Behav Res (2013) 45:1246–1258
DOI 10.3758/s13428-013-0326-1

EsPal: One-stop shopping for Spanish word properties

Andrew Duchon · Manuel Perea · Nuria Sebastián-Gallés · Antonia Martí · Manuel Carreiras

Published online: 7 March 2013

Abstract This article introduces EsPal: a Web-accessible repository containing a comprehensive set of properties of Spanish words. EsPal is based on an extensible set of data sources, beginning with a 300 million token written database and a 460 million token subtitle database. Properties available include word frequency, orthographic structure and neighborhoods, phonological structure and neighborhoods, and subjective ratings such as imageability. Subword structure properties are also available in terms of bigrams and trigrams, biphones, and bisyllables. Lemma and part-of-speech information and their corresponding frequencies are also indexed. The website enables users either to upload a set of words to receive their properties or to receive a set of words matching constraints on the properties. The properties themselves are easily extensible and will be added over time as they become available. It is freely available from the following website: http://www.bcbl.eu/databases/espal/.

Keywords Word frequency · Subtitles · Word recognition · Corpus linguistics · Psycholinguistics

A. Duchon (✉) · M. Carreiras
Basque Center on Cognition, Brain, and Language, Donostia, Spain
e-mail: a.duchon@bcbl.eu

M. Perea
Universitat de València, Valencia, Spain

N. Sebastián-Gallés
Universitat Pompeu Fabra, Barcelona, Spain

A. Martí
Universitat de Barcelona, Barcelona, Spain

M. Carreiras
IKERBASQUE. Basque Foundation for Science, Bilbao, Spain

Researchers from a wide range of disciplines (e.g., neuroscience, artificial intelligence, psychology, linguistics, and education, among others) who work in the interdisciplinary area of language research (e.g., language acquisition, language processing, language learning, bilingualism, and computational linguistics) need quick and efficient access to information about specific properties of words. For example, word frequency is a dominant factor in accounting for visual word recognition speed as measured by lexical decision times (Forster & Chambers, 1973; Monsell, 1991) and eye fixation durations during reading (Rayner, 2009). Unsurprisingly, reading behavior as measured by, for example, lexical decision, naming, fixation times, and so on is affected by a wide range of other properties of words, including orthographic neighborhood (Carreiras, Perea, & Grainger, 1997; Grainger, 1990), syllable frequency (Carreiras, Alvarez, & de Vega, 1993; Carreiras & Perea, 2004; Perea & Carreiras, 1998), and imageability (James, 1975), to cite just a few examples. Similarly, with regard to other fields that employ linguistic stimuli, such as memory research, it has been shown that word frequency plays a role in short-term memory (Hulme et al., 1997) and syllable length in working memory (Gathercole & Baddeley, 1990).

Given the wide range of word properties that can affect language and cognitive processing, it is desirable to have a single, integrated, and updateable source of data. For Spanish, there are now a variety of databases available, but some are based on a relatively small number of tokens (Davis & Perea, 2005; Sebastián-Gallés, Martí, Carreiras, & Cuetos, 2000; Taulé, Martı, & Recasens, 2008), while others provide information about a limited number of variables (Alonso, Fernandez, & Díez, 2011; Cuetos-Vega, González-Nosti, Barbón-Gutiérrez, & Brysbaert, 2011; Davies, 2005; Marian, Bartolotti, Chabal, & Shook, 2012). EsPal (Español Palabras, meaning simply "Spanish words") is a Web-based repository available at http://www.bcbl.eu/databases/espal/ that has been designed to fill this gap, providing information on a

comprehensive set of word properties from corpora with hundreds of millions of words.

The most similar effort is the Syllabarium (Duñabeitia, Cholin, Corral, Perea, & Carreiras, 2010), which is a Web-based tool accessing a database containing information on word frequencies and syllable frequencies by token and syllable position. Standalone software packages are also available for Spanish and other languages that provide subsets of the properties in EsPal (Davis, 2005; Davis & Perea, 2005; New, Pallier, Brysbaert, & Ferrand, 2004; Perea et al., 2006). However, given the size of the corpora (discussed below), some of the calculations for some of the properties take up to a week on a standard PC, so a precomputed set of properties is preferred. With EsPal, the back-end processing for the word and subword properties is conducted using a multistep program written in Java, which precomputes not only basic properties of word frequency and form, but also orthographic structure and neighborhoods, phonological structure and neighborhoods, lemma and part-of-speech properties, and subword structure properties related to letter bigrams and trigrams, bisyllables, and biphones. In addition, other data such as a word's subjective ratings (e.g., familiarity, imageability, etc.) can be easily attached to the data and made searchable.

The second important factor of EsPal is the capacity to apply the exact same processing to different corpora. A number of studies have shown that, across many languages, word frequencies derived from movie subtitle corpora provide a better account for various psycholinguistic effects (Brysbaert, New, & Keuleers, 2012; Cai & Brysbaert, 2010; Cuetos-Vega et al., 2011; Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010; Keuleers, Brysbaert, & New, 2010; New, Brysbaert, Veronis, & Pallier, 2007). However, properties from written corpora have in the past been more common and may better predict some phenomena, so it is useful to have different sources of data available for researchers, depending on their goals. EsPal currently fulfills this goal by applying the same processing to both a corpus based on movie subtitles and one based on written text (fiction, nonfiction, and Web pages).

Finally, the Spanish-speaking community is diverse, and EsPal is constructed to be able to accommodate this diversity, at least in terms of phonological representation. Standard Castilian Spanish spoken on mainland Spain differs in a number of dimensions from the Spanish spoken in the Canary Islands and in Latin America (which itself is quite diverse). EsPal therefore also allows the user to choose which phonological representation is used, for example, to derive properties related to phonological neighborhoods.

In the remainder of this article, we describe the collection and preprocessing of the written and subtitle databases currently available in EsPal; how we calculate orthographic and phonological properties, subword properties, lemma and part-of-speech properties; and the source of the subjective ratings data.

Written corpus collection and preprocessing

Written corpus collection

The EsPal Written Corpus is derived from a wide selection of texts collected from the Web or available in digital format. Table 1 provides a listing of percentages in terms of word tokens across the different sources and genres. We grouped them into nine subsets according to their content: academic, culture, law, philosophy, literature, news, politics, society, and the Spanish Wikipedia. All these texts had to meet the requirements of being freely available and not subject to copyright. Most documents were gathered from websites featuring a variety of linguistic styles, including formal, colloquial, and specialized language.

The academic texts are mainly Ph.D. theses selected from a wide range of scientific fields: anthropology, architecture, art, biology, law, economics, electronics, philology, philosophy, physics, history, humanities, engineering, mathematics, medicine, psychology, chemistry, telecommunications, and veterinary science. The set of culture texts is composed of news about cultural events from several newspapers and blogs of opinion about films. Legal texts include mainly rulings by the High Court of Justice of several autonomous regions in Spain, as well as news from the judiciary field as it appeared in popular newspapers (*El Mundo*, *El País*, and *El Periódico*). The literary texts come from several websites containing works with expired copyrights (*bdigital*, *biblioteca_ignoria*, *libroteca*, *logos*, and *scribd*). These works are both texts written in Spanish and translations into Spanish. The news is from the EFE Agency from January, February, and March 2000. The politics set contains news texts referring to Spain's 2007 autonomic elections, speeches by the Spanish President during 2008, and documents taken from political party websites. The society set is composed of Web texts about religion, abortion, and psychology. Finally, the Web data are from the whole Spanish Wikipedia, circa February 2009.

