Behav Res (2011) 43:449–458
DOI 10.3758/s13428-011-0062-3

Oral frequency norms for 67,979 Spanish words

María Angeles Alonso · Angel Fernandez · Emiliano Díez

Published online: 17 March 2011

Abstract Frequency of occurrence is an important attribute of lexical units, and one that is widely used in psychological research and theorization. Although printed frequency norms have long been available for Spanish, and subtitle-based norms have more recently been published, oral frequency norms have not been systematically compiled for a representative set of words. In this study, a corpus of over three million units, representing present-day use of the language in Spain, was used to derive a frequency count of spoken words. The corpus consisted of 913 separate documents that contained transcriptions of oral recordings obtained in a wide variety of situations, mostly radio and television programs. The resulting database, containing absolute and relative frequency values for 67,979 orally produced words, is presented. Validity analyses showed significant correlations of oral frequency with other frequency measures and suggest that oral frequency can predict some types of lexical processing with the same or higher levels of precision, when contrasted with text- or subtitle-based frequencies. In conclusion, we discuss ways in which these oral frequency norms can be put to use. The norms can be downloaded from www.springerlink.com.

Keywords Word frequency · Spoken word frequency · Frequency estimates · Spanish norms

Electronic supplementary material The online version of this article (doi:10.3758/s13428-011-0062-3) contains supplementary material, which is available to authorized users.

M. A. Alonso (✉)
Facultad de Psicologia, Universidad de La Laguna,
Campus de Guajara 38205 La Laguna, Spain
e-mail: maalonso@ull.es

A. Fernandez · E. Díez
Universidad de Salamanca,
Salamanca, Spain

Lexical frequency is a descriptive attribute that makes reference to the extent to which a particular word is used by the speakers of a given language. The findings of many diachronic linguistic studies have suggested that the frequency with which linguistic units are produced and encountered in the course of verbal interactions modulates the way in which languages change over time, affecting variations in phonology, morphology, and grammar (Bybee, 2007). For example, quantitative analyses of word usage throughout extended periods of time have revealed that, at least in Indo-European languages, terms that occur very frequently in discourse are more resistant to change than are those in lesser use (Pagel, Atkinson, & Meade, 2007). Additionally, and more relevant to the purpose of this study, the results of many psychological experiments entailing the control and manipulation of verbal materials indicate that lexical frequency is involved in important aspects of human cognition related to language and memory.

Psycholinguistic studies have consistently found that, when other variables are held constant, higher frequency words are more readily processed than lower frequency words in tasks such as word recognition, lexical decision, and naming (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Forster & Chambers, 1973; Grainger, 1990; Howes & Solomon, 1951). Lexical frequency has also been demonstrated to be an important word attribute in empirical studies focusing on memory for verbal materials. As an example, higher frequency words are typically recalled better than lower frequency words (Deese, 1960), whereas recognition accuracy is usually superior for lower than for higher frequency words (Gregg, 1976; Mandler, Goodman, & Wilkes-Gibbs, 1982). A general conclusion emerging from these and many other studies in language and memory is that word frequency is a reliable indicator of the accessibility of lexical representations stored in memory (Nelson & McEvoy,

2000), and therefore that the availability of reliable frequency indexes can be of the greatest importance for both the implementation of strict control on the characteristics of experimental verbal materials and the testing of particular hypotheses about the role played by word frequency in a number of cognitive processes.

Word frequency is usually understood as the rate with which a given word is used in everyday contexts involving processing of natural language, be it speech or printed text. Estimations are usually presented in the form of language-specific frequency norms that provide quantitative descriptions of usage for specific individual words, and they can be computed using two kinds of broadly different procedures. One approach has been to rely on subjective estimations given by samples of native speakers of the language of interest. Some of these studies (e.g., Balota, Pilotti, & Cortese, 2001, in English; Gonthier, Desrochers, Thompson, & Landry, 2009, in French) are aimed at directly obtaining subjective frequency judgments, with groups of participants usually required to provide a scale-based rating for each of a limited set of words, indicating the frequency with which they tend to encounter each word in their daily life using a limited range of numerical values (e.g., 1 to 7). A different approach, also using subjective methodology, has been to obtain scale-based judgments of word familiarity from samples of native speakers (e.g., Gernsbacher, 1984; Toglia & Battig, 1978), assuming that familiarity judgments reflect, to a large extent, variation in the number of times that a particular word has been encountered by an individual. Another, more prevailing approach has been to use objective methods, directly deriving frequency indexes from the total number of occurrences of words within a wide sample of written materials (e.g., Kučera & Francis, 1967) or spoken records (e.g., Brown, 1984). These are ratio indexes that reflect the absolute or relative number of instances of a given word within the sample of reference, and they are taken as estimators of the frequency of the word in the population at large, always bearing in mind that frequency indexes of this kind are systematically affected by sample size, and that word use in natural situations is not randomly distributed (Baayen, 2001).

