SUBTLEX-CAT: Subtitle word frequencies and contextual diversity for Catalan

Roger Boada¹ · Marc Guasch¹ · Juan Haro¹ · Josep Demestre¹ · Pilar Ferré¹

Published online: 20 March 2019
© The Psychonomic Society, Inc. 2019

Abstract

SUBTLEX-CAT is a word frequency and contextual diversity database for Catalan, obtained from a 278-million-word corpus based on subtitles supplied from broadcast Catalan television. Like all previous SUBTLEX corpora, it comprises subtitles from films and TV series. In addition, it includes a wider range of TV shows (e.g., news, documentaries, debates, and talk shows) than has been included in most previous databases. Frequency metrics were obtained for the whole corpus, on the one hand, and only for films and fiction TV series, on the other. Two lexical decision experiments revealed that the subtitle-based metrics outperformed the previously available frequency estimates, computed from either written texts or texts from the Internet. Furthermore, the metrics obtained from the whole corpus were better predictors than the ones obtained from films and fiction TV series alone. In both experiments, the best predictor of response times and accuracy was contextual diversity.

Keywords Word frequency · Contextual diversity · Catalan language · Subtitles

Word frequency influences most aspects of language processing. Together with word length, it is among the strongest predictors of lexical decision times (i.e., the time needed to decide whether a string of letters is a word or a nonword in a lexical decision task, or LDT), where high-frequency words are responded to faster than low-frequency words (e.g., Brysbaert et al., 2011). It is also important in word naming, although to a lesser extent (e.g., Balota, Cortese, Sergeant-Marshall, Spieler, & Yap, 2004). Moreover, word frequency is also relevant in memory research, in which high-frequency words are recalled better but recognized more poorly than low-frequency words (e.g., Cortese, Khanna, & Hacker, 2010). Consequently, this variable is routinely controlled for in psycholinguistic and memory studies. To do that, researchers rely on frequency databases. Such databases have traditionally been elaborated by counting the number of appearances of words within corpora obtained from written texts, including books, newspapers, and other periodical sources, such as magazines. Among these

corpora, most research in English has been based on the database of Kučera and Francis (1967), the frequency estimates coming from the British National Corpus (BNC; Kilgariff, 2006), and the CELEX Lexical Database (Baayen, Piepenbrock, & Gulikers, 1995). There are similar databases in other languages, including Spanish (Alameda & Cuetos, 1995; Sebastián-Gallés, Martí, Carreiras, & Cuetos, 2000), French (Imbs, 1971; New, Pallier, Brysbaert, & Ferrand, 2004), and Italian (De Mauro, Mancini, Vedovelli, & Voghera, 1993).

Corpora elaborated from written texts have some limitations that can affect the validity of the frequency estimates based on them. As some authors have pointed out, the topics included in those corpora are not necessarily related to daily life. Furthermore, they tap onto a linguistic style that is not representative of the language used commonly by speakers (Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010). This is because writers try to use a more formal register and to avoid repeating words. This yields greater lexical diversity, which can lead to an overestimation of the frequency of words that are hardly used by the speakers, as well as an underestimation of more frequently used words (Cuetos, González-Nosti, Barbón, & Brysbaert, 2011).

In an effort to overcome such limitations, some authors have turned to the Internet as a source of materials for

✉ Roger Boada
roger.boada@urv.cat

¹ Department of Psychology and Research Center for Behavior Assessment, Universitat Rovira i Virgili, Tarragona, Spain

elaborating frequency norms that are more suited to everyday language. For instance, Burgess and Livesay (1998) created the HAL corpus, which includes more than 130 million words, by gathering texts from online Usenet newsgroups. More recent Internet-based frequency norms have been based on even larger corpora, containing up to 500 billion words (Brants & Franz, 2006; Shaoul & Westbury, 2013).

During the last decade, some corpora have been elaborated from another type of digitized text—in particular, film and television subtitles. The first authors using such an approach were New, Brysbaert, Veronis, and Pallier (2007), who developed a database of French frequencies based on a corpus of 52 million words coming from 9,474 different films and television series. The authors compared these new measures to preexisting frequency measures obtained from written (New et al., 2004) and spoken (Reference Corpus of Spoken French; Equipe DELIC, 2004) corpora in relation to lexical decision times, either taken from a previous experiment (Bonin, Chalard, Méot, & Fayol, 2001) or collected by themselves. The authors observed that subtitle frequencies were better predictors of lexical decision times than were the more traditional frequency measures. Two years later, Brysbaert and New (2009) replicated the findings of New et al. (2007) with subtitle frequencies for American English (SUBTLEX-US). These authors used reaction times (RTs) obtained from a large pool of words (40,000), observing that frequency estimates derived from subtitles predicted lexical decision performance even better than did Internet-based frequencies.

Since those pioneering studies, frequency estimates based on film and television subtitles have been elaborated in several languages, such as Dutch (SUBTLEX-NL; Keuleers, Brysbaert, & New, 2010), Greek (SUBTLEX-GR; Dimitropoulou et al., 2010), Chinese (SUBTLEX-CH; Cai & Brysbaert, 2010), Spanish (SUBTLEX-ESP; Cuetos et al., 2011; EsPal's subtitle corpus collection: Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013), German (SUBTLEX-DE; Brysbaert et al., 2011), Brazilian Portuguese (SUBTLEX-PT-BR; Tang, 2012), Albanian (SUBTLEX-AL; Avdyli & Cuetos, 2013), British English (SUBTLEX-UK; van Heuven, Mandera, Keuleers, & Brysbaert, 2014), European Portuguese (SUBTLEX-PT; Soares et al., 2015), and Polish (SUBTLEX-PL; Mandera, Keuleers, Wodniecka, & Brysbaert, 2015). The authors of these studies have commonly obtained the raw material needed to elaborate the corpora from specialized websites. The number of subtitles considered has varied widely across databases, from the around 6,000 subtitle files in the SUBTLEX-GR corpus (Dimitropoulou et al., 2010) to the around 45,000 broadcasts in the SUBTLEX-UK corpus (van Heuven et al., 2014). Similarly, the final numbers of words (tokens) included in various corpora differ, ranging from two million words in

SUBTLEX-AL (Avdyli & Cuetos, 2013) to more than 450 million words in EsPal's subtitle corpus collection (Duchon et al., 2013).

A clear advantage of subtitle frequency norms is that they are derived from large corpora. Corpora containing large sets of words are preferable over small corpora for several reasons. One of these is that they provide better frequency estimates overall, because the standard error of the word counts diminishes as the sample size increases (Lee, 2003). The other is that larger corpora include a greater representation of low-frequency words than do smaller corpora, thus leading in particular to more accurate frequency estimates of such words. According to Brysbaert and New (2009), a corpus of 16–30 million words allows for reliable frequency estimates. Such a requirement is fulfilled by the majority of the above-mentioned databases. However, it seems that the size of the corpora from which subtitle frequency norms derive cannot entirely explain their advantage in predicting word recognition performance. Indeed, such an advantage has been demonstrated even when the subtitle frequency estimates have been compared with frequency estimates obtained from written sources involving large sets of words (e.g., Brysbaert et al., 2011). Instead, it seems that corpus representativeness (i.e., the extent to which the linguistic samples included in the corpus represent the language as a whole; Sinclair, 2005) can be more relevant. Film and television subtitles have a higher capacity to capture the language that people use in everyday life than do written corpora, because they are transcripts of oral communication. Furthermore, the majority of the studies on the effects of frequency on language processing have relied on young undergraduate students (Dimitropoulou et al., 2010), which is a population that spends more time watching TV than reading (Instituto Nacional de Estadística, 2011) and that mostly uses informal language. Thus, it becomes apparent that film and television subtitles can be a very appropriate source of word selection for the elaboration of frequency norms.

To test for the adequacy of such norms, all of the studies above validated their frequency estimates by comparing them to more traditional measures (i.e., frequencies derived from written corpora) in word recognition tasks. These studies have mainly relied on the LDT, because it is a task highly sensitive to word frequency (e.g., Balota et al., 2004; Brysbaert & New, 2009; Cortese & Khanna, 2007; Yap & Balota, 2009). To do this, distinct approaches have been adopted. In some cases, the authors have taken advantage of already published experiments and have correlated or regressed the RTs from those experiments with distinct frequency measures (e.g., Brysbaert & New, 2009; Cai & Brysbaert, 2010, Exp. 1; Cuetos et al., 2011; Duchon et al., 2013; New et al., 2007; van Heuven et al., 2014). In a few cases, those experiments

were megastudies involving large sets of words. As an example, van Heuven et al. performed correlations between the SUBTLEX-UK frequencies, the BNC frequencies (Kilgarriff, 2006), or the CELEX frequencies (Baayen et al., 1995) and the RTs and accuracy from the British Lexicon Project (BLP, which includes over 28,000 words; Keuleers, Lacey, Rastle, & Brysbaert, 2012). Similarly, Brysbaert and New (2009) used the lexical decision times from the more than 40,000 words included in the English Lexicon Project (Balota et al., 2007) to validate their American English frequency norms in comparison to frequencies obtained from distinct written sources: the Kučera and Francis norms (Kučera & Francis, 1967), CELEX (Baayen et al., 1995), the HAL corpus (Burgess & Livesay, 1998), the BNC (Leech, Rayson, & Wilson, 2001), and the Zeno corpus (Zeno, Ivens, Millard, & Duvvuri, 1995).