Table 1 Percentage of terms by source type in the EsPal written corpus

Source type	Percent of terms
Academics	1.8 %
Culture	0.2 %
Law	1.0 %
Philosophy	1.1 %
Literature	22.5 %
News	8.7 %
Politics	16.0 %
Society	4.7 %
Web/Wikipedia	43.9 %

The whole corpus underwent a process of cleaning to eliminate the metadata usually present in these types of texts. This process was both automatic and manual and was extremely time consuming.

Written corpus preprocessing

Before the data were incorporated into EsPal, all the text was first parsed using the FreeLing part-of-speech tagger (Padró, Collado, Reese, Lloberes, & Castellón, 2010) to output into a file one term per line with its lemma and its part of speech. The parsing resulted in a total of 309,530,600 terms (no punctuation was included). A "term" could be one or more words and included dates (*17 de julio de 1990* ["July 17, 1990"]), proper nouns *(Congreso de los Estados Unidos* ["United States Congress"]), or phrases (*por ejemplo* ["for example"]). These terms were then imported into a `raw_sequence` table in EsPal, with one word per row (i.e., multiword terms were separated) and columns for the lemma and the part-of-speech tag (e.g., http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html). If the word came from a multiword term, then the word itself was used as the lemma. In this manner, the part-of-speech tag is maintained for the word within its larger context—for example, *de* ["of"] will have lemma statistics as a date and a proper noun (among others) in addition to being a preposition. In the lemma processing section below, we describe further lemma information available for words. The word and lemma were changed to all lowercase using the Java string function toLowerCase with the "es" locale. This table had a total of 325,773,444 rows. Subsequent processing of the contents of the `raw_sequence` table is described later.

Subtitle corpus collection and pre-processing

Subtitle corpus collection

A total of 100,659 Spanish subtitle files were originally provided by the www.opensubtitles.org website including metadata about the file (such as author and total downloads). The Internet Movie Database (IMDb) ID was also supplied, by which genre, director, and cast information can be obtained. Subtitle file formats contain an index number, the start and stop time for which the subtitle is to be shown on screen in milliseconds, and the text of the subtitle, all of which were stored in the `subtitles` table of the database. Movies account for 65.6 % of the files, with the remainder from television episodes. A given show can be labeled with more than one genre, so the words in a subtitle file can be double counted, but across all such counts, by genre, 22.0 % of the words are from dramas, 10.9 % from comedies, 10.3 % from thrillers, 7.7 % from crime shows, 7.4 % from action shows, 7.3 % from romances, 5.8 % from mysteries, and 5.5 % from adventure shows, and the remaining are in 13 other genres, accounting for less than 5 % each. Similarly, the source show can contain more than one language, and across such counts, 52.6 % of the words are from English language shows, followed by French (8.5 %) and Spanish (5.5 %). No limits were put on the date of the source, since the subtitles themselves, uploaded by users of the website, are of recent origin. However, given the metadata maintained about the source of the words, a variety of subcorpora are possible whose properties might be more appropriate depending on the psycholinguistic question being asked.

Subtitle corpus preprocessing

For a proper parsing of text, complete sentences are needed. However, a single subtitle instance could have two speakers (usually denoted by a dash ["–"] at the beginning of each of their statements), or a single speaker's statement could continue into the next subtitle instance (usually denoted by ellipses ["…"] at then end). Therefore, a second stage of processing was run to fill a `statements` table with strings that were, at a first approximation, single statements (which could contain multiple sentences). At this stage, subtitles were removed that contained metadata (such as the author of the subtitles or translations of the credits); all HTML markings were removed; and contents within brackets (often indicating sounds) were also removed.

Each statement was submitted individually to FreeLing (Padró et al., 2010) for part-of-speech tagging and lemmatization. In this case, the lowercased word, lowercased lemma, and part-of-speech tag were stored directly in the `raw_sequence` table, along with the file ID, IMDb movie ID, statement index, and within-statement index. Thus, the provenance, or origin, of every word can be traced back, enabling further analyses, which we will be reporting in the future. In the end, words from 98,339 distinct files and 40,444 unique movies are present in this table.

Word selection and frequency processing

The `raw_sequence` table holds every individual word token from the source. The count of each unique word type is accumulated in a second table, `raw_words`. Every word type in this table is checked against the criteria below. Those that do not pass the criteria are marked as rejected. The word had to appear in at least one of these publicly available sets of Spanish words: OpenOffice,[1] AGME,[2] or SemEval.[3] For future comparison, we also allowed words present in other

[1] http://wiki.services.openoffice.org/wiki/Dictionaries.
[2] http://www.cic.ipn.mx/~sidorov/agme/.
[3] http://www.lsi.upc.edu/~nlp/semeval/msacs_download.html.

recent Spanish corpora projects (Alonso et al., 2011; Cuetos-Vega et al., 2011). In addition, we included a large number of Spanish first names, surnames, and place names from publicly available websites. Rejection criteria were that words could not be longer than 30 characters,[4] contain a nonletter (which excluded hyphenated words), have more than 3 characters in a row of the same character, nor contain non-Spanish characters that is, outside of a–z, áéíóúñü. Words that passed these filters were placed into the `word_data` table with their counts.[5] Table 2 contains the final counts of word types and word tokens for the two corpora.

The `word_data` table contains all the information about each word and, thus, what can be searched for simultaneously via the Web interface. We will be presenting the various properties available for each word with its column name in bold italics. For each ***word***, we store the count (***cnt***), the frequency per million (***frq***), log10(***cnt***+1) (***log_cnt***), and log10(***frq*** + 1/*N*), where *N* = millions of words in the database (***log_frqN***), which has been shown to be a fruitful way to compare frequencies across corpora (Brysbaert et al., 2011).

Subtitle corpus contextual diversity processing

Recent work has found that the number of different contexts in which a word occurs can be more informative than the token frequency (Adelman, Brown, & Quesada, 2006; Brysbaert & New, 2009; Dimitropoulou et al., 2010; Keuleers et al., 2010; Perea, Soares, & Comesaña, 2013). The original EsPal subtitles database described above uses all the files available, so some shows are multiply represented. Therefore, EsPal provides a third database of properties (subtitles_cdm) that are based on the number of different movies (IMDb IDs) that the word appears in. In this database, ***cnt*** refers to the count of different movies, and ***frq*** is equal to the percent of movies (i.e., 100 * ***cnt***/40,444). We also explored using the count of different subtitle files, with the expectation that this would have some relationship to popularity (e.g., there are almost 300 versions of *Lord of the Rings: Return of the King*) and, therefore, provide word frequencies that were better predictors of certain psycholinguistic variables. However, in all the cases we have explored to date, the contextual diversity based on the number of movies has given slightly better results.

