Behavior Research Methods
2005, 37 (4), 665-671

BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish

COLIN J. DAVIS
Macquarie University, Sydney, New South Wales, Australia
and
MANUEL PEREA
Universitat de València, València, Spain

This article describes a Windows program that enables users to obtain a broad range of statistics concerning the properties of word and nonword stimuli in Spanish, including word frequency, syllable frequency, bigram and biphone frequency, orthographic similarity, orthographic and phonological structure, concreteness, familiarity, imageability, valence, arousal, and age-of-acquisition measures. It is designed for use by researchers in psycholinguistics, particularly those concerned with recognition of isolated words. The program computes measures of orthographic similarity online, with respect to either a default vocabulary of 31,491 Spanish words or a vocabulary specified by the user. In addition to providing standard orthographic and phonological neighborhood measures, the program can be used to obtain information about other forms of orthographic similarity, such as transposed-letter similarity and embedded-word similarity. It is available, free of charge, from the following Web site: www.maccs.mq.edu.au/~colin/B-Pal.

A key tool for conducting well-controlled research with linguistic stimuli in a given language is an easy-to-use, comprehensive application for computing psycholinguistic statistics. As such, there are several useful databases for computing psycholinguistic statistics in English (e.g., MRC database, Coltheart, 1981; N-Watch, Davis, 2005) and in French (Lexique, New, Pallier, Brysbaert, & Ferrand, 2004). Here, we present an application for computing a large variety of psycholinguistic statistics for Spanish stimuli. This application may be useful for researchers in cognitive psychology in their monolingual studies in Spanish or, also, in their bilingual studies (e.g., with Spanish–English bilinguals).

Up to the mid-1990s, the most cited reference for Spanish researchers in psycholinguistics was the frequency count compiled by Juilland and Chang-Rodríguez (1964). This word count was based on half a million Spanish words, and the entries included only words with a frequency of at least 10 per million. The obvious limitations of this frequency database led Alameda and Cuetos (1995) to compile a frequency count based on around 2 million Spanish words.[1] As in the case of the Juilland and Chang-Rodríguez count, the Alameda and Cuetos database was not accompanied by an application, and thus researchers needed to use other applications (e.g., spreadsheets or text-processing tools) to select the appropriate stimuli from the original text files containing lexical entries and their corresponding written frequencies. More recently, Sebastián-Gallés, Martí, Cuetos, & Carreiras (2000) compiled LEXESP, a frequency database based on approximately 5 million Spanish words.[2] LEXESP not only extends the existing corpus of the Alameda and Cuetos database, but also provides other relevant indices, such as subjective norms (imageability, concreteness, and familiarity), number of syllables, stress location, and pronunciation (in DISC codes), among others. Furthermore, LEXESP is accompanied by an application (named *Corco*) for accessing the different indices contained in the database. Thus, the LEXESP/Corco database has been a clear advance for researchers using linguistic stimuli in Spanish.

However, one limitation of the LEXESP/Corco database is the closed architecture and lack of versatility of Corco. Indices that are not included in the LEXESP files cannot be integrated within Corco. Statistics such as age of acquisition, orthographic neighborhood measures, phonological neighborhood measures, and syllable frequency, among others, are of considerable interest to researchers in psy-

This work was partially supported by an Australian Research Council Post-Doctoral Grant to the first author. The authors thank Manuel Carreiras, Fernando Cuetos, Maria Antonia Martí, Jaime Redondo, and Núria Sebastián-Gallés for their permission to include lexical information they had previously collected within this software package. Thanks are also due Edurne Laseka and Hans Stadthagen-González. Correspondence concerning this article should be addressed to C. J. Davis, Department of Experimental Psychology, University of Bristol, 8 Woodland Rd., Bristol BS8 1TN, England (e-mail: colin.davis@bristol.ac.uk) or to M. Perea, Departament de Metodologia, Facultat de Psicología, Universitat de València, Av. Blasco Ibáñez, 21, 46010 València, Spain (e-mail: mperea@uv.es).

cholinguistics. However, these indices are not collected in the LEXESP database and cannot be integrated within the Corco program. Thus, researchers need to make use of other tools and applications to select their stimuli. Furthermore, computing some of the relevant indices (e.g., orthographic and/or phonological neighborhood size) requires researchers to write their own programs.