It should be noted, however, that word recognition megastudies are not available in all languages. In such cases, the preferred option by researchers has been to conduct an experiment involving a few hundred words as a validation of their frequency estimates in comparison to written-based frequencies (Avdyli & Cuetos, 2013; Brysbaert et al., 2011; Cai & Brysbaert, 2010, Exp. 2; Dimitropoulou et al., 2010; Mandera et al., 2015; Soares et al., 2015). As Brysbaert et al. pointed out, this is a good approach, provided that the experimental words are sampled from the entire frequency range. The experimental designs chosen have varied across studies. For instance, Brysbaert et al. did several lexical decision experiments manipulating the difficulty of the task by including distinct types of pseudowords. A different approach was followed by both Mandera et al. and Dimitropoulou et al., who selected the words for which the written-based corpora and the subtitle-based corpora gave the most divergent frequency estimates. Regardless of the approach taken (i.e., data coming from megastudies or from experiments involving a few hundred words), the conclusion of the reviewed studies has been unanimous: Frequency estimates derived from subtitles clearly outperform those derived from written texts in predicting word recognition performance.

Apart from demonstrating the superior predictive capacity of subtitle-derived measures in word recognition experiments, the above-reviewed studies have consistently revealed that one particular frequency measure is even more predictive: so-called *contextual diversity* (CD), defined by Adelman, Brown, and Quesada (2006) as the proportion of contexts in which a word appears in a lexical database. In an influential article, those authors demonstrated that CD predicted RTs in both lexical decision tasks and naming tasks better than the traditional word frequency measure (i.e., the number of times a word appears in a lexical database). Although these authors derived their measures from

written corpora, the results of the studies that have relied on subtitles, as mentioned above, have reached the same conclusion: CD (defined here as the number of subtitles in which a particular word appears) is a better predictor of word recognition RTs than is word frequency (Cai & Brysbaert, 2010; Dimitropoulou et al., 2010; Keuleers et al., 2012; van Heuven et al., 2014).

The facilitative effect of the number of contexts in which a word appears is a widespread finding. Indeed, RTs are faster for high-CD words (i.e., words appearing in many contexts) than for low-CD words in single-word identification tasks (e.g., Perea, Soares, & Comesaña, 2013; Vergara-Martínez, Comesaña, & Perea, 2017). Furthermore, fixation durations are shorter for the former than for the latter in sentence-reading tasks (Plummer, Perea, & Rayner, 2014). An advantage for high-CD words has also been observed in the acquisition of new vocabulary (Rosa, Tapia, & Perea, 2017; see also Huang, 2017, for a review). As for the nature of the CD effect, considering that word frequency and CD are highly correlated, a common mechanism for both effects might be expected. Alternatively, it might be that CD effects have an origin distinct from that of word frequency effects. This is what Vergara-Martínez et al. have recently demonstrated in an event-related potential study. Concretely, these authors showed that these two effects have distinct neural signatures. In particular, the direction of the N400 component associated with CD effects resembles that obtained with the manipulation of “semantic richness” variables (e.g., Rabovsky, Sommer, & Abdel Rahman, 2012), suggesting that encountering words in different contexts may increase their semantic richness (Vergara-Martínez et al., 2017). Considering all of the factors above, it is relevant to include both traditional word frequency measures and CD measures in databases of frequency estimates.

In this article, we introduce SUBTLEX-CAT, a database of subtitle frequency estimates for Catalan. Catalan is a Romance language spoken by approximately nine million people (Simons & Fennig, 2018). It is a co-official language in Catalonia, Valencia, and the Balearic Islands, together with Spanish. Hence, virtually all Catalan speakers are bilinguals of Catalan and Spanish. Due to the particularities of this population—they are highly proficient bilinguals who commonly acquire both languages during early childhood and live immersed in both languages—it has been the focus of interest of a large number of psycholinguistic studies dealing with topics including bilingual memory (e.g., Ferré, Sánchez-Casas, & Guasch, 2006; Guasch, Sánchez-Casas, Ferré, & García-Albea, 2008, 2011; Moldovan, Demestre, Ferré, & Sánchez-Casas, 2016); parallel activation of languages in bilinguals (e.g., Comesaña et al., 2015; Guasch, Ferré, & Haro, 2017); emotional processing in the two languages (e.g., Ferré,

Anglada-Tort, & Guasch, 2018; Ferré, García, Fraga, Sánchez-Casas, & Molero, 2010; Ferré, Sánchez-Casas, & Fraga, 2013); the linguistic, cognitive, and neural consequences of bilingualism (e.g., Branzi, Calabria, Boscarino, & Costa, 2016; Calabria, Branzi, Marne, Hernández, & Costa, 2015; Kandel et al., 2016; Martin et al., 2013; Rodríguez-Pujadas et al., 2013); and language deterioration in demented bilinguals (e.g., Calabria et al., 2017; Calabria, Marne, Romero-Pinel, Juncadella, & Costa, 2014), among others.

Considering the interest in Catalan–Spanish bilinguals, it is very relevant to have Catalan frequency norms that will allow researchers to perform monolingual experiments in Catalan as well as bilingual experiments in which both Catalan and Spanish are involved. To date, researchers have relied on the Corpus Textual Informatitzat de la Llengua Catalana (CTILC; Rafel, 1998) in order to get the frequency values for Catalan words. It is a corpus composed of 51,253,669 words, compiled from sources whose publication dates range from 1833 to 1988. As a corpus obtained from written texts, it has the limitations mentioned above. Furthermore, considering the year range covered, it is clear that the frequency estimates for some of the words might be outdated nowadays. In relation to that problem, it is worth mentioning the study of Brysbaert and New (2009), in which they observed that word frequency estimates based on subtitles previous to 1990 explained 3% less of the variance in lexical decision times for young participants than do post-1990 subtitles. Interestingly, the effect is reversed for older participants. This finding points out that, to get language-representative norms, it is highly desirable to derive frequency estimates from recent sources if the aim is to test young participants (which is the most common case in psycholinguistic studies).

Another corpus that has recently appeared is worth to be mentioned here. It is WorldLex (Gimenes & New, 2016), which compiles word frequency and contextual diversity norms for 66 languages (including Catalan), based on Internet sources. The Catalan corpus consists of 19.7 million words gathered in 2013 from blog posts (8.2 million), Twitter (6.5 million), and newspapers (5 million), 25 years later than the newest CTILC documents.

In this study, we developed SUBTLEX-CAT, which is based on the subtitles of television programs broadcast from 2011 to 2016 by four Catalan TV channels. In developing the database, we used a procedure similar to that followed in the subtitle databases elaborated in other languages. Our database, however, has some particularities that are worth mentioning here. First, with 278.6 million words, it is one of the largest available subtitle corpus in any language. Second, all of the subtitles were transcribed by professional translators or transcribers, whereas in most previous SUBTLEX databases the subtitles were collected

from websites constructed by the Internet community. Moreover, more than half of the subtitles correspond to broadcasts originally filmed in Catalan, whereas in the previous non-English subtitle corpora, the majority of their contents were adaptations from broadcasts (i.e., movies and TV series) originally filmed in English. Third, the variety of contents is significantly higher than in most SUBTLEX databases, which have exclusively included films and fiction TV shows (with only the exception, to our knowledge, of the UK version). In our corpus, this kind of content represents only about a quarter of the subtitles. The SUBTLEX-CAT corpus includes a wider range of contents (i.e., documentaries, news, talk shows, debates, sports broadcasts, etc.). We consider especially interesting that live shows were also subtitled and included in the corpus, thus offering a more natural approach to oral language than scripted shows. The last advantage of our corpus is the availability of metadata provided for most files, including production type, year, title, duration, and original language, among other information, which has allowed us to examine the corpus in more depth. Moreover, it has allowed us to construct several subcorpora by selecting the properties of the subtitles we wanted to include and to compare their performance as predictors on experimental psycholinguistic tasks.

Apart from the characteristics of SUBTLEX-CAT, some particularities of the speaking Catalan environment should be commented on. In particular, it should be noted that although Catalan is spoken in three Spanish autonomous communities (i.e., Catalonia, Valencia, and the Balearic Islands), there are several differences between the dialectal variants spoken in these regions. These differences are mainly lexical and phonological. Because the subtitles were obtained from Catalan broadcast television, the variety reflected in the subtitles is the normative Catalan spoken in the autonomous community of Catalonia. Thus, any researcher intending to use the SUBTLEX-CAT database should be aware of that. On the other hand, researchers using the present norms should also be aware that Spanish is present to a certain extent in the corpus. The reason is that, although the subtitles were written by professional translators, bilingual conversations took place in a small number of the broadcasts. This is a reflection of the pattern of speaking in Catalonia, where bilingual conversations are common. Hence, although the corpus did not include any entire show in Spanish, it was not possible to eliminate all of the Spanish from the corpus. In any case, the presence of non-Catalan words in our corpus was very low.