Subjective and objective methods can each have advantages and disadvantages, and arguments have been advanced in favor of each approach. For example, Gernsbacher (1984) defended the use of subjective estimators of frequency, on the evidence that familiarity ratings did not always match printed frequency estimators and were, nonetheless, potent predictors of word recognition latencies. More recently, Thompson and Desrochers (2009) reviewed reliability and validity issues affecting objective and subjective frequency estimators and concluded that, while both are likely to be valid indicators of the same psychological construct, both tend to be biased, each in its own way. Thus, the choice between oral or written indicators, or even the option of their conjoint use, should depend on the specific assumptions and goals of particular lines of research. In any case, the fact is that objectively derived frequency norms have been more widely used in psychological experimentation. Among the reasons for their greater use, an important one is that some early objective frequency indexes (e.g., Kučera & Francis, 1967; Thorndike & Lorge, 1944) have become standardized materials, shared by numerous researchers, that facilitate comparisons across different studies. A more practical reason is that, overall, objective norms tend to contain indexes for larger sets of words.

As mentioned above, objective frequency norms can be of two different types: written and oral. Written frequency values are usually extracted from a sample of records selected from printed materials of varied origins (such as works of fiction, magazines, and newspapers), and they represent, for each word, the total number of times it appears within the whole sample of selected texts. Early and widely used sources for written frequency in English have been provided by Thorndike and Lorge (1944), who published printed frequency range values for a set of 30,000 words, and by Kučera and Francis (1967), who obtained frequency counts for 50,406 words appearing in a sample of approximately 1 million words. Frequency norms derived from printed materials also exist in other languages, such as Italian (Barca, Burani, & Arduino, 2002), French (New, Pallier, Brysbaert, & Ferrand, 2004), German and Dutch (Baayen, Piepenbrock, & Gulikers, 1995), Greek (Ktori, van Heuven, & Pitchford, 2008), Chinese (Liu, Shu, & Li, 2007), and Arabic (Boudelaa & Marslen-Wilson, 2010). The same basic counting procedure has been used to obtain oral frequency values, although the sampled materials have in this case been transcripts of spoken interventions by people in different situations (e.g., talks, interviews, and telephone conversations). Using this methodological approach, Brown (1984) published values for nearly 5,000 English words appearing in spontaneous conversation. In a more recent study, Pastizzo and Carbone (2007) computed spoken frequency counts for 34,922 English words appearing in a sample of 1.6 million words included in transcripts of oral exchanges recorded in an academic setting. Similar spoken frequency estimates are available in other languages, such as French (Equipe DELIC, 2004) or Italian (De Mauro, Mancini, Vedovelli, & Voghera, 1993).

Assembling objective frequency norms is a laborious task and, in general, written counts have been easier to obtain, in part because of the availability of written records, and in part because of the simpler way in which

computations can be performed, quite directly from machine-readable text files. This point is clearly illustrated by the possibility of computing frequency counts from very large corpora containing over 15 million words from written documents (e.g., Baayen et al., 1995; Zeno, Ivens, Millard, & Duvvuri, 1995) or over 100 million words from written samples available on the Internet (see Balota, Yap, Cortese, Hutchison, Kessler, Loftis, et al., 2007; Burgess & Livesay, 1998). Publicly available records of spoken interventions have been harder to obtain until now, because their computational treatment requires an intermediate stage of transcription, from sound to print, that cannot be fully automatized with current technologies. Therefore, it is not surprising that written norms have more commonly been the choice when researchers have developed objective frequency norms. However, New, Brysbaert, Veronis, and Pallier (2007) recently showed that computational limitations affecting the transformation from oral to written transcriptions could be overcome by using subtitles belonging to verbal exchanges in film and television series. In this seminal article, they reported the construction of a corpus of 52 million words in French, and the results of analyses showing that subtitle-based frequency estimators had high correlations with other measures of frequency (spoken and printed) and were good predictors of performance in a lexical processing task. Since then, the same type of frequency norms, all based on large-size corpora, have been elaborated for English (Brysbaert & New, 2009), Dutch (Keuleers, Brysbaert, & New, 2010), and Chinese (Cai & Brysbaert, 2010), and in all cases the subtitle-based frequency values have demonstrated remarkable congruency and criterion validities. As a matter of fact, these newly developed frequency estimators have demonstrated higher predictive value than widely used printed frequency indexes in tasks such as lexical decision, leading to the suggestion that in choosing frequency norms, one should give preference to the ones extracted from the largest sample size (Brysbaert & New, 2009).

However, correlations are typically not perfect, indicating that there may be important aspects in which written and spoken norms can differ. For example, Allwood (1998) analyzed the differences in word frequency between spoken and written Swedish using two corpora of similar size, and found that there are fewer unique words in the case of the spoken sample, that common words are more common in spoken language, and that some parts of speech (e.g., pronouns, conjunctions, and adverbs) are more common in oral language, where others (e.g., prepositions, nouns, and adjectives) are more common in written language. It is also reasonable to assume that some words, such as vulgarisms, may be rarely used in print, whereas other types of words, such as technical terms, are more likely to be used in print (Dahl, 1979). More importantly, the results of some experimental studies (e.g., Brown & Watson, 1987; Pastizzo & Carbone, 2007) have indicated that spoken and written frequency have different effects on some tasks, suggesting that they may not be indicators of a single, unique dimension. Therefore, the decision about which type of norms are likely to be more appropriate may reasonably be made not only on the basis of corpus size, but also by taking into consideration the particular aims, demands, and assumptions of specific studies, and the availability of norms for both spoken and written language may increase opportunities for control and manipulation in experimental research involving verbal materials.