[4] A cutoff was made for processing and memory considerations. Out of over 460 million tokens in the raw subtitle data set, only 735 tokens have a length greater than 30.

[5] The system is designed such that at this stage, it would also have been possible to further reduce the words by removing accents or tildes and collapsing the counts across the subsequent word forms. Some psycholinguistic research questions, such as studies focused on stress assignment (e.g., Shelton, Gerfen, & Gutiérrez-Palma, 2011), might benefit from this type of frequency data. However, the first version of these sources has the actual form of the word.

Table 2 Counts of word types and word tokens in each corpus

Corpus	Word types	Word tokens
Written	277,771	307,772,547
Subtitles	244,983	462,611,693

Orthographic properties processing

Orthographic structure

The basics of the orthographic structure are, of course, present in the ***word*** column itself. In addition, the number of letters (***num_letters***) and whether or not there are repeated letters (***rep_letters***) within the word (0 = false, and 1 = true) are stored. A straightforward consonant–vowel structure (***orth_cv_structure***) was also created by replacing each vowel character (*a*,*e*,*i*,*o*,*u*, with or without accents, but not y) with "V" and all other characters with "C." Note, however, that there are certain limitations to this simple heuristic, especially with regard to the letters *y* and *h*.

Orthographic neighborhoods

Orthographic neighborhood size affects a large number of psycholinguistic phenomena (Carreiras et al., 1997; Davis, Perea, & Acha, 2009; Grainger, 1990; Yarkoni, Balota, & Yap, 2008). For EsPal, each word was compared with all other words in the same source in order to provide an array of neighborhood properties. For single-change substitution, addition, deletion, and transpose letter neighbors, data are provided such as the list of neighbors and the frequency of the highest frequency neighbor. The average edit distance (Levenshtein distance) of the 20 closest words (no matter how far) is also provided (***Lev_N***). Another way to compare a word with all the others is the character in the word at which it is no longer like any other word (***orth_uniq_point***), which is a factor in reading studies (e.g., Miller, Juhasz, & Rayner, 2006) . If the word is completely unique, a secondary orthographic uniqueness point (***orth_sec_uniq_point***) is determined as well in case the uniqueness is simply due to, for example, the plural form of the word. Table 3 contains all of the measures available concerning orthographic neighborhoods.

Phonological properties processing

Phonological structure

Spanish is a relatively transparent language, so syllable and phonological structure can be derived from the orthography in a rule-based fashion. To derive the syllable structure, we

Table 3 Orthographic neighborhood variable names and meanings

Variable name	Variable meaning
N	Number of substitution neighbors
NHF	Number of higher frequency substitution neighbors
frq_hf_s	Frequency of the highest frequency substitution neighbor
hf_s	Highest frequency substitution neighbor
hf_s_list	List of substitution neighbors in descending frequency, with the place of the word itself marked by "OOOOOO"
P	Number of positions with substitution neighbors
PHF	Number of positions with higher frequency substitution neighbors
avg_frq_Ns	Average frequency of substitution neighbors
N_TL	Number of transposed-letter neighbors
frq_hf_tl	Frequency of the highest frequency transposed-letter neighbor
hf_tl	Highest frequency transposed-letter neighbor
hf_tl_list	List of transposed-letter neighbors in descending frequency, with the place of the word itself marked by "OOOOOO"
N_A	Number of addition-letter neighbors
frq_hf_A	Frequency of the highest frequency addition-letter neighbor
hf_A	Highest frequency addition-letter neighbor
hf_A_list	List of addition-letter neighbors in descending frequency, with the place of the word itself marked by "OOOOOO"
N_D	Number of deletion-letter neighbors
frq_hf_D	Frequency of the highest frequency deletion-letter neighbor
hf_D	Highest frequency deletion-letter neighbor
hf_D_list	List of substitution neighbors in descending frequency, with the place of the word itself marked by "OOOOOO"
orth_uniq_point	The character in the word at which it is no longer like any other word. Set to 0 if it is subsumed by some other word and not unique
orth_sec_uniq_point	If the word is unique, then the uniqueness point with the last letter removed
Lev_N	Average Levenshtein distance of the 20 closest words (OLD20)

implemented, with some minor changes, the rules in Silabeador TIP (Hernández-Figueroa, Rodríguez-Rodríguez, & Carreras-Riudavets, 2009) to obtain orthographic syllable boundaries (***orth_syll_structure***). The most notable change was the addition of the onset, nucleus, and coda information being stored for each character. From this information, the number of syllables (***num_syll***) and the position of the syllable with the accent was also derived (***syll_accent***).

The phonetic transcription of the word (***phon_structure***) was derived using a Java implementation of the rules in the SAGA project (Nogueiras & Mariño, 2009) taking advantage, when necessary, of the syllabification described above. For example, the letter *t* is phonetically transcribed as *t* (*toro* ["bull"] → *toro*), except when it is syllable final (*etnia* ["ethnicity"] → *eDnja*). The codes were modified to be a single character and are shown in Table 4. From this information, the number of phonemes (***num_phon***), the initial phoneme (***init_phon***), and the phonetically based CV structure (***phon_cv_structure***) were derived.[6]

[6] Note that exceptions to the rules have not been implemented, and we are investigating other methods by which to derive phonetic transcriptions.

Two phonetic representations were derived, one for Castilian Spanish and one for Latin American Spanish. Although this is a complex topic and pronunciation varies dramatically within and between countries (Moreno & Mariño, 1998), for this introduction of EsPal, the only difference between these two representations are that *z* and *c* (followed by *e* or *i*) are transcribed as *T* in Castilian and *s* in Latin American Spanish. However, the software and website are capable of accommodating any number of phonetic representations, and more accurate representations can be added over time. In the database and website output, these columns and the neighborhood columns described below are prepended by either ***es*** or ***sa*** for Castilian and Latin American Spanish, respectively, depending on which representation is chosen.

Phonological neighborhoods

With a single-character representation of the phonemes of each word, we can use exactly the same neighborhood processing as was used for the orthographic neighborhoods. However, in the spoken word recognition literature, slightly different variables are typically investigated, so the properties provided are different from those for the orthographic

Table 4 Phonetic transcription codes used in EsPal

SAGA code	EsPal code	Sound
p	p	voiceless bilabial plosive
b	b	voiced bilabial plosive
t	t	voiceless dental plosive
d	d	voiced dental plosive
k	k	voiceless velar plosive
g	g	voiced velar plosive
m	m	voiced bilabial nasal
n	n	voiced alveolar nasal
N	N	voiced velar nasal (preceding a velar consonant)
J	J	voiced palatal nasal
tS	C	voiceless palatal affricate
f	f	voiceless labiodental fricative
T	T	voiceless interdental fricative
s	s	voiceless alveolar fricative
z	z	voiced alveolar fricative (preceding a voiced consonant)
jj	H	voiced palatal fricative
x	x	voiceless velar fricative
l	l	voiced alveolar lateral
L	L	voiced lateral palatal
rr	R	voiced alveolar trill
j	j	palatal semivowel
w	w	labiovelar semivowel
B	B	voiced bilabial approximant
D	D	voiced dental approximant
G	G	voiced velar approximant
r	r	simple vibrating voiced alveolar
a	a	open central vowel
e	e	front half vowel
i	i	front closed vowel
o	o	half rounded back vowel
u	u	closed rounded back vowel

neighborhoods. Table 5 contains a listing of those phonological neighborhood variables currently available.