We believe that the best solution to this situation is to provide a single, user-friendly application in which the files from the Spanish databases (e.g., LEXESP and age-of-acquisition ratings, among others) are combined, so that researchers can easily obtain the relevant indices to manipulate/control the linguistic stimuli in their experiments. One such solution is *BuscaPalabras* (*Wordsearch* in English).

BuscaPalabras (henceforth, we will use the abbreviated name *B-Pal*) is the Spanish version of the original N-Watch program for English stimuli (Davis, 2005). It takes into account the particular characteristics of the Spanish orthographic system (e.g., accent marks, diacritic marks, and the letter *ñ*), that cannot be used with N-Watch. Furthermore, unlike the original N-Watch application, B-Pal provides indices related to syllable measures (i.e., token and type syllable frequency), since the syllable seems to play a key role in the processing of Spanish stimuli (see Carreiras, Álvarez, & de Vega, 1993; Carreiras & Perea, 2002, 2004; Perea & Carreiras, 1998). Another relevant feature of B-Pal is that it allows researchers to employ user-defined indices. This is especially useful because new norms on different potentially relevant variables are currently being compiled (e.g., valence and arousal [Redondo, Fraga, Comesaña, & Perea, 2005] and objective age-of-acquisition norms [Pérez, 2004], among others), and as such, there is no "ultimate" database. As an example, in the present version of B-Pal, we have included the subjective age-of-acquisition norms from Cuetos, Ellis, and Álvarez (1999) as a user-defined index.

As in the original N-Watch program, B-Pal computes statistics such as neighborhood size, neighborhood frequency, transposed-letter neighbors (e.g., *calvo–clavo*), and related measures of orthographic similarity online; furthermore, these outputs can be obtained for both word and nonword inputs, as will be explained below. In addition, to our knowledge, this is the first database to provide information on phonological neighborhood with Spanish stimuli.[3] That is, the program is useful for researchers in both written and spoken word processing domains.

In sum, leaving aside the comprehensive role of B-Pal for researchers working with Spanish stimuli, the program provides a number of orthographic and phonological indices of special interest for researchers, as we will describe below. The program runs on Windows PCs (preferably with at least 64 MB of RAM), and the full package (including data files) requires approximately 3.5 MB of hard disk space. It is available, free of charge, from the following Web site: www.maccs.mq.edu.au/~colin/B-Pal.

The Default Vocabulary

The most updated and comprehensive lexical database in Spanish, LEXESP, has a total of 166,494 separate entries with their corresponding frequencies (the corpus is around 5 million words). However, many entries are proper nouns (e.g., *Martínez*), words linked with hyphens (e.g., *rey-de-españa*), or words that are not Spanish (e.g., *oosterschelde*, *where*, *vrai*). Furthermore, many entries contain nonalphabetic characters (e.g., *'frica*, *&se*), and some of them are not even pronounceable (e.g., *grrrrrrrrr*, *zzpldos*). To filter the raw LEXESP corpus, we cross-checked these entries against the lexical entries in the official Spanish dictionary, the *Real Academia Española* (RAE) dictionary (electronic edition; Real Academia de la Lengua, 1995). In the first step, we removed any duplicates in the RAE dictionary (e.g., *heroína* is a homonymic word that has two lexical entries in the RAE dictionary, one corresponding to the English word *heroine* and the other corresponding to the English word *heroin*). Second, we eliminated any corpus entries in the LEXESP database that were not contained in the RAE dictionary. This eliminated misspellings, nonlexical abbreviations, and other linguistic oddities (see Pérez, Alameda, & Cuetos, 2003, for a similar approach). Third, lexical entries in the RAE dictionary with a frequency of zero were also eliminated, which included very low frequency words that are not likely to be in most people's internal lexicons (e.g., the word *mizo*). Fourth, given that most experiments in which verbal stimuli are used employ words that are 3–12 letters in length, only words in this range were included in the dictionary. The total number of entries included in the default vocabulary is 31,491 entries. For each entry, we collected a number of objective indices taken from the LEXESP database: word frequency, pronunciation (using DISC codes), and position of the lexical stress. As we will describe below, B-Pal provides a number of other indices, especially ones relating to orthographic/phonological neighborhoods.