Objective frequency norms have a long history and, as detailed above, have been elaborated for a variety of languages. In the case of the Spanish language, the compilation of objective frequency norms had a relatively early start, with the work of Buchanan (1927), who reported the written frequency of 6,702 words on the basis of their presence in a sample of 1,200,000 words extracted from a variety of texts published in Spain and Latin America. Since then, several printed frequency counts have been published (see Pérez, Campoy, & Navalón, 2001, for an extensive list of normative studies in Spanish). In recent times, there are two databases for printed frequency that are widely used by researchers conducting studies with verbal materials in Spanish. Alameda and Cuetos (1995) used an extensive sample of texts, containing 2 million words, to derive frequency counts for a total of 81,323 different words. Building on this effort, Sebastián, Martí, Carreiras, and Cuetos (2000) further expanded the source sample to produce LESEXP. This computerized database incorporated the complete Alameda and Cuetos text sample plus an extra sample of 3 million words, providing improved frequency counts and additional objective and subjective lexical indices for a much larger set of individual words. In comparison with previous printed-frequency norms, these two contributions have the advantage of being based on textual samples that are considerably larger, more diverse, and more contemporary, resulting in more reliable normative information for a larger set of words.

The availability of the other type of objectively derived frequency norms in Spanish, the ones based on spoken productions, has been more restricted. One reason is that the majority of studies, often driven by applied educational goals, provide frequency data obtained from the oral production of small groups of children (e.g., Serra, Serrat, Solé, Bel, & Aparici, 2000). With respect to norms based on the spoken productions of adults, the only database known to us for oral frequency norms was compiled by Ávila (1999), based on a corpus of 500,000 words extracted from the oral productions of 291 speakers in a single city in Spain (Málaga). The situation has changed very recently, with the publication of a subtitle-based frequency database (Cuetos, González-Nosti,

Barbón, & Brysbaert, in press) obtained from a corpus of over 41 million words used in films and television series. As in the case of the recent, similar norms for other languages, these new frequency indicators have been reported to possess good overall validity, both in terms of correlations with other Spanish printed frequency norms and in terms of predicting performance in two lexical processing tasks (naming and lexical decision).

The oral norms described in the present report were conceived as a potentially valuable research tool, contributing to the remediation of an important absence of normative materials that was apparent at the start of the project. The publication of the subtitle-based norms is a very important step in the achievement of this goal, but there are reasons to assume that the availability of the norms presented here may still constitute a worthy contribution. First, discrepancies in the contents and specific frequency values are likely to be found when the two types of norms are compared (see New et al., 2007), opening the opportunity to conduct more refined and better-informed selections of materials when oral frequency is of importance. Second, the spoken norms presented here can be a useful addition to the set of frequency norms already available for use in Spain (based on subjective estimations, written documents, or subtitles), facilitating the application of selection strategies based on the use of multiple-source, matched-frequency indexes. Finally, but no less importantly, these norms can contribute to further investigative efforts that could, in the future, be aimed at determining the comparative virtues of each type of norm in relation to lexical-processing tasks other than word naming and lexical decision, or to still-unexplored applications to recall and recognition memory tasks.

In what follows, the source corpus, the data extraction procedure, and the characteristics of the normative database are described, followed by a report of validity analyses.

The source corpus

The oral frequency norms presented in this article were constructed by extracting single-word counts from the transcriptions of oral documents included in the Corpus de Referencia del Español Actual (CREA) [Reference Corpus for Present-Day Spanish], an extensive and rich database created and maintained by the Real Academia Española. In its current state, CREA consists of a large collection of records, from varied written and oral sources dated between 1975 and 2004, that altogether contains approximately 170 million Spanish words (see Real Academia Española, 2010b).

The oral component of CREA presently incorporates written verbatim transcripts corresponding to 2,000 spoken documents, the total number of words being approximately 10 million. Each transcribed document contains standardized orthographic renditions of all of the verbal utterances contained in the spoken document. The transcription process was carried out by trained coders, who listened to the spoken source recordings and manually entered the corresponding orthographic versions of the units into computers. Coders used customized word-processing software, for documents prior to 1999, and a specialized transcription program (Transcriber; see Barras, Geoffrois, Wu, & Liberman, 2001) for later documents. The transcription procedure was exhaustive, and the database contains the recorded orthographic forms of all utterances in each spoken document, including production errors, truncated words, repeated words, foreign words, and so forth. In addition, phenomena such as speaker overlap in conversations, turn taking, and pauses were coded using markup languages such as SGML and XML (see Pino & Sánchez, 1999, and Sánchez, 2005, for detailed descriptions of the selection, acquisition, and transcription procedures).