Subword processing

Infralexical, or subword, features are known to influence lexical decision and naming times (Carreiras et al., 1993; Carreiras & Perea, 2004). The processing was very similar for bigrams, trigrams, biphones, and bisyllables, but for exposition we will describe only bigram processing. A new table `bigram_raw` is created to hold for each bigram–word-position combination the sum of word token frequencies (***frq***) and word type counts from the `word_data` table. For instance, when the word *casa* ["house"] is encountered, it is found to contain three bigrams (*ca*, *as*, *sa*) with positions 1, 2, and 3, respectively.[7] An entry is made in the `bigram_raw` table for each of these bigrams at their positions, and the frequency per million (***frq***) of *caso* is added to the token frequency column and 1 is added to the type count column. When the word *caso* ["case"] is encountered, *ca* at position 1 and *as* at position 2 have their token frequency and type count columns incremented by the frequency per million of *caso* and 1, respectively; and a new entry for *so* is made at position 3.

After information from all the words was added to the `bigram_raw` table, each word was reanalyzed to obtain properties of its bigrams. For example, across the entire word *casa*, we can sum or average, in terms of token frequency or type count, its three bigram frequencies. These sums and averages can also either respect the position of the bigram or not (e.g., *ca* at position 1 vs. at any position). Thus, there are eight bigram values that are available for each word as a whole.

For a given word, EsPal also provides each bigram's token frequency and type count, either for the bigram in that position only or for the bigram in that position found anywhere in a word. So *caso* has three nonzero bigram data sets, and the first data set has the token frequency and type count of *ca* at position 1 and of *ca* at any position. Bigram and trigram data are calculated for words with up to 20 characters. Similar processing is done for biphones on the basis of the phonetic structure (***phon_structure***) up to 20 phonemes and for bisyllables on the basis of the individual syllables in the orthographic syllable structure (***orth_syll_structure***) up to eight syllables.

To provide this large amount of infralexical information, we created a systematic method for deriving property names. Property name affixes are added for each n-gram length (bigram [***B***] or trigram [***T***]), and for each n-gram modality (orthographic [***O***], phonemic [***P***], syllabic [***S***]). So, bigram = ***BO***; trigram = ***TO***; biphone = ***BP***; and bisyllable = ***BS***. The system is designed to be extensible, so any other combination of interest could be added. Currently, the frequency per million (***frq***) is used and denoted by ***F*** in the variable name, but the count (***cnt***) could also be used, as well as the log of either. We can add such versions of the calculations as they are requested. Eight variables are made for each length–type combination. These have combinations that are position sensitive (***pos_***) or independent (***abs_***) sums (***S***) or means (***M***) of the token frequency (***tok_***) or type count (***type_***). The previous code is then appended to the property name. For example, the position-independent mean of biphone token frequencies is ***abs_tok_MBPF***.

[7] Note that it is common to have markers for the beginning and end of words as well; for example, *casa* would also produce the bigrams *_c* and *a_* and the trigrams *_ca* and *sa_*. This information will be available in a subsequent version of the database.

Table 5 Phonological neighborhood variable names and meanings

Variable name	Variable meaning
NP	Number of phonological neighbors (all kinds)
NPHF	Number of higher frequency phonological neighbors
frq_hfp	Frequency of the highest frequency phonological neighbor
hfp	Phonological neighbor with the highest frequency
hfp_list	List of phonological neighbors in descending frequency, with the place of the word itself marked by "OOOOOO"
pf	Number of phonemes/positions with phonological neighbors
pf_hf	Number of phonemes/positions with higher frequency phonological neighbors
avg_frq_Np	Average frequency of phonological neighbors
phon_uniq_point	Phoneme position at which it is no longer like any other word. Set to 0 if it is subsumed by some other word and not unique
homoph	Number of other word entries with the same ***phon_structure***
homoph_list	List of homophones in descending frequency

Lemma and part-of-speech processing

While word-form frequencies have tended to dominate analyses, the lemma and part-of-speech frequencies may also influence behavior (Baayen, Dijkstra, & Schreuder, 1997; Taft, 1979). To set the values for the lemma and part-of-speech properties, we return to the `raw_sequence` table. Counts were made of every unique combination of word, lemma, and part-of-speech tag, rejecting combinations where the lemma contains non-Spanish characters or is too long (> 255 characters). For the written database, there were 388,270 word–lemma–code types, and for the subtitles database, there were 404,394 word–lemma–code types. Since there was more than one row per word, these data were stored in a separate `lemma_data` table for searching (cf. Brysbaert et al., 2012).

For each word, EsPal gives the percentage of occurrences with each lemma–code combination. For example, the word *caso* most often appears as a common masculine singular noun ["case"] but can also appear as a conjunction (*caso de que* ["if"]), an adverb (*en todo caso* ["in any case"]), a preposition (*en caso de* ["in case of"]), a verb (*yo me caso* ["I marry"]), as well as a proper noun and URL. Similarly, for each lemma, EsPal gives the percentage of occurrences with each word–code combination. For example, the lemma *caso*, besides occurring with the previous parts of speech, also occurs with the masculine plural noun *casos*. The variable ***percent_word*** gives the percentage of each word (by ***_type*** or ***_tok***) that has that word–lemma–code, and ***percent_lemma*** gives the percentage of each lemma (by ***_type*** or ***_tok***) that has that word–lemma–code. For example, for the word–lemma–code combinations with *caso* as the word, ***percent_word_type*** = 16.76 % in the written database, since *caso* appears with six different lemma–code combinations, and the ***percent_word_tok*** for the masculine singular noun lemma–code = 81.5 %, and for the simple preposition = 5.6 %.

Fig. 1 Screenshot of the choice of database and phonology from the EsPal website

The part-of-speech tags are also expanded to allow searching and organization of results. The part-of-speech information includes ***Category***, ***Type***, ***Degree***, ***Appreciative***, ***Diminutive***, ***Person***, ***Mode***, ***Tense***, ***Form***, ***Gender***, ***Number***, ***Function***, ***Possessor***, and ***Politeness***. A full list for Spanish can be found on the FreeLing website,[8] which shows, for example, how the different attributes of an adjective are specified.

