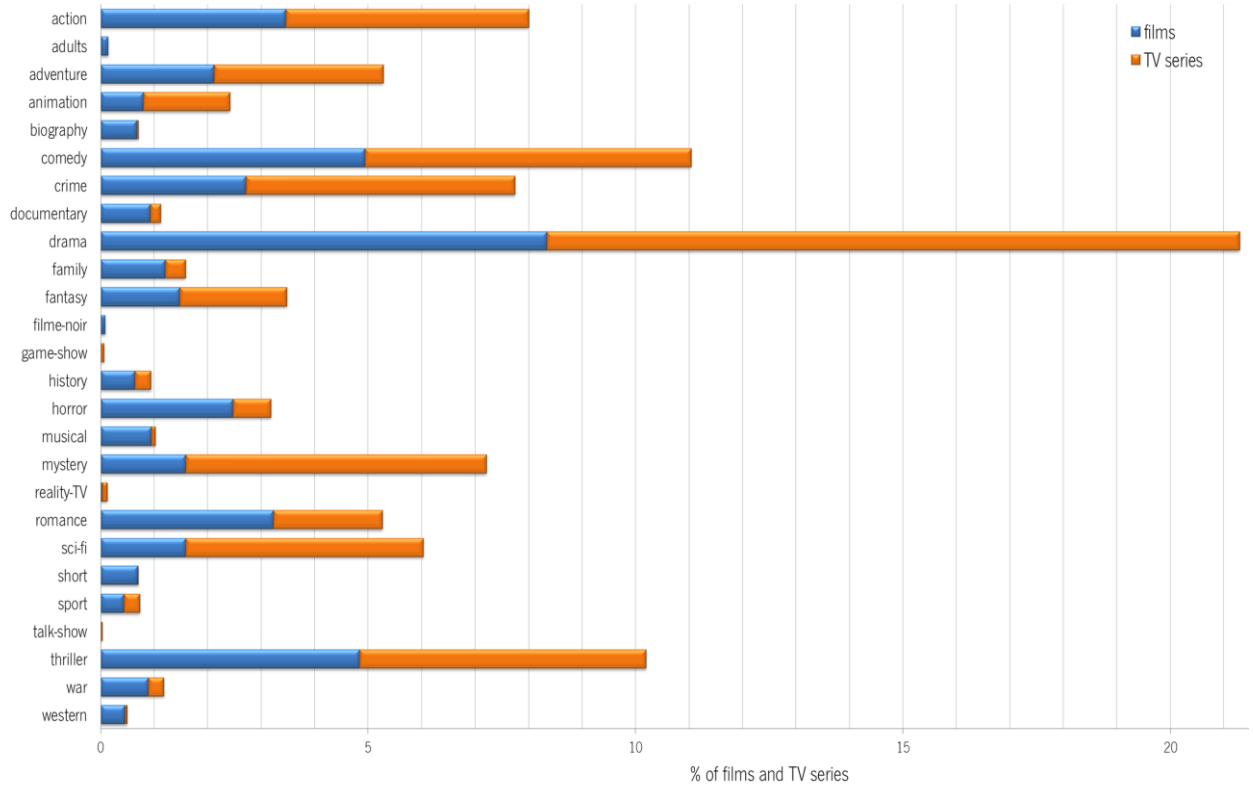

Figure 1. Film and television series subtitle distribution in the SUBTLEX-PT corpus by IMDb genre.

Table 1. Percentages of variance in Reaction Times (RT) and Accuracy (Acc) accounted by P-PAL and SUBTLEX-PT word frequency measures.

Frequency measures			RT (%)	Acc (%)
P-PAL$_{wf}$		LOG10	19.6	7.9
		LOG10+LOG10^2	21.2	10.0
SUBTLEX-PT	wf	LOG10	34.7	7.9
		LOG10+LOG10^2	36.1	9.3
	cd	LOG10	37.5	9.3
		LOG10+LOG102	37.9	10.2
	zipf	Zipf	34.7	7.9
		Zipf+Zipf^2	36.0	9.3

Note: LOG10 is the base 10 logarithm computed from FREQ$_{count}$+1 from P-PAL word frequency (wf) and from SUBTLEX-PT word-frequency (wf) and contextual diversity (cd) measures. Zipf measure is a logarithmic scale resulting from [log10 FREQcount+1/N] +3 – in which *N* corresponds to the number of words (types) in the corpus. LOG10^2 and Zipf^2 is the square value of each of these measures. All R^2s have *p*s < .001.