Because the aim of CREA is to reflect present-day word usage in as many dialectal variations of the Spanish language as possible, the source documents are of varied geographical origin, including transcripts of oral samples obtained in Spain and also in most Spanish-speaking countries in America. However, in order to preserve homogeneity in the norms, only the documents originating in Spain were used for the present study. As a result, the target segment of the corpus used to elaborate the present norms was a subset of the general oral corpus, consisting of a total of 913 transcribed oral documents of varied nature, most of them sound recordings of radio and television programs. The oldest and largest set of documents (85%) corresponds to transcriptions of recordings created between 1975 and 1999, which formed the initial CREA corpus that was available in 2000. The remaining 15% of the documents are transcriptions of later recordings created between 2000 and 2004, and have been incorporated into the extended new version of the corpus made available in 2008. The genres represented in this set of documents were general news (8%), feature reports and documentaries (9%), interviews (18%), debates and talk shows (18%), live sport broadcasts and sport shows (9%), magazines and variety shows (35%), and lottery and game shows (2%). An additional 1% of the documents were a miscellany of recordings of both formal and informal verbal situations (e.g., lectures, sermons, supermarket PA messages, and phone conversations).

The spoken frequency database

The total number of separate transcribed string units contained in the 913 documents was 3,199,614, a pool from which a word-count software program (Scott, 2004)

identified an initial raw set of 79,855 unique entries and computed the number of occurrences for each of them. With the aim of eliminating transcription errors and instances of word fragments, a multistep depuration procedure on this initial set of types was followed, with automatic filtering processes being used first. Single letters or letter strings were selected as valid words if they were included in electronically accessible databases of current use among Spanish researchers, such as printed frequency norms (Alameda & Cuetos, 1995; Sebastián et al., 2000), category norms (Soto, Sebastián, García, & del Amo, 1994), and association norms (Fernandez, Díez, & Alonso, 2010; Fernandez, Díez, Alonso, & Beato, 2004). Additionally, units were selected as words if they corresponded to entries in the online version of the 22nd edition of the Spanish dictionary published by Real Academia Española (2001), to names listed in an online dictionary of male and female names (www.scribd.com/doc/448666), to surnames included in an online list of over 4,000 family names (www.atienza.org/apellido.htm), and to foreign words included in the English, French, Italian, and German dictionaries available for online search in the Free Dictionary (www.thefreedictionary.com). Finally, the remaining set of candidate units were visually inspected by the three authors, who corrected unambiguous misspellings and considered that the units were proper words if there was a 100% agreement on their status. These words were mostly inflected forms (plurals and verb forms) not included in the searched databases and dictionaries. As a result of this process, a total of 11,876 units in the initial raw set were definitely excluded. These were all low-frequency units that corresponded to a total of 25,454 entries in the original corpus. Although their exclusion had a low impact in the overall contents of the initial corpus (a reduction of 0.79% of the string units, not affecting the number of documents), the relative-frequency values provided in the present norms were computed taking this reduction into account. The final normative database was therefore formed by a set of 67,979 unique entries (*types*) corresponding to inflected forms in Spanish, drawn from a refined corpus of 3,174,189 transcribed units (*tokens*).

Unsurprisingly, there was large variability in the number of times these words were repeated throughout the source pool; the absolute raw frequencies ranged from a minimum of 1 to a maximum of 149,996, with an overall mean frequency of 46.67 and a median of 2.00. As is characteristic of word frequency distributions, words were not normally distributed across the range of frequencies, and low-frequency words constituted a vast majority (a sizable subset of 25,940 words had a raw frequency of 1, and words with a raw frequency of less than 20 [or 6 per million] represented 87.5% of the total). Accordingly, the distribution showed a marked positive skewness of 79.30 (SE = 0.009) and a kurtosis value of 7.75 (SE = 0.02). Word length, in number of characters, varied between 1 and 24, with a mean value of 8.46 (SD = 2.57).

The database is available for downloading from www.springerlink.com in two different file formats, text (spanish_oral_freq.csv) and spreadsheet (spanish_oral_freq.xls), both with identical contents and general structure. In both files, all of the words are listed in alphabetical order, and each of them is followed by three quantitative indexes. The first of them is the absolute frequency of the word in the source corpus; the second index represents frequency per million; and the third index represents $\log_{10}$(absolute frequency +1). These last two indicators are provided because they are very generally used when word frequency is a variable in psychological research.

Congruent and criterion validity

One way to assess the validity of the collected oral frequency values is to observe their relation with other frequency indicators in the same language. To accomplish this analysis, we computed the Pearson correlation coefficient between our oral frequency values and three different estimates of frequency available in Spanish: subjective frequency (Desrochers, Liceras, Fernández-Fuertes, & Thomson, 2010), written frequency (Sebastián et al., 2000), and subtitle-based frequency (Cuetos et al., in press). Table 1 displays the results of this analysis, showing, as expected, that the oral frequency estimates were positively correlated with the three other estimates.