Some of the lemma information is also added to the `word_data` table—namely, information about the most common part of speech associated with the word (the "maximum lemma") and the "lemma frequency" of the word, which is based on the sum of the counts of all the words that have the same lemma as any of the lemmas of the word (Keuleers et al., 2010). For the maximum lemma of a word, EsPal provides the lemma itself (***max_lem_lemma***), the detailed part-of-speech code (***max_lem_code***), and the percentage of all the word's tokens with that code (***max_lem_perc***), the category (***max_lem_cat***), and the percentage as that category (***max_lem_cat_sum_perc***). So for example, in the subtitles database, the word *caso* mentioned above appears 90.15 % as a common masculine singular noun and 90.55 % as a noun overall (the additional appearances probably labeled as a

[8] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html.

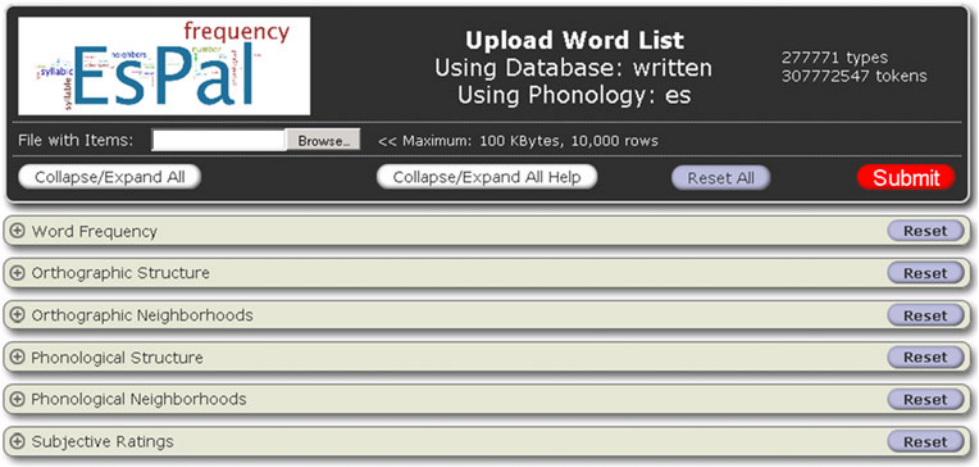


Fig. 2 Screenshot of the Word to Properties page where one can upload a list of words to receive the properties of the types shown

proper noun). For the lemma frequencies, EsPal makes available the log(***cnt*** +1), as well as the $\log^2$(***cnt*** + 1), which Keuleers et al. (2010) found helped account for more variance in lexical decision times in Dutch.

Subjective ratings

Subjective ratings, such as the imageability of the thing that a word refers to, also modulate the process of lexical access

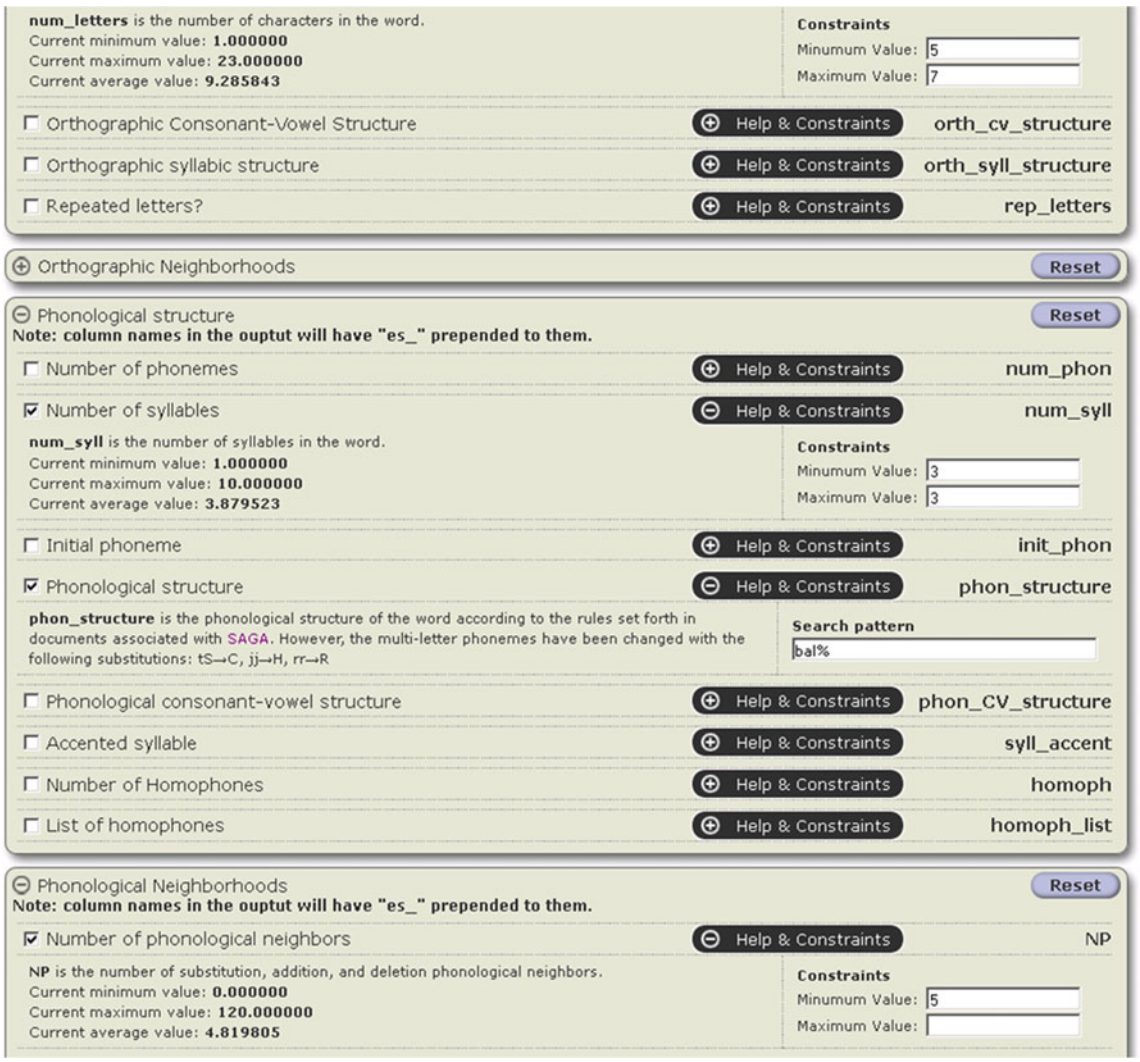
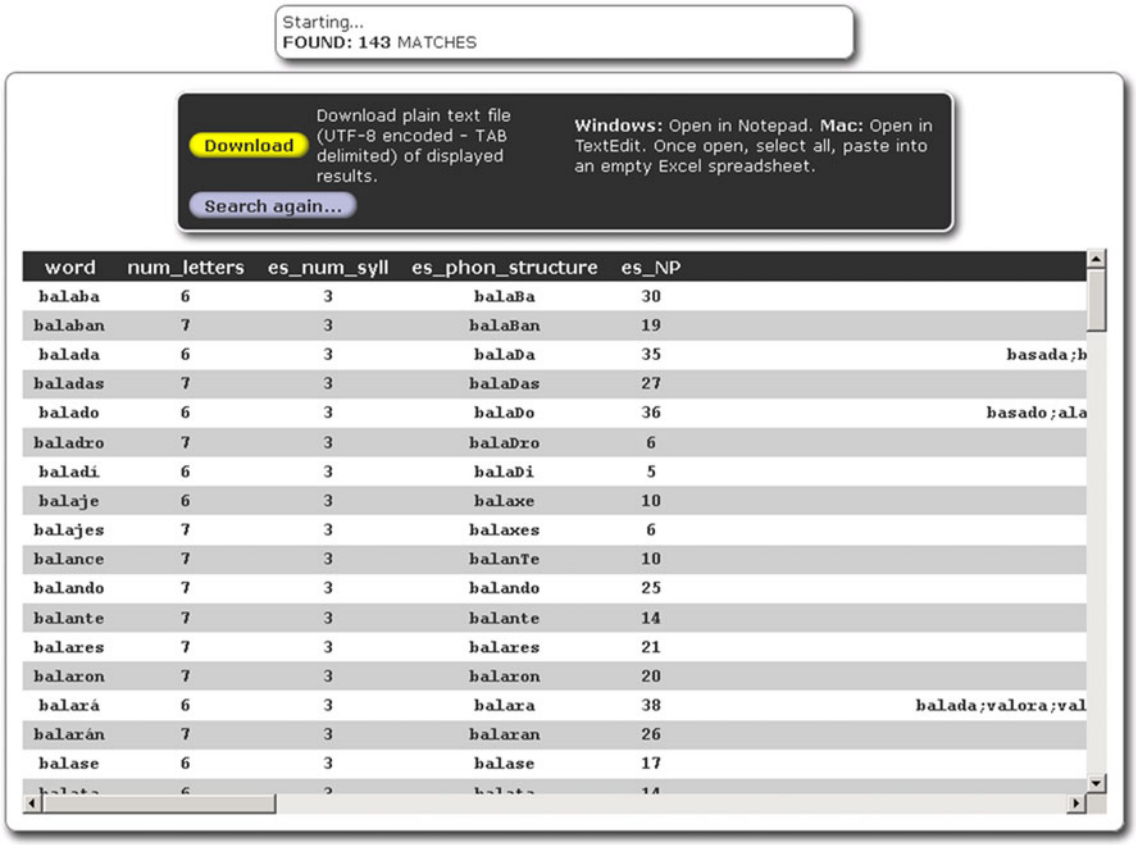