Of course, we acknowledge that the present default vocabulary is not perfect. For instance, the RAE dictionary includes many words that are used in only one—or a few—Spanish-speaking countries and, as such, are unknown for the large majority of speakers (e.g., the word *quila*, which is a type of bamboo in some countries of South America, has one occurrence in the database). Likewise, some of the entries reflect very low frequency words that are probably unknown to most well-educated Spanish speakers (e.g., the word *tarra*, which is a term that denotes an old person, has one occurrence in the database). Finally, the present database does not include the large variety of inflected forms of each Spanish verb, and thus researchers whose focus is on verb processing are referred to the whole LEXESP database and to a Spanish dictionary that lists all the potential verbal forms. In any case, it should be noted that B-Pal can use any user-defined vocabulary (the file containing the database is in .txt format), so that researchers can readily use a vocabulary file other than the default vocabulary that is provided with the program.

Specifying the Stimuli to Be Analyzed

To simplify things for nonnative Spanish-speaking users, the menus are presented in English, although the Help menu includes an option to provide descriptions of

the output fields in Spanish. The use of the program is essentially the same as that for the English version (Davis, 2005). The program's main window resembles a spreadsheet, in which each of the stimuli specified by the user occupies a separate row and the statistics for that stimulus are displayed in separate columns. There are three ways to input stimuli to the program: (1) by typing individual stimuli into the edit line at the top of the screen, (2) by using the File|Open menu option to read in a text file (a list of stimuli, with one stimulus per line), or (3) by pasting a list of stimuli from the clipboard (by using the Edit|Paste menu option, the right-click pop-up menu, or just the shortcut Ctrl-V). The latter option is particularly useful when one has a list of words in another open document (e.g., an Excel spreadsheet or a text file); the list can be selected, copied onto the clipboard, and then pasted directly into the program. Users with non-Spanish keyboards can input one of the accented characters (á, é, í, ó, ú, ü, or ñ) by selecting the letter from a drop-down list (situated near the top right of the display) and adding it to the edit line by clicking the button next to this list.

Available Statistics

When the program starts, the only reported statistic is the LEXESP frequency per million words. This is just the value from the LEXESP database divided by five. In some cases, it may be appropriate to match items on log frequency. One of the program's output fields (LOG10_FRQ) returns the (base 10) logarithm of a word's frequency (plus 1). Additional output fields can be selected by clicking the *Analyse Options* button. This brings up a list of available statistics. Other than word frequency, these statistics can be divided into the following four broad categories: orthographic statistics, phonological statistics, neighborhood statistics, and assorted other statistics. In the following description, output fields are denoted in italicized capitals (e.g., *LEXESP*).

Orthographic statistics. Most of the statistics in this category are bigram frequency measures, which are both position and length sensitive. These bigram frequencies were computed on the basis of the LEXESP word frequency corpus.[4] For example, the stimulus *gato* (Spanish for *cat*) contains three bigrams (*ga*, *at*, and *to*). For the first of these (*ga*), the corresponding bigram frequency is based on the number (and frequency) of four-letter words that begin with *ga*; for example, the type frequency for *ga* is 15 (these types including *gafa*, *gafe*, *gala*, etc.), and the token frequency is the sum of the word frequencies for these 15 types (equals 118.2). The token frequency of the *n*th bigram is obtained by selecting the field BF*n* (e.g., selecting BF1 for the stimulus *gato* gives a value of 118.2, representing the token frequency of the first bigram, *ga*). B-Pal can also use these bigram frequencies to compute a variety of summary measures for the entire string. The BF_TK field outputs the average bigram token frequency across the entire letter string; for example, for *gato* the BF_TK value equals (118.2 + 257.5 + 1070.7) / 3 = 482.1. The BF_TP field is the average bigram type frequency across the entire letter string; for example, for *gato* the BF_TP value equals (15 + 29 + 45) / 3 = 29.67. Summed log bigram frequency (SLBF) is the sum of the logarithms of the token frequencies of each of the bigrams contained in the letter string. Mean log bigram frequency (MLBF) is simply SLBF divided by the number of bigrams in the stimulus (i.e., the number of letters minus one). Finally, LEN_L is the number of letters in the stimulus, and the CV_O field provides a simple description of the letter string's orthographic consonant–vowel structure (e.g., *gato* has a CVCV structure).