In the following sections, we first detail the procedures followed to develop the corpus and compute the frequency measures. After that, we present a validation of the database. To do this validation, we tested the predictive capacity of

Fig. 1 Histogram of the years of production of the shows whose subtitles were included in the corpus

frequency measures derived from SUBTLEX-CAT in comparison to those derived from other sources in two lexical decision experiments.

SUBTLEX-CAT

Corpus description

SUBTLEX-CAT was constructed from 84,141 subtitles in Catalan provided by the CCMA (the Catalan Audio-Visual Media Corporation), which include material from four broadcast TV channels (TV3, 3/24, Canal 33, and Esports 3). The shows corresponding to the subtitles provided were broadcast at least once between October 2011 and July 2016, although the shows could have been produced before this (see Fig. 1). In total, the subtitles in the corpus include 278.6 million word occurrences (tokens) of 751,078 different word forms (types) extracted from a total of 44,784 hours of television programming. The corpus includes 51.5% TV shows whose original language is Catalan, followed by 30.8% in English, 7.5% in Japanese, 6.4% in French, 1.4% in German, 1.2% in Spanish, and 1.3% in other languages.

Text cleaning

Previous to the final word count, a minimal cleaning of the subtitles was necessary. Since all the subtitles came from the same source, they had a very consistent structure. First, all the time codes and style tags were removed in order to leave only the raw text. Then, all the stage directions that were used to describe some aspect of the scene (e.g., to indicate who was speaking in a conversation in which the characters were off-screen) were also eliminated. Furthermore, in the subtitles provided, all song lyrics were subtitled and marked with a particular tag. TV shows

always begin with the same song. For that reason, we decided to remove all song lyrics, because otherwise the frequencies of the words in the lyrics would be overestimated. Additionally, all words were converted to lowercase before adding up the frequencies of the words, regardless of whether they appeared in the middle of a sentence or at its beginning. Finally, because punctuation marks were not going to be taken into account, they were deleted.

A special remark is worth making in reference to a particularity of the Catalan language. In Catalan, some pronouns (i.e., clitic pronouns) can be attached to either the front or the end of a verb with the character ' or -, (e.g., *dóna-m'ho* "give it to me," where both *m* and *ho* are pronouns). In addition, sometimes the article can also be attached at the front of a noun with the character ' (e.g., *l'home* "the man," where *l* refers to the article *el* "the"). In such cases, we decided to split these words into their constituents, because in fact such words can be considered compounds. In doing so, we ensured that the frequencies of the nouns, pronouns, and inflected verb forms were computed correctly.

After text cleaning, but before final processing, we extracted a list of all the characters still remaining in the subtitle files (see Appendix 1). The list includes all the valid letters in Catalan, plus all their diacritical marks and the ten digits. These data show that the presence of nonlegal characters in Catalan in the corpus is marginal (0.04%). Furthermore, 95.84% of the appearances of such characters correspond either to the letter "ñ," a nonlegal character in Catalan but present in several Spanish surnames and words, or to the character "à", which is a typical typo in Catalan, because the letter "a" can only carry the accent mark in the opposite direction (i.e., "â").

Text cleaning and the final word count were both carried out with a computer program devised for this purpose in the Python language.

Division of the corpus

We will henceforth refer to the corpus that includes all 84,141 subtitle files and 278.6 million words as *SUBTLEX-CAT_{ALL}*. Only the text-cleaning procedure described in the previous section was applied to the files.

Furthermore, thanks to the metadata provided with the subtitles, it was possible to build a second corpus, including only the subtitles from films and TV shows. We will refer to this corpus and the metrics derived from it as *SUBTLEX-CAT_{FILM}*. This corpus consisted of 15,097 subtitles that include 73.7 million words (26.5% of the original corpus), greatly exceeding both the sizes of many existing subtitle corpora in other languages and the requirement for a corpus of 16–30 million words for reliable word frequency norms (Brysbaert & New, 2009). Our goal was to build a corpus that was as similar as possible to those in which the subtitles were downloaded from the Internet. It is possible that fiction content, such as in films and TV series depicting everyday situations, more resembles the natural language to which speakers are exposed than do other kinds of content. Many nonfictional contents are very topic-specific (e.g., concerning savannah lions, modernist architecture, folk music, . . .), whereas others are highly related to people's everyday life (news) or are very topic-diverse (talk shows). Whether the inclusion of more diverse kinds of content improves or worsens the frequency measures is an empirical issue that will be addressed in the next sections.

Since we were provided with metadata for each subtitle file, we built other, additional subcorpora apart from *SUBTLEX-CAT_{ALL}* and *SUBTLEX-CAT_{FILM}*. For instance, we built a subcorpus eliminating around 20% of the subtitle files that were very short (less than 10 min) or that contained a great number of words in languages other than Catalan (mostly Spanish). We also computed frequency estimates for a subcorpus that included only broadcasts that were originally in Catalan, only live shows, excluding shows for children, and so forth. It is worth mentioning that none of these corpora outperformed the metrics derived from *SUBTLEX-CAT_{ALL}* in predicting lexical decision times and accuracy. For that reason, the analyses made with those subcorpora are not included here.

Obtaining the frequency measures

Following the previous SUBTLEX databases, we collected data for word frequency and contextual diversity. Word frequency (WF) has traditionally been expressed as the number of occurrences per million words in the corpus (relative frequency). Over the years, it has been increasingly common to use the logarithm of the relative frequency, plus one (Howes & Solomon, 1951), to mitigate the

exponential nature of word frequency expressed by Zipf's (1949) law. Following van Heuven et al. (2014), we used the Zipf measure for WF, computed as 3 plus $\log_{10}$ of the relative frequency. This measure not only improves the parametric properties of the WF values, because it is a logarithmic transformation, but it also offers a scale from 0 to 7 with a more natural and straightforward interpretation. Zipf values were computed for all the corpora considered in the following experiments.

Contextual diversity (CD) can be defined as the proportion of different contexts in which a word appears (Adelman et al., 2006). Following Brysbaert and New (2009), we considered each subtitle as a context and counted the number of contexts per million in which each word appeared. For the same reasons as for WF, the logarithm of the CD + 1 was used in all the analyses.

Apart from the frequency measures, we also included other indexes of interest in psycholinguistic studies, such as number of syllables, Coltheart's *N* (Coltheart, Davelaar, Jonasson, & Besner, 1977), number of higher-frequency neighbors, OLD20 (Yarkoni, Balota, & Yap, 2008), and mean bigram frequencies (by types and by tokens). Coltheart's *N* and OLD20 were computed using the *vwr* package in R (Keuleers, 2013). Coltheart's *N* refers to the number of words present in the corpus that can be obtained by changing just one letter from a given word (being all the words of the same length). The number of higher frequency neighbors, in turn, is the number of such words that have higher lexical frequency than that of the given word. OLD20 is the average Levenshtein distance between a word and its 20 nearest neighbors (including single-letter substitutions, but also taking into account letter insertions, deletions, and transpositions). Finally, to compute the mean bigram frequencies, first we obtained a list of all the bigrams existing in the corpus. For instance, the word *casa* ("house") contains three bigrams: *ca*, *as*, and *sa*. Then we counted the number of words in which each bigram appeared (type bigram frequency). We also computed the sum of the relative frequencies of all the words in which each bigram appeared (token bigram frequency). Thus, the mean bigram frequency of a word is the sum of the type or token bigram frequencies of its bigrams, divided by the number of bigrams that word has. It is worth mentioning that when computing these last two variables, and also the variables relating to neighborhood, we considered characters with diacritic marks as being different from their base characters (e.g., the *à* character was considered as different from the *a* character).