Fig. 3 Screenshot of the Constraints to Words page where a variety of constraints have been applied



word	num_letters	es_num_syll	es_phon_structure	es_NP	
balaba	6	3	balaBa	30	
balaban	7	3	balaBan	19	
balada	6	3	balaDa	35	basada;b
baladas	7	3	balaDas	27	
balado	6	3	balaDo	36	basado;ala
baladro	7	3	balaDro	6	
baladi	6	3	balaDi	5	
balaje	6	3	balaxe	10	
balajes	7	3	balaxes	6	
balance	7	3	balanTe	10	
balando	7	3	balando	25	
balante	7	3	balante	14	
balares	7	3	balares	21	
balaron	7	3	balaron	20	
balará	6	3	balara	38	balada;valora;val
balarán	7	3	balaran	26	
balase	6	3	balase	17	

Fig. 4 Screenshot of the results of the query from Fig. 3. All checked properties are returned, which here include ***num_letters, es_num_syll, es_phon_structure, es_NP*** (number of phonological neighbors), and ***es_hfp_list*** (list of higher frequency phonological neighbors)

(Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004). For EsPal, 6,500 words were selected (mostly nouns and verbs, although some nouns could be considered also adjectives). The words corresponded to those with the highest frequencies in the Alameda and Cuetos (1995) and the Juilland and Chang-Rodríguez (1964) word frequency lists. Nouns with gender (e.g., *niña*, *niño*) and number (e.g., *corte* and *cortes*) inflections were generally both included for evaluation. We decided to include both since, in many cases, the different usages with the two gender forms or the two number forms hold different semantic features. For instance, the word *corte* suggests more clearly the action of cutting than does the plural form *cortes*. In addition, each form involves different semantic meanings: *Cortes* is a term that can be used to refer to the parliament of Spain (*Cortes Generales*), while *corte* is linked more to the royalty. On the other hand, some nouns could also be considered adjectives; for example, the word *rojo* ["red"] can refer to the color itself (as well as a Communist) or be used as an adjective. Finally, we have included nonreflexive and reflexive verbal forms when the two are common, such as *aplicar* ["to apply/attach"] and *aplicarse* ["to apply oneself/work hard"], because there are important semantic differences between them.

From the 6,500 words, we created 130 questionnaires of 100 words each. This way, each word appeared in a different position in two questionnaires and was embedded in a different context of other words. Then we created three forms for each of the 130 questionnaires, so that each word was evaluated on a scale of 1–7 for three different values: concreteness, familiarity, and imageability. Subjective ratings were obtained in two different time windows. The first wave was obtained in 1998–1999 and corresponds to the data appearing in LEXESP (Sebastian-Gallés et al., 2000). The questionnaires were answered by undergraduates from 12 different Spanish universities, including Universitat Autònoma de Barcelona, Universidad Autónoma de Madrid, Universitat de Barcelona, Universidad Complutense de Madrid, Universidad de Granada, Universidad de Oviedo, Universidad de La Laguna, Universitat Rovira i Virgili, Universitat de València, Universidad de Santiago de Compostela, Universidad de Málaga, and Universidad de Salamanca. Due to the random sampling, not all words were equally evaluated, and around 2,000 words in each dimension did not reach the minimum of 30 responses. In a second wave (taking place between 2007 and 2009), an additional set of undergraduate students from the Universitat de Barcelona and Universidad de La Laguna answered new questionnaires so that a minimum of 30 responses for each word were finally reached. The data present in EsPal are the