Phonological Statistics

Most of the phonological statistics output by the program are specific to words or, more correctly, those words for which a pronunciation is listed in the vocabulary file (unlisted stimuli return values of −1); the program's default vocabulary of 31,491 words includes pronunciations for each word. These output fields include the word's pronunciation (DISC_PRON), its initial phoneme (P1), its stress patterns (STRESS), the number of phonemes (LEN_P) and syllables (LEN_S) that it contains, and whether it has any homophones (HOM). If the word has a homophone, the spelling of this homophone is output (e.g., *bello* for the word *vello*; note that the letters *b* and *v* are pronounced as /b/ in Spanish); otherwise, a value of −1 is returned. The pronunciation of a word is transcribed in DISC phonetic codes, in which each phoneme is coded by a single character. Syllable boundaries are indicated by hyphens (e.g., *gato* is coded as gA–tO). The STRESS field returns the number of the syllable that is stressed in the word (this is always 1 for monosyllabic words). The CV_P field provides a simple description of the letter string's phonological consonant–vowel structure. For example, *gato* has a CVCV structure, whereas *hato* (/ɑ–tɔ/) has a VCV structure.

The program's default phonology is the one included in LEXESP, which is that common to most regions in Spain. The letter *z* is pronounced like /θ/ in most of Spain (and in the LEXESP phonological codes), but it is pronounced like /s/ in the southern regions of Spain and the Canary Islands, as well as throughout Latin America. That is, the words *caza* (*hunt*) and *casa* (*house*) are homophones for a speaker from Mexico, but not for a speaker from Madrid. Likewise, when appearing in the combinations *ce* and *ci*, the letter *c* is pronounced like /θ/ in most regions of Spain (and in the LEXESP phonological codes), but like /s/ in the southern regions, the Canary Islands, and Latin America. To accommodate these regional variations, it is possible to switch between the two phonologies by selecting the Options|Change Phonology menu option.

B-Pal also offers biphone frequency statistics that are computed in much the same way as those for bigram frequency, except that they are based on phonological codes. For example, selecting the MLBPF field gives the mean log frequency of the biphones in a word (e.g., MLBPF for *gato* is 2.52).

Neighborhood statistics. There are several statistics in this category. *N* is the standard measure of orthographic neighborhood size, determined by counting the number

of words that can be formed by substituting a single letter at any of the letter positions within the string (Coltheart, Davelaar, Jonasson, & Besner, 1977). This metric has been found to be related to measures of performance in a variety of reading tasks, including lexical decision, naming, perceptual identification, and semantic categorization (for reviews, see Andrews, 1997; Perea & Rosa, 2000). It should be noted that the program counts a word as an orthographic neighbor only if that word is included in the currently selected vocabulary (e.g., the default vocabulary does not include the word *zulo*, and so this word is not counted as an orthographic neighbor of *mulo*).[5] A list of the orthographic neighbors for each stimulus can be seen by switching to a different window (Window|Orth Neighbour List); choose Window|Main Form to return to the main window.

Several fields provide information about the distribution of neighbors. The N1 through N5 fields display the number of neighbors at Positions 1 through 5, respectively (if applicable; i.e., 1 through 4 for four-letter stimuli); for example, *gato* has nine neighbors at Position 1 (*dato*, *hato*, *lato*, *mato*, *nato*, *pato*, *rato*, *tato*, and *ñato*), five at Position 3 (*gago*, *galo*, *gamo*, *gano*, and *gayo*), and one at Position 4 (*gata*), but none at Position 2. *P* is a count of the number of positions at which legal neighbors can be formed (e.g., $P = 3$ for the stimulus *gato*). Pugh, Rexer, and Katz (1994; see also Mathey & Zagar, 2000) found that this metric, which they referred to as *spread*, was inversely correlated with lexical decision latency.