Supplementary materials

The norms are available as a Microsoft Excel .xlsx file, as supplementary materials to this article, from <https://psico.fcep.urv.cat/projectes/gip/papers/SUBTLEX-CAT.xlsx>, or

can be accessed online using the web tool NIM (Guasch, Boada, Ferré, & Sánchez-Casas, 2013; <https://psico.fcep.urv.cat/utilitats/nim/>). For correct computation of the variables related to lexical neighborhood and bigram frequencies, the raw database was trimmed. Concretely, the final set only contained those words included in one of the following online Catalan dictionaries or lists of words: the online dictionary of the Institut d'Estudis Catalans (1995, DIEC2; <http://mdlc.iec.cat/>), the online dictionary of the Enciclopèdia Catalana (<http://www.diccionari.cat/>), and the Catalan Softcatalà spellchecker (<https://www.softcatala.org/corrector/>¹). Additionally, we removed those words that contain either numbers or nonlegal characters in Catalan. Thus, the final supplementary materials include data from 202,204 Catalan words, along with the following information:

- **Num_letters**: the length of a word counted as number of letters.
- **Num_Syl**: the number of syllables of the word.
- **Abs_WF**: the absolute number of appearances of a word in the corpus.
- **Rel_WF**: the relative WF per million—that is, $(\text{Abs_WF} / 278,617,824) * 1,000,000$.
- **Zipf**: the value of a word in the Zipf scale computed as $\log(\text{Rel_WF}) + 3$.
- **Abs_CD**: the absolute number of subtitle files in which a word appears.
- **Rel_CD**: the relative CD per million—that is, $(\text{Abs_CD} / 84,141) * 1,000,000$.
- **log(Rel_CD+1)**: the log frequency of Rel_CD + 1.
- **N**: the number of words in the corpus that can be obtained by changing one letter from a word (i.e., Coltheart's *N*).
- **NHF**: the number of words in the Coltheart's *N* that have a higher lexical frequency than the given word.
- **OLD20**: mean Levenshtein distance of the 20 nearest neighbors of a word.
- **MBF_To**: mean token frequency of the bigrams of the word.
- **MBF_Ty**: mean type frequency of the bigrams of the word.

Validation

The LDT has proven to be especially sensitive to frequency measures and is commonly used to compare the proportions of variance explained by several lexical predictors. Due to the lack of previous published lexical decision studies in Catalan, we designed two different experiments in

order to test the predictive power of the different measures of WF and CD. We compared the SUBTLEX-CAT frequency estimates to the already available CTILC and WorldLex frequency norms.² In Experiment 1, we used a set of randomly selected words for each frequency level. In Experiment 2, we selected a set of stimuli that gave the most divergent frequency estimates between CTILC, SUBTLEX-CAT_{FILM}, and SUBTLEX-CAT_{ALL}. Of note, due to the characteristics of the speaking environment in Catalonia, all the participants were bilinguals of Catalan and Spanish. Although they were highly proficient in both languages, they considered themselves to be either Catalan- or Spanish-dominant. For that reason, we analyzed the results considering all the participants together as well as dividing them by dominance.

Experiment 1

Method

Participants A total of 47 Catalan college students (38 women and nine men) from the Universitat Rovira i Virgili (Tarragona, Spain) with ages between 18 and 30 ($M = 21.3$ years; $SD = 2.5$), took part in the experiment. All participants had normal or corrected-to-normal vision, and they participated in exchange for extra course credits. Participants were living in Catalonia at the moment of the test, had acquired the Catalan language before the age of 6 ($M = 2.86$, $SD = 1.48$), and rated themselves as being highly proficient on a 7-point Likert scale ($M = 6.72$, $SD = 0.42$). All the participants were bilinguals of Catalan and Spanish. We divided participants, according to their responses to a language history questionnaire, into two groups: 26 reported Catalan as their dominant language, and 21 considered Spanish their dominant language. Additionally, participants rated their proficiency, age of acquisition, and preference and frequency of use in listening, speaking, reading, and writing in both Catalan and Spanish. All the details about the characteristics of the Catalan-dominant and Spanish-dominant bilinguals may be found in Appendix 2.

Materials The stimuli consisted of 475 Catalan words selected from the set of words in common between CTILC and SUBTLEX-CAT. Because not many studies have used Catalan in a lexical decision task, we used a pseudo-

¹ A regularly updated version of the wordlist is available at <https://github.com/Softcatala/catalan-dict-tools/tree/master/resultats/lt>.

² The WorldLex frequency metrics have not been used, to date, in any published study involving the Catalan language. We were not aware of their existence when the experiments of the present study were designed. However, we have included WorldLex in the analyses in order to compare its predictive power with that of the other frequency measures. We thank one of the reviewers of this article for letting us know about these norms.

random method to select the final list of stimuli. We firstly removed all words longer than 15 letters or shorter than three letters. We divided that set into five different lists according to their Zipf values in the SUBTLEX-CAT corpus (between 5 and 6, between 4 and 5, between 3 and 4, between 2 and 3, and less than 2). The lists were then randomized, and the first 95 words of each list were selected. Misspelled words and proper nouns were removed and replaced by the next word on that list. The final 475-word lists had mean word lengths of 8.2 letters ($SD = 2.54$) and 3.2 syllables ($SD = 1.19$) (see Table 1).

Finally, a total of 475 nonwords were either selected from previous studies or constructed by replacing one or two letters from an existing Catalan word. All of them were pronounceable and legal letter sequences in Catalan and did not correspond to any word in the corpora.

Procedure The lexical decision experiment was run in the laboratory using the DMDX software (Forster & Forster, 2003). Participants were individually tested in separate soundproof booths. Each trial began with a fixation cross in the center of the screen for 500 ms, followed by the presentation of each stimulus for 2,000 ms or until a response was given. Participants were instructed to respond as quickly and accurately as possible by pressing a “yes” or “no” button on a keypad. They were told to select “yes” if they believed that the word existed in the Catalan language, and “no” if they believed that it did not. After participants had responded, the next fixation point appeared.

After a ten-trial practice block, the experimental trials were presented in a continuous running mode and stopped every 100 trials (approximately every 5 min) until the participant resumed the task. The order of item presentation was randomized for every participant. It took around 45 min to complete the task. After the experiment, participants filled out the language history questionnaire.

Results and discussion

Incorrect responses (approximately 12.9% of the data) and RTs shorter than 250 ms and longer than 1,500 ms (1% of the data) were excluded from the latency data analyses. Trials with response latencies three standard deviations above or below each participant’s average were also removed (1.4% of the data). Data from nine items with accuracy rates under 75% were also excluded. Therefore, all further analyses included 466 items.

We used separate linear regression analysis in which RTs or accuracy rates were the dependent variables and the different frequency measures were the predictors. All word frequency measures were transformed to the Zipf scale (van Heuven et al., 2014). The square of Zipf value for each word was included, too, as a predictor in the model to account for very high frequency effects, following Balota et al. (2004). We also constructed regression models using $\log_{10}(CD + 1)$ and its squared value instead of WF. For each regression analysis, we included the number of letters, the number of syllables, Coltheart’s N , NHF,

Table 1 Distribution of mean word length, WF, and CD in the randomly sampled words for each of the five tiers (standard deviations in parentheses)

Variable	WF 2 or less	WF 2–3	WF 3–4	WF 4–5	WF 5–6
Number of letters	10.18 (2.01)	9.46 (2.03)	8.27 (1.86)	7.29 (2.45)	5.84 (1.70)
Number of syllables	4.15 (0.86)	3.76 (1.04)	3.29 (1.10)	2.82 (1.03)	2.17 (0.81)
Coltheart’s N	0.71 (1.24)	1.01 (1.4)	2.28 (3.55)	4.88 (5.84)	8.03 (8.29)
NHF	0.41 (0.96)	0.39 (0.67)	0.45 (1.16)	0.48 (1.45)	0.40 (0.84)
OLD20	2.77 (0.7)	2.42 (0.59)	2.12 (0.56)	1.88 (0.73)	1.59 (0.58)
MBF Token	21,378 (7,531)	22,437 (7,762)	23,811 (9,414)	24,363 (9,655)	23,826 (9,554)
MBF Type	12,651 (3,811)	13,213 (3,883)	13,506 (5,051)	13,542 (5,105)	13,325 (5,341)
ZIPF _{CTILC}	2.23 (0.60)	2.75 (0.61)	3.52 (0.53)	4.14 (0.59)	5.11 (0.40)
ZIPF _{WorldLex-BLOG}	1.36 (1.19)	2.54 (0.82)	3.53 (0.44)	4.3 (0.59)	5.14 (0.65)
ZIPF _{WorldLex-TWITTER}	0.68 (1.1)	1.64 (1.33)	3.05 (1.16)	4.13 (0.7)	5.01 (0.73)
ZIPF _{WorldLex-NEWS}	0.61 (1.14)	1.37 (1.54)	2.8 (1.46)	3.98 (0.95)	4.95 (0.74)
ZIPF _{FILM}	0.84 (0.81)	2.05 (0.71)	3.13 (0.46)	4.07 (0.48)	5.25 (0.40)
ZIPF _{ALL}	1.49 (0.36)	2.50 (0.25)	3.47 (0.27)	4.34 (0.26)	5.30 (0.25)
CD _{WorldLex-BLOG}	0.59 (0.56)	1.28 (0.52)	2.15 (0.43)	2.93 (0.48)	3.76 (0.54)
CD _{WorldLex-TWITTER}	0.21 (0.36)	0.59 (0.53)	1.45 (0.74)	2.36 (0.59)	3.23 (0.61)
CD _{WorldLex-NEWS}	0.33 (0.65)	0.81 (0.96)	1.8 (1.05)	2.78 (0.79)	3.72 (0.66)
CD _{FILM}	1.19 (1.12)	2.65 (0.82)	3.72 (0.47)	4.60 (0.46)	5.59 (0.29)
CD _{ALL}	1.97 (0.35)	2.95 (0.24)	3.86 (0.28)	4.64 (0.26)	5.46 (0.20)

OLD20, MBF token, and MBF type as control variables. We carried out the analyses considering all the participants together, as well as separating them by dominance. We present adjusted $R^2 * 100$ as a measure of the percentage of explained variance in the following analyses.