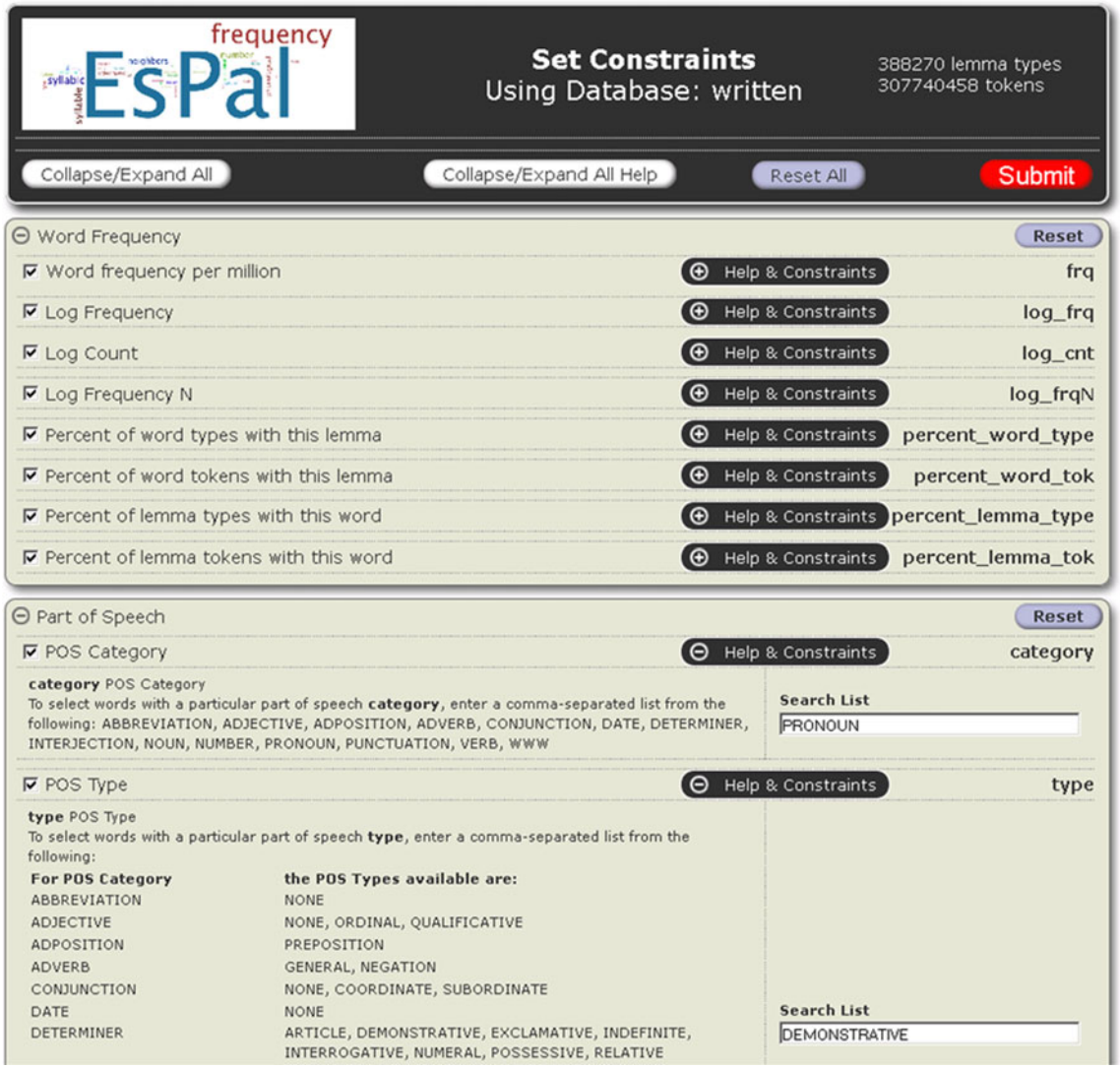


Fig. 5 Screenshot of the POS Constraints to Words page from which frequency properties will be returned for all demonstrative pronouns

average ratings for over 6,400 words from at least 30 participants and from at least 2 universities.

EsPal website

The EsPal website can be accessed at http://www.bcbl.eu/databases/espal/. When the user first goes to the website (Fig. 1), the user must first choose a source database and phonology via the radio buttons. There are then four ways to obtain information from EsPal. The user can upload a file of words, one word per line, to receive chosen properties on those words, or the user can set constraints on the properties to receive a list of words having those constraints. These two actions can be performed on data in either the `word_data` table or the `lemma_data` table.

Table 6 Frequency correlations: Correlations of frequency—that is, log(count + 1)—between different corpora (number of common words)

	EsPal-written	EsPal-subtitles	EsPal-subtitles CDM	LEXESP (B-Pal)	Oral frequency
EsPal-subtitles	.693 (193,757)				
EsPal-subtitles CDM	.713 (193,757)	.977 (244,947)			
LEXESP (B-Pal)	.855 (30,277)	.794 (28,727)	.799 (28,727)		
Oral frequency	.700 (65,388)	.655 (62,271)	.649 (62,271)	.827 (18,723)	
SUBTLEX-ESP	.663 (88,303)	.938 (93,949)	.936 (93,949)	.777 (20,316)	.725 (44,374)

For example, clicking "Words to Properties" brings the user to a Web page (Fig. 2) where he or she can upload a file of words, one word per line, and within each of the subpanels, choose which properties to receive. Clicking "Submit" brings the user to the results page, which contains a table of the results, as well as a button to download a file containing the results (returned in the order of the original file). Instructions on the page specify how best to convert the downloaded file into a spreadsheet program.

Clicking "Constraints to Words" from the EsPal homepage allows the user to set constraints for returned words. The example in Fig. 3 shows how the user might search for words with five to seven letters and three syllables that start with the phonemes "bal" and have at least five phonological neighbors. In the written database, this returns 143 words (Fig. 4).

Clicking "Words to Lemma and POS Properties" from the EsPal homepage allows the user to receive lemma and part-of-speech information for a list of words. Starting with "POS Constraints to Words" the user can request, for example, the frequencies of all demonstrative pronouns (Fig. 5).

Index comparisons and validity

While the main purpose of this article is to describe the source of the word frequency data and how it has been processed and made available, readers may wish to note how it compares to other corpora with regard to the psycholinguistic data mentioned in the introduction. We compare the three EsPal corpora with three other Spanish data sources: LEXESP (Sebastián-Gallés et al., 2000), although in the form of B-Pal (Davis & Perea, 2005), which did extensive cleaning of the data; SUBTLEX-ESP (Cuetos-Vega et al., 2011), which also used subtitles (although from different online sources); and oral frequency data from Alonso et al. (2011). Table 6 shows the overall frequency correlations between each of these sources based on the number of words they have in common, although better means may be available for such comparisons (Brysbaert & Diependaele, 2012). As one would expect, the written databases (EsPal-Written and LEXESP [B-Pal]) are most similar to each other, and the subtitle databases are most similar to each other, with the oral frequency data somewhere in between.

Given that the lexical decision times used in the SUBTLEX-ESP paper are not yet available and our own are still forthcoming, we provide some basic comparisons with two other data sets currently available: word-naming times (Cuetos & Barbón, 2006) and picture-naming times (Cuetos, Ellis, & Alvarez, 1999), which are shown in Tables 7 and 8, respectively, along with word length as an added factor in the multiple regression, as previous authors have done. Among the EsPal corpora, the subtitles CDM database performs best and reinforces previous findings in other languages. The EsPal subtitle data sets (both by token and CDM) are very similar to the SUBTLEX-ESP and oral frequency data sets with respect to word-naming times and account for slightly more variance than SUBTLEX-ESP with respect to the picture-naming times.

Table 7 Word naming: Regression analysis results using word length and the frequency, log(count + 1), from different corpora on word naming times (Cuetos & Barbón, 2006)

Factors	Weights	Adjusted R^2
LEXESP (BPAL)	−8.464 *	.302
Length	10.786 ***	(N = 240)
Oral frequency	−8.263 **	.301
Length	10.774 ***	(N = 235)
SUBTLEX-ESP	−8.353 **	.312
Length	10.390 ***	(N = 239)
EsPal-written	−5.870 †	.298
Length	10.700 ***	(N = 240)
EsPal-subtitle tokens	−7.430 *	.304
Length	10.604 ***	(N = 240)
EsPal-subtitle CDM	−9.074 *	.305
Length	10.611 ***	(N = 240)

Note. All adjusted R^2 have ps < .001.