Furthermore, and in order to explore the predictive validity of our oral frequency estimates, we conducted regression analyses with two performance measures, word naming times and picture naming times, as dependent variables and word length and three frequency estimates (oral frequency, written frequency, and subtitle-based frequency) as predictors.

Table 1 Correlations between oral frequency and subjective frequency, objective word frequency from written texts (LEXESP), and objective frequency from subtitles (SUBTLEX-ESP)

	Oral Frequency	Number of Common Words
Subjective Frequency	.68	171[a]
Written Frequency (LEXESP)	.79	52,257
Subtitle Frequency (SUBTLEX-ESP)	.67	42,609

All frequencies were transformed to $\log_{10}$(frequency + 1). All correlation coefficients are significant ($p < .001$). [a] Compound words were excluded from the data set for this analysis

Word naming times were obtained from Cuetos and Barbón (2006), who reported reading latencies for a set of 240 common Spanish words visually presented in isolation. Picture naming times were drawn from Cuetos, Ellis, and Álvarez (1999), who reported naming latencies to a set of 140 individually presented black-and-white standardized drawings of common objects. These performance data sets are admittedly small (cf. Brysbaert & New, 2009, who used a set of over 40,000 lexical decision and word naming times in similarly oriented analyses), but they are the only ones that have been found for Spanish. Lexical decision and naming times for a larger set of 2,764 Spanish words have been used in their validation of subtitle frequencies by Cuetos et al. (in press), but these lexical processing data have not yet been published and were still not available from their developers at the time of writing the present report. The results of the analyses with the available data sets are shown in Table 2. As can be observed, the predictive values of the different frequency measures in the case of word naming times were quite similar, all of them accounting for nearly 30% of the variance, and were quantitatively similar to estimations obtained with larger data sets in Spanish (Cuetos et al., in press) and other languages (e.g., Brysbaert & New, 2009; Cai & Brysbaert, 2010). In the case of picture naming, the predictive values of the three frequency measures were lower, but in this case oral frequency estimates did best, with a gain of nearly 6% relative to written frequency, and a gain of 17.7% relative to subtitle-based frequency. These results indicate, first, that oral frequency estimates are as good as other frequency measures at predicting word reading times and, second, that in some tasks (e.g., picture naming), oral frequency may be a better predictor than other frequency measures.

Table 2 Results of the regression analyses, with word frequency and word length as predictors, and word and picture naming times as dependent variables

Model	Word Naming Times (Cuetos & Barbón, 2006)		Picture Naming Times (Cuetos, Ellis, & Álvarez, 1999)	
	Weight	Adjusted R^2	Weight	Adjusted R^2
Oral Frequency	–8.35	.299**	–71.94	.210**
Length	10.71		5.82	
	(N = 234)		(N = 135)	
LEXESP	–8.85	.293**	–63.08	.154**
Length	10.78		7.60	
	(N = 234)		(N = 135)	
SUBTLEX-ESP	–5.58	.290**	–2.17	.033*
Length	11.05		14.46	
	(N = 233)		(N = 135)	

All frequencies were transformed to $\log_{10}$(frequency + 1). * $p < .05$. ** $p < .001$.

Conclusion

The frequency norms presented here were elaborated with the aim of adding new oral frequency estimators for a substantial number of stimuli to the existing repertoire of normative studies available for linguistic stimuli in Spanish. In what follows, issues relevant to the interpretation of the norms and their possible limitations are discussed, and ways in which the provided normative data could facilitate the work of researchers are considered.

The source corpus from which the norms were derived consisted of oral productions recorded in a period covering the last few decades and, according to its developers—linguistic experts at the Real Academia Española—it reflects present-day use of Spanish. Nonetheless, some specific entries may have varied in use over the years and may, at present, be more or less frequent than is indicated in the present database. This particular problem can affect most of the linguistic norms habitually used in psychological research that are based on lexical productions by human participants. For example, free association norms usually provide some strong stimulus–response relationships that result from overexposure to word combinations that are particularly frequent at the time of data collection. Even norms that are more likely to reflect rather crystallized lexical characteristics, such as category membership, may be influenced by transitory shifts in the use of concepts and words (Novick, 2003). The issue of stability is also relevant for any type of objectively derived frequency norms (e.g., those based on printed materials or those based on subtitles), and could be more critical when dealing with oral frequency estimates, because of the presence of colloquialisms that may show a rapid pattern of variation in use over time. Although no perfect solution to this problem is likely to exist, steps to minimize some of its effects could be taken in the future by making periodic updates of the corpus, focused on increasing the number and variety of documents from more recent periods of time. The ongoing plan of the Real Academia, which aims at incorporating numerous oral documents into the CREA database every 5 years (Real Academia Española, 2010a), will certainly facilitate future revisions of the norms.

It should also be noted that the spoken documents constituting the source corpus were all specifically selected with the criterion that they originated in Spain, with the aim of producing normative data closely adjusted to a particular population. Thus, generalization of the frequency indexes to other dialectal variations of Spanish, such as those spoken in the American continent, cannot be taken for granted. Although this may be an issue for many normative

studies, it is particularly noteworthy here, because it is very possible that there is more dialect-determined variation in spoken uses than in written uses of the same language. Fortunately, as mentioned in the description of the corpus, the CREA database also contains transcripts of oral documents recorded in many other Spanish-speaking countries, and elaboration of population-specific oral frequency norms could be a valuable contribution of future work in this area.