Considering all participants together, the analyses revealed that both the WF and CD measures from the two SUBTLEX-CAT corpora improved the percentage of explained variance when compared with the CTILC frequency metrics. The SUBTLEX-CAT_{FILM} word frequency explained 2.7% and 5% more variability in the latency and accuracy analyses, respectively. This result shows that frequency estimates from film and fiction TV series subtitles are reliable predictors of lexical access, in line with the findings from other SUBTLEX databases. Interestingly, SUBTLEX-CAT_{ALL} benefited either from the larger corpus size or the variety of contents, outperforming SUBTLEX-CAT_{FILM} by an additional 3%–4% for both RTs and accuracy. As in previous studies, CD measures were better predictors of response latencies and accuracy than were their WF counterparts, but the increase was lower here, not reaching 1% in any case (see Table 2). The three measures of WF and CD obtained from WorldLex (from blogs, Twitter, and newspapers) were also added to the control variables as predictors in the corresponding analysis. The metrics from WorldLex were effective predictors of RTs, improving on those from the CTILC and SUBTLEX-CAT_{FILM}, but explaining 2% less variance than SUBTLEX-CAT_{ALL} for both WF and CD. Nonetheless, WorldLex predicted an additional 0.6% of error rate variance as compared to SUBTLEX-CAT. All analyses comparing the two dominance groups revealed the same pattern: a higher proportion of variance explained in the Catalan-dominant than in the Spanish-dominant group for both the RT (ranging between 2.4% and 4.7%) and error (between 0.7% and 1.7%) analyses.

The results of Experiment 1 clearly show that the subtitle metrics explained a greater variability of the dependent variables than did the frequency measures based on

written texts (i.e., the CTILC norms). The best tested model included the control variables and CD_{ALL}, explaining 63.2% of the RT variance and 33.1% of the accuracy variance. Those values represent a 7.3% and 8.8% increase, respectively, as compared to the CTILC frequency estimates. CD_{ALL} also explained 2% more variance in RTs than did CD_{WORLDEX}. Notwithstanding, Worldlex CD was a better predictor of accuracy than was SUBTLEX-CAT, by 0.6%. These results are along the same lines as those described in the previous literature (e.g., Avdyli & Cuetos, 2013; Brysbaert et al., 2011; Cai & Brysbaert, 2010; Cuetos et al., 2011; Dimitropoulou et al., 2010; Duchon et al., 2013; Keuleers et al., 2010; Mander et al., 2015; Soares et al., 2015; Tang, 2012; van Heuven et al., 2014).

Experiment 2

In Experiment 1, we found that any of the SUBTLEX-CAT measures was a better predictor of lexical decision performance than were the previously available norms (CTILC and WorldLex). In Experiment 2, we aimed to explore the differences between the available frequency measures by selecting those words giving the most divergent frequency estimates across the SUBTLEX-CAT and CTILC corpora—among WF_{CTILC}, WF_{FILM}, and WF_{ALL}, on the one hand, and CD_{FILM} and CD_{ALL}, on the other. In this way, we tried to replicate the results of Experiment 1 with a more constrained item sample.

Method

Participants A total of 55 Catalan college students (45 women and 10 men) from the Universitat Rovira i Virgili (Tarragona, Spain), with ages between 18 and 28 ($M = 20.38$ years, $SD = 2.93$), took part in the experiment. All participants had normal or corrected-to-normal vision, and they participated in exchange for extra course credit. The

Table 2 Percentages of explained variance in response times and accuracy by each group of predictors

Predictors	All participants		Catalan-dominant bilinguals		Spanish-dominant bilinguals	
	RT	Accuracy	RT	Accuracy	RT	Accuracy
WF _{CTILC}	55.9	24.3	55.3	23.3	50.7	22.6
WF _{WORLDEX}	60.7	32.4	58.7	31.1	56.3	30.1
WF _{FILM}	58.6	29.3	57.3	28.4	53.6	27.0
WF _{ALL}	62.8	32.4	61.7	31.4	57.2	29.8
CD _{WORLDEX}	61.2	33.7	59.2	32.3	56.6	31.2
CD _{FILM}	59.1	29.3	57.9	28.4	54.1	27.1
CD _{ALL}	63.2	33.1	62.2	32.1	57.5	30.4

participants were living in Catalonia at the moment of the test and had acquired the Catalan language before the age of 6 ($M = 2.88$, $SD = 1.33$). Participants filled out the same language history questionnaire as in Experiment 1. Their average proficiency in Catalan was very high on a 1-to-7 scale ($M = 6.6$, $SD = 0.58$). There were 36 Catalan-dominant bilinguals, and the remaining 19 participants were Spanish-dominant bilinguals (see Appendix 2).

Materials We began the selection of the critical words with a list of all the words that appeared at least once in the CTILC and each of the two SUBTLEX-CAT corpora, and removed all the words shorter than three or longer than ten letters. To select those words that showed the highest differences across the corpora, we carried out several regression analyses among the three measures of WF, on the one hand, and among the two CD measures, on the other, and selected those words that showed the highest residuals (i.e., the ones that differed most between measures). We also excluded inflected verbal forms, proper nouns, propositions, conjunctions, diminutives, numbers, place names, demonyms, and compound words. The final set of 475 items contained those words that maximized differences across the corpora regarding the metrics of lexical frequency and CD. Table 3 summarizes the descriptive statistics for the variables included in the analyses. Examples of the ten words with higher and lower residuals in each

regression analysis are presented in Appendix 3. Finally, the same 475 pseudowords as in Experiment 1 were used as “no” responses in the lexical decision task.

Procedure We used the same procedure as in Experiment 1.

Results and discussion

Incorrect responses (approximately 8.15% of the data) and RTs shorter than 250 ms and longer than 1,500 ms (1% of the data) were excluded from the latency data analyses. Trials with response latencies three standard deviations above or below each participant’s average were also removed (1.5% of the data). The data from one item with an accuracy rate under 75% were also removed, so all subsequent analyses included 474 items. As in Experiment 1, linear regression models were constructed in order to predict latencies and accuracy from the same control variables as in Experiment 1 and from the distinct WF and CD measures considered in the present study.

Table 4 shows the percentages of variance in the latency and error data explained by the predictors of each model. Contrary to what was observed in Experiment 1, WF_{FILM} and CD_{FILM} did not perform better than WF_{CTILC} . $WF_{WORLDLEX}$ explained a roughly 2% more variance than the CTILC norms for both RTs and accuracy, whereas this difference increased to 3% when $CD_{WORLDLEX}$ was used. More importantly, WF_{ALL} explained an extra 4% of variance in RTs and 1.5% in accuracy, when compared with CTILC. In line with previous results, CD_{ALL} was the best RT predictor, surpassing WF_{CTILC} by 5% and $CD_{WORLDLEX}$ by 2%. Accuracy was predicted slightly better by WorldLex than by SUBTLEX (by 0.2% for CD). The pattern of RT results was similar to that from Experiment 1 when comparing Catalan- and Spanish-dominant bilinguals. Frequency norms were better predictors for the former than for the later for both RTs (differences ranging between 6.4% and 8.6%) and accuracy (between 1.4% and 3.4%). The only noticeable difference between the two groups in Experiment 2 was that in the accuracy analysis, the WorldLex metrics outperformed SUBTLEX-CAT in the Spanish-dominant group, whereas the advantage was reversed in the Catalan-dominant group.