† $p < .1$

* $p < .05$

** $p < .01$

*** $p < .001$

Table 8 Picture naming: Regression analysis results using word length and the frequency, log(count + 1), from different corpora on picture naming times (Cuetos, Ellis, & Alvarez, 1999)

Factors	Weights	Adjusted R^2
LEXESP (BPAL)	−61.187 ***	.161
Length	9.634 †	(N = 139)
Oral frequency	−65.848 ***	.188
Length	7.446	(N = 137)
SUBTLEX-ESP	−44.897 ***	.118
Length	12.244 *	(N = 138)
EsPal-written	−39.378 **	.100
Length	11.748 *	(N = 139)
EsPal-subtitle tokens	−46.61 ***	.123
Length	11.84 *	(N = 139)
EsPal-subtitle CDM	−59.008 ***	.133
Length	11.050 †	(N = 139)

Note. All adjusted R^2 s have ps < .001.

† $p < .1$

* $p < .05$

** $p < .01$

*** $p < .001$

EsPal currently provides the properties of two data sources, one written and one based on subtitles, with additional information based on the contextual diversity (by movie) of the subtitles data. We provide initial evidence that these data sources, the latter especially, are comparable to other corpora in Spanish in terms of their frequency data helping to predict some psycholinguistic phenomena. We should note, however, that there are some limitations that researchers should keep in mind when using the data contained in EsPal, especially the subtitle data. These data are based on a large number of amateur translations of media that are most often English, not Spanish, in source, and since proper nouns are typically not translated (e.g., "John" is not renamed "Juan"), such terms will appear with some frequency. We have used publicly available lists of "Spanish words" in order to restrict what is inserted into our databases, as well as allow comparison with other experimental data. Even so, when using EsPal to generate Spanish words for an experiment, one should have a native speaker, from the same culture as the subjects, cull out these perhaps undesirable elements. Nevertheless, our initial validation results suggest that despite what pollution may occur because of these foreign words, the frequencies given for the "true" Spanish words are useful.

Conclusion

EsPal is a free online application that makes available a wide range of frequency, orthographic, phonological, and subjective information about Spanish words. EsPal provides an extensible, ever-improving, and accurate set of data sources and analyses. Initial testing of the current data indicates that they are at least comparable to extant sources. This system may, therefore, assist the research communities of many disciplines to accelerate selection of stimuli for their experiments and thereby increase the rate of scientific progress.

Acknowledgments We would like to thank Daniel Diaz for his technical help during the initial phases of the project. Our reviewers have been extremely helpful as well. This work was partially funded by a grant, HUM2007–30271–E/FILO, from the Spanish Ministry of Science and Innovation.

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*(9), 814–823.

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

Alonso, M. A., Fernandez, A., & Díez, E. (2011). Oral frequency norms for 67,979 Spanish words. *Behavior Research Methods, 43,* 449–458.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37*(1), 94–117.

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; Journal of Experimental Psychology: General, 133*(2), 283–316.

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*(5), 412.

Brysbaert, M., & Diependaele, K. (2012). Dealing with zero word frequencies: a review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*. doi:10.3758/s13428-012-0270-5

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*(4), 977–990.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*(4), 991–997.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One, 5*(6), e10729. doi:10.1371/journal.pone.0010729

Carreiras, M., Alvarez, C. J., & de Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language, 32*(6), 766–780.

Carreiras, M., & Perea, M. (2004). Naming pseudowords in Spanish: Effects of syllable frequency. *Brain and Language, 90*(1–3), 393–400.

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of the orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 857–871.

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

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

Cuetos-Vega, F., González-Nosti, M., Barbón-Gutiérrez, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica: Revista de Metodología y Psicología Experimental, 32*(2), 133–143.

Davies, M. (2005). The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics, 10*(3), 307–334.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37*(1), 65–70.

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods, 37*(4), 665–671.

Davis, C. J., Perea, M., & Acha, J. (2009). Re (de) fining the orthographic neighborhood: The role of addition and deletion neighbors in lexical decision and reading. *Journal of Experimental Psychology. Human Perception and Performance, 35*(5), 1550–1570.

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).

Duñabeitia, J. A., Cholin, J., Corral, J., Perea, M., & Carreiras, M. (2010). SYLLABARIUM: An online application for deriving complete statistics for Basque and Spanish orthographic syllables. *Behavior Research Methods, 42*(1), 118–125.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 627–635.

Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology, 81*(4), 439–454.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29*(2), 228–244.

Hernández-Figueroa, Z., Rodríguez-Rodríguez, G., & Carreras-Riudavets, F. (2009). *Separador de sílabas del español - Silabeador TIP*. Retrieved from http://tip.dis.ulpgc.es

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1217.

James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology. Human Perception and Performance, 1*(2), 130–136.

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

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE, 7*(8), e43230.

Miller, B., Juhasz, B. J., & Rayner, K. (2006). The orthographic uniqueness point and eye movements during reading. *British Journal of Psychology, 97*(2), 191–216.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148–197). Hillsdale, NJ: Lawrence Erlbaum Associates.

Moreno, A., & Mariño, J. B. (1998). Spanish dialects: Phonetic transcription. *Fifth International Conference on Spoken Language Processing (ICSLP '98)* (pp. 189–192). Sydney, Australia.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied PsychoLinguistics, 28*(4), 661.

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

Nogueiras, A., & Mariño, J. (2009). *SAGA: Transcriptor fonético de las variedades dialectales del español*. Retrieved from http://www.talp.upc.edu/index.php/technology/tools/signal-processing-tools/81-saga

Padró, L., Collado, M., Reese, S., Lloberes, M., & Castellón, I. (2010). Freeling 2.1: Five years of open-source language processing tools. *Proceedings of 7th Language Resources and Evaluation Conference*. La Valletta, Malta.

Perea, M., & Carreiras, M. (1998). Effects of syllable frequency and syllable neighborhood frequency in visual word recognition. *Journal of Experimental Psychology. Human Perception and Performance, 24*(1), 134–144.

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.* doi:10.1016/j.jecp.2012.10.014

Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M. (2006). E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods, 38*(4), 610–615.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506.

Sebastián-Gallés, N., Martí, M., Carreiras, M., & Cuetos, F. (2000). *LEXESP: Léxico Informatizado del Español*. Barcelona: Universitat de Barcelona.

Shelton, M., Gerfen, C., & Gutiérrez-Palma, N. (2011). The interaction of subsyllabic encoding and stress assignment: A new examination of an old problem in Spanish. *Language & Cognitive Processes, 27*(10), 1459–1478.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition, 7*(4), 263–272.

Taulé, M., Martı, M. A., & Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979.