One relevant issue that deserves consideration is the size of the corpus from which the frequency estimators were extracted, which contained a total of 3,174,189 transcribed tokens. The same type of norms for English words, compiled by Brown (1984), were based on a corpus of 191,918 units; those compiled more recently by Pastizzo and Carbone (2007) for the same language were based on a corpus of 1,630,376 units; norms for French words (Equipe DELIC, 2004) have been obtained from a corpus of 438,378 units; and norms for Italian words (De Mauro et al., 1993) from a corpus of 500,000 units. Therefore, in relative terms, when compared to the corpora used to obtain equivalent norms in other languages, our source corpus can be considered large, virtually doubling in size the largest of the mentioned corpora. However, some studies have obtained results indicating that extensive corpora have advantages over smaller ones (e.g., Balota et al., 2004; Burgess & Livesay, 1998; Zevin & Seidenberg, 2002). Also, recent research analyzing the effect of corpus size on the validity of frequency estimations in predicting lexical-processing times has led to the conclusion that, because measurement errors particularly affect the low frequency values obtained from smaller word samples, the ideal corpus should contain between 10 and 30 million words (Brysbaert & New, 2009). In light of these considerations, the corpus used to compute our frequency norms has to be considered small in absolute terms, and our frequency estimators are likely to account for approximately 4% less of the variance tied to lexical processing than are those derived from a 30-million-word corpus, when low-frequency words are considered (extrapolating from data in Table 3 of Brysbaert & New, 2009). Nonetheless, as demonstrated by the validity analyses, the present frequency norms have significant correlations with other available and commonly used frequency norms in Spanish, and thus are able to significantly predict performance in basic lexical-processing tasks, such as word and picture naming, to a reasonable extent.

Furthermore, these norms can also be considered valuable if more qualitative aspects are taken into consideration. In this regard, it is worth noting that when the total list of words included in our norms was matched with the complete contents of SUBTLEX-ESP (Cuetos et al., in press), there were 25,376 words from our norms that were not present in the subtitle-based listing. As has been observed in other cases (e.g., New et al., 2007), these were mostly low-frequency words, and many could be classified as regionalisms, onomatopoeias, parts of idioms, and so on. However, a substantial set of high-frequency words were also not included in SUBTLEX-ESP. Focusing on the 114 of these words that had a frequency over 30 per million in our norms, we found that 9 words (8%) corresponded to category exemplars that are part of habitual lexical productions in Spanish, according to available category norms (Soto et al., 1994). The average frequency production of these exemplars was 32.4 in a sample of 356 subjects, which places these words among the first 20 produced in the standard category production task. Also, 82 of these high-frequency words (72%) were found to correspond to free association responses collected in an independent normative study that contained associative responses to over 4,000 stimuli (Fernandez et al., 2010; Fernandez et al., 2004). These words varied in terms of the number of different cues to which they were produced as responses, going from a minimum of 1 to a maximum of 156. The following provide illustrative examples of how common these 82 words are in everyday use: *abuela* ("grandmother") was a free association response to 156 different cues; *amiga* ("friend," in feminine), a response to 138 different cues; *falso* ("false"), a response to 102 cues; and *ciencia* ("science"), a response to 92 cues. In summary, a sizable number of significant words in the norms presented here are frequently used, yet are absent in the subtitle norms based on a much larger source corpus. While this finding does not compromise the overall validity and usefulness of SUBTLEX-ESP, it highlights the virtues of having alternative sources of normative materials from which researchers can draw quantitative information according to their needs. In this regard, it is interesting to note that some of the reliability and validity problems typically afflicting frequency norms could be, at least to some extent, alleviated by adopting the strategy of using multiple frequency indexes for the selection of stimuli (Thompson & Desrochers, 2009).

In spite of the limitations mentioned above, which both help to qualify our findings and point to future necessary and interesting research efforts, these and recently available oral frequency norms could be of help in a number of ways to cognitive researchers who habitually employ words in experiments conducted in Spain. On the one hand, stimulus control could be more finely achieved by having a more varied set of quantitative lexical indicators; on the other hand, and perhaps more importantly, new manipulations could now be included in experimental designs, allowing for the testing of more specific frequency-related hypotheses

Oral frequency estimators would also be particularly useful in experimental situations in which the characteristics of participant samples suggest that using printed frequency might be inappropriate—for example, when participants have few years of formal education or are not frequently exposed to printed media in their daily life. This could be an important consideration to take into account when working with samples of older participants because, according to recent survey data reported by the Spanish Publisher's Federation (Federación de Gremios de Editores de España, 2010), more than 50% of people over 55 years of age have to be classified as nonreaders of books (they never or almost never read a book), and 35% of the people in this population segment do not even read newspapers or magazines. Under such circumstances, it seems reasonable to suggest spoken frequency as the variable of choice. A similar reasoning applies to clinical situations that may require cognitive assessment through tests and tasks that demand processing of frequency-adjusted verbal stimuli (e.g., recall and recognition of frequent and infrequent words): Spoken frequency would be the choice in the case of patients who may be illiterate (e.g., the very old) or underexposed to textual information in Spanish (e.g., recent adult immigrants).