The comparison of the results of Experiments 1 and 2 reveals that all available frequency norms predicted RTs better in Experiment 2 than in Experiment 1. A possible cause of such an advantage could be the more constrained item sample in the former experiment, which resulted in a list of words known by more participants, as can be observed in the overall error rates (13% vs. 8%) and the numbers of items with more than 75% errors (nine vs. one). However, the same was not true for accuracy, in that frequency estimates were better predictors in Experiment

Table 3 Descriptive statistics for the variables considered in Experiment 2, including the maximum and minimum values, mean, and standard deviation

Variables	Minimum	Maximum	Mean	SD
Number of letters	3.00	10.00	6.54	2.00
Number of syllables	1.00	6.00	2.73	1.06
N	0	42	5.69	8.10
N_{HF}	0	14	0.50	1.47
OLD20	1	3.80	1.83	0.60
MBF_Token	2,667	65,873	20,784	8,896
MBF_Type	2,179	40,673	11,938	4,928
WF_{CTILC}	1.29	6.28	4.08	1.18
$WF_{WORLDLEX-Blog}$	0	6.29	4.09	1.25
$WF_{WORLDLEX-Twitter}$	0	6.58	3.69	1.71
$WF_{WORLDLEX-News}$	0	6.17	3.53	1.88
WF_{ALL}	1.99	6.30	4.26	1.16
WF_{FILM}	1.43	6.52	4.10	1.38
$CD_{WORLDLEX-Blog}$	0	4.82	2.72	1.18
$CD_{WORLDLEX-Twitter}$	0	4.78	2.11	1.32
$CD_{WORLDLEX-News}$	0	4.93	2.50	1.50
CD_{ALL}	2.44	5.92	4.48	1.04
CD_{FILM}	1.83	6.00	4.51	1.21

Table 4 Percentages of explained variance of response times and accuracy by each group of predictors

Predictors	All participants		Catalan-dominant bilinguals		Spanish-dominant bilinguals	
	RT	Accuracy	RT	Accuracy	RT	Accuracy
WF _{CTILC}	63.0	30.1	63.3	28.2	54.7	26.2
WF _{WORLDLEX}	65.1	32.1	64.9	29.7	57.9	28.3
WF _{FILM}	60.9	27.0	60.7	25.9	53.6	22.7
WF _{ALL}	66.9	31.6	66.2	30.1	59.7	27.1
CD _{WORLDLEX}	66.0	32.9	65.9	30.7	58.6	28.9
CD _{FILM}	63.1	28.7	62.9	27.5	55.6	24.1
CD _{ALL}	68.1	32.7	67.3	31.1	60.9	28.1

1 than in Experiment 2. Despite this difference, both experiments show very similar patterns of results: The best predictors of lexical decision RTs were those metrics obtained from the whole subtitle corpus, followed by WorldLex, the film subtitle corpus, and the CTILC metrics. When the two dominance groups were analyzed separately, the same pattern was observed for the Spanish-dominant groups in both experiments. Nevertheless, the analysis of accuracy in the Catalan-dominant group revealed some differences between experiments: whereas in Experiment 1 CD_{WORLDLEX} was the best predictor, in Experiment 2 SUBTLEX-CAT explained more error variance. What is more interesting is that in both experiments performance was predicted better by the frequency norms in Catalan-dominant bilinguals than in Spanish-dominant bilinguals.

Conclusion

With SUBTLEX-CAT, we present word frequency and contextual diversity measures for the Catalan language based on one of the largest subtitle corpora in any language to the present date. We replicated two of the common findings with subtitle-based word frequency metrics: first, the superiority of these norms over the previously available norms, and second, the advantage of contextual diversity measures when compared with superficial word frequency estimates. As to the first point, the superiority of SUBTLEX-CAT over norms derived from written texts (CTILC) was observed in two separate lexical decision experiments. In addition to their different origins, another distinctive aspect of the two corpora to highlight is the temporal difference between them, as CTILC is about 20 years older than SUBTLEX-CAT. This fact can be qualitatively corroborated by the presence of relatively frequent words in the present corpus that are not present in the previous one (e.g., *videojoc*, video game; *càsting*, casting;

alienígena, alien; *zombi*, zombie . . .). Along the same lines, when comparing the subtitle-based frequency measures to those coming from other online sources, such as Twitter, blogs, and newspapers, SUBTLEX-CAT was a better predictor of lexical access than the WorldLex estimates. Both corpora include texts that are supposed to better reflect the everyday use of language than do corpora based only on written texts (i.e., CTILC). Both SUBTLEX and WorldLex are also based on newer materials than those in CTILC. However, it remains unclear whether the advantage of subtitle-based norms simply arises from the bigger corpus size (278 million vs. 20 million words).

Furthermore, we empirically tested the benefits of including nonfictional contexts in a subtitle corpus. We observed that metrics derived from a corpus that includes exclusively subtitles from films and fiction TV series are underpowered as compared with a larger and more varied corpus that includes other types of nonfiction broadcasts. Although the observed improvement is modest and might be produced simply by the difference in size between the two corpora, it is more likely that the cause is the greater representativeness of the larger corpus. In fact, New et al. (2007) already pointed out that corpora that include only fictional TV materials are biased, and therefore the sampled materials should be made more representative of the natural language. This recommendation has already been followed by van Heuven et al. (2014) for British English, using subtitles from the BBC. To our knowledge, SUBTLEX-CAT is the second subtitle-based frequency metric from this “second generation” to include broadcasts from all kind of fictional and nonfictional TV broadcasts. We hope that this new database will facilitate psycholinguistic research in the Catalan language.

Author note This research was funded by grant PCIN-2015-165-C02-02S from MINECO/FEDER and by the Research Promotion Program of the Universitat Rovira i Virgili (2017PFR-URV-B2-32). We thank the Catalan Audio-Visual Media Corporation for kindly providing us with the subtitles.

Appendix 1

Table 5 List of all characters present in the subtitles after cleaning, with their absolute frequencies and percentages of appearance

Character	Absolute frequency	Percentage	Character	Absolute frequency	Percentage
<i>space</i>	239,608,098	17.47%	q	18,672,938	1.36%
a	140,845,168	10.27%	r	76,004,136	5.54%
à	4,566,528	0.33%	s	88,619,052	6.46%
b	15,345,505	1.12%	t	67,445,329	4.92%
c	36,339,480	2.65%	u	51,683,224	3.77%
ç	1,055,690	0.08%	ú	1,102,388	0.08%
d	37,592,104	2.74%	ü	172,629	0.01%
e	138,209,800	10.08%	v	17,366,996	1.27%
é	10,143,654	0.74%	w	494,445	0.04%
è	4,152,176	0.30%	x	6,029,826	0.44%
f	11,049,819	0.81%	y	3,890,672	0.28%
g	15,460,022	1.13%	z	986,164	0.07%
h	15,766,501	1.15%	·*	478,989	0.03%
i	71,809,338	5.24%	0	1,125,195	0.08%
í	4,950,383	0.36%	1	637,601	0.05%
ï	309,691	0.02%	2	443,721	0.03%
j	5,133,093	0.37%	3	395,398	0.03%
k	1,037,966	0.08%	4	276,105	0.02%
l	64,607,323	4.71%	5	364,989	0.03%
m	39,656,623	2.89%	6	191,821	0.01%
n	72,931,825	5.32%	7	171,889	0.01%
o	63,056,644	4.60%	8	178,688	0.01%
ó	4,045,173	0.29%	9	192,060	0.01%
ò	3,888,437	0.28%	Other**	542,041	0.04%
p	32,258,066	2.35%	Total	1,371,285,403	100 %

* This character is the middle dot used in Catalan for the geminate l (i.e., “l·l”). ** Other refers to the following non-Catalan characters: & á â ã ä å æ ê ë ì î ï ò ó ô õ ø ù û ý ÿ

Appendix 2

Table 6 Results from the language history questionnaire for all participants in Experiments 1 and 2

	Experiment 1			Experiment 2		
	All participants N = 47	L1 Catalan N = 26	L1 Spanish N = 21	All participants N = 55	L1 Catalan N = 36	L1 Spanish N = 19
RT	670.43 (111.3)	661.99 (115.5)	679.43 (114.7)	661.90 (95.1)	656.89 (93.0)	671.75 (106.6)
Accuracy	12.87 (19.8)	12.53 (19.8)	13.27 (20.7)	8.97 (12.6)	8.88 (12.3)	9.14 (15.0)
Age	21.30 (2.5)	21.15 (2.7)	21.48 (2.2)	20.38 (2.9)	20.28 (3.2)	20.58 (2.4)
Catalan proficiency	6.72 (0.4)	6.85 (0.3)	6.57 (0.5)	6.60 (0.6)	6.74 (0.6)	6.60 (0.6)
Listening	6.98 (0.1)	7.00 (0.0)	6.95 (0.2)	6.91 (0.3)	6.97 (0.2)	6.91 (0.4)
Speaking	6.53 (0.7)	6.73 (0.6)	6.29 (0.8)	6.55 (0.7)	6.86 (0.4)	6.55 (0.8)
Reading	6.85 (0.4)	6.92 (0.3)	6.76 (0.4)	6.65 (0.9)	6.72 (1.0)	6.65 (0.6)
Writing	6.53 (0.8)	6.73 (0.5)	6.29 (1.0)	6.31 (1.1)	6.42 (1.2)	6.31 (0.9)
Spanish proficiency	6.86 (0.3)	6.76 (0.4)	6.99 (0.1)	6.66 (0.7)	6.53 (0.8)	6.66 (0.2)
Listening	6.98 (0.1)	6.96 (0.2)	7.00 (0.0)	6.85 (0.5)	6.81 (0.6)	6.85 (0.2)
Speaking	6.72 (0.6)	6.50 (0.8)	7.00 (0.0)	6.58 (0.8)	6.39 (0.9)	6.58 (0.2)
Reading	6.94 (0.4)	6.88 (0.6)	7.00 (0.0)	6.71 (1.0)	6.58 (1.2)	6.71 (0.2)
Writing	6.81 (0.5)	6.69 (0.6)	6.95 (0.2)	6.49 (1.0)	6.36 (1.2)	6.49 (0.5)
Age of acquisition: Catalan	2.86 (1.5)	2.50 (1.4)	3.30 (1.3)	2.88 (1.3)	2.30 (1.1)	2.88 (1.1)
Speaking	1.26 (1.6)	0.42 (1.0)	2.29 (1.6)	1.09 (1.6)	0.14 (0.4)	1.09 (1.5)
Reading	3.47 (1.5)	3.35 (1.5)	3.62 (1.4)	3.62 (1.5)	3.22 (1.6)	3.62 (1.2)
Writing	3.85 (1.4)	3.73 (1.6)	4.00 (1.0)	3.93 (1.7)	3.53 (1.7)	3.93 (1.4)
Age of acquisition: Spanish	2.85 (1.4)	2.99 (1.6)	2.68 (1.0)	3.02 (1.4)	3.15 (1.6)	3.02 (1.1)
Speaking	0.98 (1.4)	1.31 (1.7)	0.57 (0.9)	1.04 (1.7)	1.33 (1.9)	1.04 (1.0)
Reading	3.66 (1.4)	3.73 (1.6)	3.57 (1.2)	3.78 (1.7)	3.89 (1.8)	3.78 (1.5)
Writing	3.91 (1.4)	3.92 (1.6)	3.90 (1.0)	4.25 (1.7)	4.22 (1.8)	4.25 (1.5)
Frequency of use	4.13 (1.1)	3.63 (1.1)	4.75 (0.8)	4.07 (1.3)	3.49 (1.2)	4.07 (0.8)
Listening	4.02 (1.0)	3.69 (0.9)	4.43 (0.9)	3.93 (1.3)	3.47 (1.2)	3.93 (0.8)
Speaking	4.02 (1.6)	3.12 (1.4)	5.14 (1.1)	4.04 (1.7)	3.22 (1.5)	4.04 (0.8)
Reading	4.53 (1.3)	4.31 (1.3)	4.81 (1.2)	4.47 (1.4)	4.11 (1.4)	4.47 (1.2)
Writing	3.96 (1.4)	3.42 (1.3)	4.62 (1.2)	3.85 (1.7)	3.17 (1.4)	3.85 (1.4)
Preference of use	4.23 (1.2)	3.71 (1.3)	4.88 (1.3)	4.18 (1.4)	3.65 (1.2)	4.18 (1.1)
Listening	4.06 (1.0)	3.81 (1.0)	4.38 (0.9)	3.93 (1.3)	3.47 (1.2)	3.93 (1.1)
Speaking	4.02 (1.9)	3.04 (1.6)	5.24 (1.4)	4.00 (1.9)	3.08 (1.5)	4.00 (1.0)
Reading	4.53 (1.2)	4.15 (1.0)	5.00 (1.3)	4.42 (1.5)	4.19 (1.5)	4.42 (1.4)
Writing	4.32 (1.5)	3.85 (1.4)	4.90 (1.4)	4.38 (1.6)	3.86 (1.5)	4.38 (1.3)

We include the values for the whole group as well as for each one of the two dominance groups. Proficiency was assessed in a 7-point Likert Scale. Age of acquisition is in years. Frequency and preference of use were presented in a 7-point scale, on which 1 means *only Catalan* and 7 *only Spanish*

Appendix 3

Table 7 Examples of ten words with higher and lower residuals in each regression analysis

$WF_{CTILC} < WF_{FILM}$		$WF_{CTILC} < WF_{ALL}$		$WF_{FILM} < WF_{ALL}$		$CD_{FILM} < CD_{ALL}$	
<i>comissària</i>	(commissioned)	<i>megafonia</i>	(megaphone)	<i>rovelló</i>	(milk mushroom)	<i>geminada</i>	(geminated)
<i>megafonia</i>	(megaphone)	<i>lletrejar</i>	(spell)	<i>lèmur</i>	(lemur)	<i>investir</i>	(inaugurate)
<i>guanterà</i>	(glove box)	<i>comissària</i>	(commissioned)	<i>geminada</i>	(geminated)	<i>autonòmic</i>	(autonomic)
<i>condó</i>	(condom)	<i>córner</i>	(corner)	<i>autonòmic</i>	(autonomic)	<i>unitària</i>	(unitary)
<i>motel</i>	(motel)	<i>cancellera</i>	(chancellor)	<i>lèxic</i>	(lexic)	<i>revalidar</i>	(revalidate)
<i>dislèxia</i>	(dyslexia)	<i>desempatar</i>	(tear off)	<i>dictat</i>	(dictation)	<i>doblet</i>	(doubled)
<i>lesbiana</i>	(lesbian)	<i>galàctic</i>	(galactic)	<i>unitària</i>	(unitary)	<i>hominids</i>	(hominids)
<i>bugaderia</i>	(laundry room)	<i>marató</i>	(marathon)	<i>doblet</i>	(doubled)	<i>rovelló</i>	(mushroom)
<i>emprenyar</i>	(embroidering)	<i>lèmur</i>	(lemur)	<i>secà</i>	(dry land)	<i>secà</i>	(dry land)
<i>sòcia</i>	(partner)	<i>robot</i>	(robot)	<i>córner</i>	(corner)	<i>cancellera</i>	(chancellor)
$WF_{FILM} < WF_{CTILC}$		$WF_{ALL} < WF_{CTILC}$		$WF_{ALL} < WF_{FILM}$		$CD_{ALL} < CD_{FILM}$	
<i>hegemonia</i>	(hegemony)	<i>ponderació</i>	(weighting)	<i>epilèptic</i>	(epileptic)	<i>epilèptic</i>	(epileptic)
<i>catalana</i>	(Catalan)	<i>epígraf</i>	(epigraph)	<i>intravenós</i>	(intravenous)	<i>sincerar</i>	(to sincere)
<i>secà</i>	(dry land)	<i>infinitiu</i>	(infinitive)	<i>altercat</i>	(scuffle)	<i>intravenós</i>	(intravenous)
<i>pastoral</i>	(pastoral)	<i>fingida</i>	(simulated)	<i>sincerar</i>	(to sincere)	<i>altercat</i>	(scuffle)
<i>grafia</i>	(spelling)	<i>restitució</i>	(restitution)	<i>procurador</i>	(attorney)	<i>radiologia</i>	(radiology)
<i>allí</i>	(there)	<i>grafia</i>	(spelling)	<i>comissària</i>	(commissioned)	<i>encobridor</i>	(abettor)
<i>lèxic</i>	(lexicon)	<i>tremolosa</i>	(trembling)	<i>promiscua</i>	(promiscuous)	<i>guanterà</i>	(glove box)
<i>infinitiu</i>	(infinitive)	<i>arbitrari</i>	(arbitrary)	<i>lletrat</i>	(lawyer)	<i>condó</i>	(condom)
<i>pertanyent</i>	(pertaining)	<i>allí</i>	(there)	<i>albarà</i>	(delivery note)	<i>incriminar</i>	(incriminating)
<i>nombre</i>	(number)	<i>cerimoniós</i>	(ceremonial)	<i>radiologia</i>	(radiology)	<i>albarà</i>	(delivery note)