Another potential area of use for the oral frequency norms is related to the use of stimuli in the auditory modality. When not precluded by the testing of specific hypotheses, most experimental procedures have typically made use of the visual presentation modality, because it is relatively easy to handle in the standard computerized research laboratory. However, recent technological advances have made speech recording, storing, and manipulation more accessible and efficient (Oard, 2008), and therefore norms obtained from corpora of spoken records may become increasingly useful for investigating modality-dependent effects.

Beyond the psychological laboratory, word analyses have been performed many times with the aim of contributing to the design of procedures and materials in the field of second-language teaching. Frequency counts, based on both written and oral documents and aimed at the educational community, have often been assembled from a variety of sources (e.g., Davis, 2006). Some ways in which oral frequency indexes and other corpus-derived information can be used in designing course content and sequencing in second-language teaching have been explored by, among others, McCarten (2007) and McCarthy (2004).

Finally, the present database can be seen as a significant contribution to the recent development of normative studies of verbal stimuli in Spanish, an effort by several researchers toward the characterization of relatively large sets of stimuli in several important dimensions. As an example, there have been recent additions to the pool of printed frequency norms, consisting of two computerized databases of words appearing in school books widely used by beginning readers (Corral, Ferrero, & Goikoetxea, 2009) and by children under 12 (Martínez & García, 2004). Of special interest, because of its close relation to the norms presented here, is the recent contribution of Cuetos et al. (in press) and the subtitle-based frequency database SUBTLEX-ESP. Access to such a rich body of quantitative information about words in Spanish will allow for potentially illuminating comparisons between different types of frequency estimators and a more thorough understanding of the ways in which frequency, alone and interacting with other features and dimensions, determines the way in which we interact with words.

Author Note This work was supported by the Spanish Ministry of Science and Innovation (Project PSI2008-05607/PSIC) and by Junta de Castilla y León (Project SA031A/06). The authors express their gratitude to Guillermo Rojo, former Secretary of the Real Academia Española (Madrid, Spain), and to Mercedes Sánchez, at that institution's Departamento de Banco de Datos, for their generous help and support throughout the development of this project. The suggestions provided by anonymous reviewers of a previous version of this article and by the journal's Editor are also gratefully acknowledged. Supplemental materials may be downloaded along with this article from www.springerlink.com.

References

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

Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. In T. Haukioja (Ed.), *Proceedings of the 16th Scandinavian Conference of Linguistics* (pp. 18–29). Turku: University of Turku, Department of Linguistics.

Ávila, A. M. (1999). *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Universidad de Málaga.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

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

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2, 938 monosyllabic words. *Memory & Cognition, 29*, 639–647. doi:10.3758/BF03200465

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. doi:10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445–459. doi:10.3758/BF03193014

Barca, L., Burani, C., & Arduino, L. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers, 34*, 424–434. doi:10.3758/BF03195471

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication, 33*, 5–22. doi:10.1016/S0167-6393(00)00067-4

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for modern standard arabic. *Behavior Research Methods, 42*, 481–487. doi:10.3758/BRM.42.2.481

Brown, G. D. A. (1984). A frequency count of 190, 000 words in the London–Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers, 16*, 502–532.

Brown, G. D. A., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition, 15*, 208–216. doi:10.3758/BF03197718

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. doi:10.3758/BRM.41.4.977

Buchanan, M. A. (1927). *A graded Spanish word book*. Toronto: University of Toronto Press.

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. doi:10.3758/BF03200655

Bybee, J. (2007). *Frequency of use and the organization of language*. New York: Oxford University Press.

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

Corral, S., Ferrero, M., & Goikoetxea, E. (2009). LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behavior Research Methods, 41*, 1009–1017. doi:10.3758/BRM.41.4.1009

Cuetos, F., & Barbón, A. (2006). Word naming in Spanish. *European Journal of Cognitive Psychology, 18*, 415–436. doi:10.1080/13594320500165896

Cuetos, F., Ellis, A. W., & Álvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments, & Computers, 31*, 650–658. doi:10.3758/917BF03200741

Cuetos, F., González-Nosti, M., Barbón, A., & Brysbaert, M. (in press). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica*.

Dahl, H. (1979). *Word frequencies of spoken American English*. Essex: Verbatim.

Davis, M. (2006). *A frequency dictionary of Spanish: Core vocabulary for learners*. New York: Routledge.

De Mauro, T., Mancini, F., Vedovelli, M., & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato*. Milan: ESTALIBRI.

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports, 7*, 337–344. doi:10.2466/PR0.7.6.337-344

Desrochers, A., Liceras, J. M., Fernández-Fuertes, R., & Thomson, G. L. (2010). Subjective frequency norms for 330 Spanish simple and compound words. *Behavior Research Methods, 42*, 109–117. doi:10.3758/BRM.42.1.109

Equipe DELIC. (2004). Présentation du corpus de référence du français parlé. *Recherches sur le Français Parlé, 18*, 11–42.

Federación de Gremios de Editores de España (2010). *Hábitos de lectura y compra de libros en España 2009: Informe de resultados*. Retrieved from www.conectarc.com/ZOCO_Articulos.htm, July 24 2010

Fernandez, A., Díez, E., Alonso, M. A., & Beato, M. S. (2004). Free-association norms for the Spanish names of the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, & Computers, 36*, 577–583. doi:10.3758/BF03195604

Fernandez, A., Díez, E., & Alonso, M. A. (2010). Normas de Asociación libre en castellano de la Universidad de Salamanca [Online database]. Available at www.usal.es/gimc/nalc

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627–635. doi:10.1016/S0022-5371(73)80042-8

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*, 256–281. doi:10.1037/0096-3445.113.2.256

Gonthier, I., Desrochers, A., Thompson, G., & Landry, D. (2009). Normes d'imagerie et de fréquence subjective pour 1 760 mots monosyllabiques de la langue française. *Canadian Journal of Experimental Psychology, 63*, 139–149. doi:10.1037/a0015386

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29*, 228–244. doi:10.1016/0749-596X(90)90074-A

Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 183–216). New York: Wiley.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology, 41*, 401–410. doi:10.1037/h0056020

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. doi:10.3758/BRM.42.3.643

Ktori, M., van Heuven, W. J. B., & Pitchford, N. J. (2008). GreekLex: A lexical database of modern Greek. *Behavior Research Methods, 40*, 773–783. doi:10.3758/BRM.40.3.773

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

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39*, 192–198. doi:10.3758/BF03193147

Mandler, G., Goodman, G. O., & Wilkes-Gibbs, D. L. (1982). The word-frequency paradox in recognition. *Memory & Cognition, 10*, 33–42. doi:10.3758/BF03197623

Martínez, J. A., & García, E. (2004). *Diccionario de frecuencias del castellano escrito en niños de 6 a 12 años*. Salamanca: Universidad Pontificia de Salamanca.

McCarten, J. (2007). *Teaching vocabulary: Lessons from the corpus, lessons for the classroom*. Cambridge: Cambridge University Press.

McCarthy, M. J. (2004). *Touchstone: From corpus to course book*. Cambridge: Cambridge University Press.

Nelson, D. L., & McEvoy, C. L. (2000). What is this thing called frequency? *Memory & Cognition, 28*, 509–522. doi:10.3758/BF03201241

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28*, 661–677. doi: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. doi:10.3758/BF03195598

Novick, L. R. (2003). At the forefront of thought: The effect of media exposure on airplane typicality. *Psychonomic Bulletin & Review, 10*, 971–974. doi:10.3758/BF03196560

Oard, D. W. (2008). Unlocking the potential of the spoken word. *Science, 321*, 1787–1788. doi:10.1126/science.1157353

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature, 449*, 717–721. doi:10.1038/nature06176

Pastizzo, M. J., & Carbone, R. F., Jr. (2007). Spoken word frequency counts based on 1.6 million words in American English. *Behavior Research Methods, 39*, 1025–1028. doi:10.3758/BF03193000

Pérez, M. A., Campoy, G., & Navalón, C. (2001). Índice de estudios normativos en idioma español. *Revista Electrónica de Metodología Aplicada, 6*, 85–105.

Pino, M., & Sánchez, M. (1999). El subcorpus oral del Banco de Datos CREA-CORDE (Real Academia Española): Procedimientos de transcripción y codificación. *Oralia, 2*, 83–138.

Real Academia Española (2001). *Diccionario de la lengua española, vigésima segunda edición*. Available at http://buscon.rae.es/draeI

Real Academia Española (2010a). *El corpus del español del siglo XXI*. Retrieved November 10, 2010, from www.rae.es

Real Academia Española (2010b). *Corpus de referencia del español actual*. Retrieved November 10, 2010, from www.rae.es

Sánchez, M. (2005). El Corpus de Referencia del Español Actual (CREA): El CREA oral. *Oralia, 8*, 37–56.

Scott, M. (2004). *WordSmith Tools version 4*. Oxford: Oxford University Press.

Sebastián, N., Martí, M. A., Carreiras, M. F., & Cuetos, F. (2000). *LEXESP, léxico informatizado del español*. Barcelona: Ediciones de la Universitat de Barcelona.

Serra, M., Serrat, E., Solé, R., Bel, A., & Aparici, M. (2000). *La adquisición del lenguaje*. Barcelona: Ariel.

Soto, P., Sebastián, M. V., García, E., & del Amo, T. (1994). *Las categorías y sus normas en castellano*. Madrid: Visor.

Thompson, G. L., & Desrochers, A. (2009). Corroborating biased indicators: Global local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods, 41*, 452–471. doi:10.3758/BRM.41.2.452

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

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale: Erlbaum.

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

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47*, 1–29. doi:10.1006/jmla.2001.2834