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. <https://doi.org/10.1111/j.1467-9280.2006.01787.x>
- Alameda, J. R., & Cuetos, F. (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Oviedo, Spain: Servicio de Publicaciones de la Universidad de Oviedo.
- Avdyl, S. R., & Cuetos, S. F. (2013). SUBTLEX-AL: Albanian word frequencies based on film subtitles. *ILIRIA International Review*, 3, 285–292. <https://doi.org/10.3389/fpsyg.2010.00218>
- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (CD ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- 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. <https://doi.org/10.1037/0096-3445.133.2.283>
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459. <https://doi.org/10.3758/BF03193014>
- 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, 401–443.
- Brants, T., & Franz, A. (2006). *Web 1T 5-gram, version 1*. Philadelphia, PA: Linguistic Data Consortium.
- Branzi, F. M., Calabria, M., Boscarino, M. L., & Costa, A. (2016). On the overlap between bilingual language control and domain-general executive control. *Acta Psychologica*, 166, 21–30. <https://doi.org/10.1016/j.actpsy.2016.03.001>
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58, 412–424. <https://doi.org/10.1027/1618-3169/a000123>
- 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. <https://doi.org/10.3758/BRM.41.4.977>
- Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, 30, 272–277. <https://doi.org/10.3758/BF03200655>
- Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, 5, e10729. <https://doi.org/10.1371/journal.pone.0010729>
- Calabria, M., Branzi, F. M., Marne, P., Hernández, M., & Costa, A. (2015). Age-related effects over bilingual language control and executive control. *Bilingualism: Language and Cognition*, 18, 65–78. <https://doi.org/10.1017/S1366728913000138>
- Calabria, M., Cattaneo, G., Marne, P., Hernández, M., Juncadella, M., Gascón-Bayarri, J., . . . Costa, A. (2017). Language deterioration in bilingual Alzheimer's disease patients: A longitudinal study. *Journal of Neurolinguistics*, 43, 59–74. <https://doi.org/10.1016/j.jneuroling.2016.06.005>
- Calabria, M., Marne, P., Romero-Pinel, L., Juncadella, M., & Costa, A. (2014). Losing control of your languages: A case study. *Cognitive Neuropsychology*, 31, 266–286. <https://doi.org/10.1080/02643294.2013.879443>
- Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.
- Comesaña, M., Ferré, P., Romero, J., Guasch, M., Soares, A. P., & García-Chico, T. (2015). Facilitative effect of cognate words vanishes when reducing the orthographic overlap: The effect of stimuli list composition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 614–635. <https://doi.org/10.1037/xlm0000065>
- Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, 60, 1072–1082. <https://doi.org/10.1080/17470210701315467>
- Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory*, 18, 595–609. <https://doi.org/10.1080/09658211.2010.493892>
- Cuetos, F., González-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32, 133–143.
- De Mauro, T., Mancini, F., Vedovelli, M., & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato (LIP)*. Milan, Italy: Etaslibri.
- Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in Psychology*, 1, 218. <https://doi.org/10.3389/fpsyg.2010.00218>
- 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. <https://doi.org/10.3758/s13428-013-0326-1>
- Equipe DELIC. (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le Français Parlé*, 18, 11–42.
- Ferré, P., Anglada-Tort, M., & Guasch, M. (2018). Processing of emotional words in bilinguals: Testing the effects of word concreteness, task type and language status. *Second Language Research*, 34, 371–394. <https://doi.org/10.1177/0267658317744008>
- Ferré, P., García, T., Fraga, I., Sánchez-Casas, R., & Molero, M. (2010). Memory for emotional words in bilinguals: Do words have the same emotional intensity in the first and in the second language? *Cognition and Emotion*, 24, 760–785. <https://doi.org/10.1080/02699930902985779>
- Ferré, P., Sánchez-Casas, R., & Fraga, I. (2013). Memory for emotional words in the first and the second language: Effects of the encoding task. *Bilingualism: Language and Cognition*, 16, 495–507. <https://doi.org/10.1017/S1366728912000314>
- Ferré, P., Sánchez-Casas, R., & Guasch, M. (2006). Can a horse be a donkey? Semantic and form interference effects in translation recognition in early and late proficient and non-proficient Spanish–Catalan bilinguals. *Language Learning*, 56, 571–608. <https://doi.org/10.1111/j.1467-9922.2006.00389.x>
- Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35, 116–124. <https://doi.org/10.3758/BF03195503>
- Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, 48, 963–972. <https://doi.org/10.3758/s13428-015-0621-0>
- Guasch, M., Boada, R., Ferré, P., & Sánchez-Casas, R. (2013). NIM: A Web-based Swiss Army knife to select stimuli for psycholinguistic studies. *Behavior Research Methods*, 45, 765–771. <https://doi.org/10.3758/s13428-012-0296-8>
- Guasch, M., Ferré, P., & Haro, J. (2017). Pupil dilation is sensitive to the cognate status of words: Further evidence for non-selectivity in bilingual lexical access. *Bilingualism: Language and Cognition*, 20, 49–54. <https://doi.org/10.1017/S1366728916000651>
- Guasch, M., Sánchez-Casas, R., Ferré, P., & García-Albea, J. E. (2008). Translation performance of beginning, intermediate and proficient Spanish–Catalan bilinguals: Effects of form and semantic relations. *Mental Lexicon*, 3, 289–308. <https://doi.org/10.1075/ml.3.3.03gua>
- Guasch, M., Sánchez-Casas, R., Ferré, P., & García-Albea, J. E. (2011). Effects of the degree of meaning similarity on cross-language

- semantic priming in highly proficient bilinguals. *Journal of Cognitive Psychology*, 23, 942–961. <https://doi.org/10.1080/20445911.2011.589382>
- Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41, 401–410. <https://doi.org/10.1037/h0056020>
- Huang, X. (2017). The role of word frequency and contextual diversity in visual word recognition: A mini review. *New Frontiers in Ophthalmology*, 3, 1–4. <https://doi.org/10.15761/NFO.1000185>
- Imbs, P. (1971). *Etudes statistiques sur le vocabulaire français. Dictionnaire des fréquences, Vocabulaire littéraire des XIX^e et XX^e siècles*. Centre de la Recherche pour un Trésor de La Langue française (CNRS), Nancy, Paris, Librairie Marcel-Didier.
- Institut d'Estudis Catalans. (1995). *Diccionari de la llengua catalana*. Barcelona, Spain: IEC.
- Instituto Nacional de Estadística. (2011). *Spanish Time Use Survey 2009–2010*. Retrieved July 16, 2018, from <http://www.ine.es/inebmenu/indiceAZ.htm>
- Kandel, S., Burfin, S., Méary, D., Ruiz-Tada, E., Costa, A., & Pascalis, O. (2016). The impact of early bilingualism on face recognition processes. *Frontiers in Psychology*, 7, 1080. <https://doi.org/10.3389/fpsyg.2016.01080>
- Keuleers, E. (2013). vwr: Useful functions for visual word recognition research (R package version 0.3.0). Retrieved from <https://CRAN.R-project.org/package=vwr>
- Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42, 643–650. <https://doi.org/10.3758/BRM.42.3.643>
- 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. <https://doi.org/10.3758/s13428-011-0118-4>
- Kilgarriff, A. (2006). *BNC database and word frequency lists*. Retrieved Jul 16, 2018, from <http://www.kilgarriff.co.uk/bnc-readme.html>
- 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, 172–175.
- Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London, UK: Longman.
- Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, 47, 471–483. <https://doi.org/10.3758/s13428-014-0489-4>
- Martin, C. D., Strijkers, K., Santesteban, M., Escera, C., Hartsuiker, R. J., & Costa, A. (2013). The impact of early bilingualism on controlling a language learned late: An ERP study. *Frontiers in Psychology*, 4, 815. <https://doi.org/10.3389/fpsyg.2013.00815>
- Moldovan, C. D., Demestre, J., Ferré, P., & Sánchez-Casas, R. (2016). The role of meaning and form similarity in translation recognition in highly proficient balanced bilinguals: A behavioral and ERP study. *Journal of Neurolinguistics*, 37, 1–11. <https://doi.org/10.1016/j.jneuroling.2015.07.002>
- New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28, 661–667. <https://doi.org/10.1017/S014271640707035X>
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36, 516–524. <https://doi.org/10.3758/BF03195598>
- 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. <https://doi.org/10.1016/j.jecp.2012.10.014>
- 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. <https://doi.org/10.1037/a0034058>
- Rabovsky, M., Sommer, W., & Abdel Rahman, R. A. (2012). The time course of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience*, 6, 11. <https://doi.org/10.3389/fnhum.2012.00011>
- Rafel, J. (1998). *Diccionari de freqüències*. Barcelona, Spain: Institut d'Estudis Catalans.
- Rodríguez-Pujadas, A., Sanjuán, A., Ventura-Campos, N., Román, P., Martín, C., Barceló, F., . . . Ávila, C. (2013). Bilinguals use language-control brain areas more than monolinguals to perform non-linguistic switching tasks. *PLoS ONE*, 8, e73028. <https://doi.org/10.1371/journal.pone.0073028>
- Rosa, E., Tapia, J. L., & Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS ONE*, 12, e0179004. <https://doi.org/10.1371/journal.pone.0179004>
- Sebastián-Gallés, N., Martí, M. A., Carreiras, M., & Cuetos, F. (2000). *LEXESP: Una base de datos informatizada del español*. Barcelona, Spain: Universitat de Barcelona.
- Shaoul, C., & Westbury, C. (2013). A reduced redundancy USNET corpus (2005–2011). Edmonton, AB: University of Alberta. Retrieved July 18, 2018, from www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html
- Simons, G. F., & Fennig, C. D. (2018). *Ethnologue: Languages of the world* (21st ed.). Dallas, TX: SIL International. Retrieved July 16, 2018, from <http://www.ethnologue.com>
- 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.
- Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J. J., . . . Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology*, 68, 680–696. <https://doi.org/10.1080/17470218.2014.964271>
- Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, 24, 208–214.
- van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67, 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Vergara-Martínez, M., Comesaña, M., & Perea, M. (2017). The ERP signature of the contextual diversity effect in visual word recognition. *Cognitive, Affective, & Behavioral Neuroscience*, 17, 461–474. <https://doi.org/10.3758/s13415-016-0491-7>
- Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, 60, 502–529. <https://doi.org/10.1016/j.jml.2009.02.001>
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15, 971–979. <https://doi.org/10.3758/PBR.15.5.971>
- Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.
- Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.