Other fields provide information about the frequency of a letter string's neighbors. The average frequency of the letter string's neighbors is measured by NF_MU. The standard deviation of these neighbor frequencies is measured by the NF_SIG field. NF_MAX is the frequency of the highest frequency neighbor (e.g., *gato*'s highest frequency neighbor is *rato*, which has a frequency of 72.5). NF_MIN is the frequency of the lowest frequency neighbor (e.g., *gato*'s lowest frequency neighbor is *gago*, which has a frequency of 0.2). Finally, it has been suggested that the critical neighbor frequency variable is relative frequency, rather than absolute frequency (e.g., Grainger, O'Regan, Jacobs, & Segui, 1989). Two output fields provide measures of relative frequency: HFN is the number of neighbors of the input that have higher frequencies, and LFN is the number of neighbors of the input that have lower frequencies than the input string. For example, of the 15 neighbors of *gato*, one is a higher frequency neighbor, and 14 are lower frequency neighbors. In the case in which the input is a nonword, HFN $= N$ and LFN $= 0$.

Other measures of orthographic similarity. Although investigations of orthographic similarity effects have focused mainly on neighbors formed by letter substitution, there are other forms of similarity relationship that have also been shown to influence performance in standard reading tasks. For example, perception of the word *cera* is affected not only by the presence of orthographic neighbors such as *cara*, but also by the presence of the orthographically similar word *crea*. This type of similarity relationship, in which two letter strings differ with respect to a single pair of adjacent letters, is known as transposed-letter similarity. Empirical work has shown that transposed-letter similarity affects performance in a variety of reading tasks, including lexical decision, naming, and semantic categorization (e.g., Andrews, 1996; Chambers, 1979; Perea & Carreiras, in press; Perea & Lupker, 2003, 2004; Taft & van Graan, 1998). Selecting the TL field causes the program to check whether the input string is a member of a transposed-letter pair—that is, whether a word can be formed by transposing an adjacent pair of letters in the input string. If a transposed-letter competitor is found, the identity of this competitor is reported in the TL field (e.g., given the input *calvo*, the output of the TL field is *clavo*). If the TL_FRQ field is selected, the frequency of the other member of the transposed-letter pair is reported. The TL_POS field records the (initial) position of the letter transposition (e.g., for the word *calvo*, TL_POS $= 2$).

Recent research has extended the work above by showing effects of transposed-letter similarity across nonadjacent letters—for example, between nonwords such as *caniso* and words such as *casino* (e.g., Perea & Lupker, 2004). Selecting the NATL field causes the program to check whether a word can be formed by transposing a pair of letters that are separated by one intervening letter (e.g., *molar* and *moral*).

A further form of orthographic similarity that has recently been shown to influence reading performance is subset/superset similarity. For example, research in which English stimuli have been used has shown that the presence of embedded words (e.g., *arm* within the word *army*) interferes with both lexical decision (Davis & Taft, 2005) and semantic categorization (Bowers, Davis, & Hanley, 2005). There is also evidence of an inhibitory effect of *addition neighbors*—words that involve the addition of a letter (e.g., *gato* and *gasto*; Bowers et al., 2005; Schoonbaert & Grainger, 2004; van Heuven & Dijkstra, 2005). Selecting the SUB and SUP fields causes the program to identify deletion neighbors (subsets) and addition neighbors (supersets) of the input stimulus, respectively; the frequency of these neighbors can be obtained by selecting the SUB_FRQ and SUP_FRQ fields. These neighbors are also displayed in the Neighbour List window (by selecting Window|Orth Neighbour List), provided that the option to show them is selected in the Analysis Options form. Finally, the N* field returns a count of all of a word's substitution, addition, and deletion neighbors. For example, *N** is 17 for the stimulus *gato*, because it has 15 substitution neighbors (i.e., $N = 15$), as well as 2 addition neighbors (*gasto* and *grato*).

Phonological Neighbors

The fields in this category are directly analogous to those for the orthographic neighborhood statistics, with one important exception, which is that the PN field includes not only substitution neighbors, but also deletion and addition neighbors, following the usual convention for computing phonological neighborhoods. Thus, the phonological neighbors of *gato* (/ɡɑ–tɔ/) include the deletion neighbor *hato* (/ɑ–tɔ/) and the addition neighbors *grato* (/ɡrɑ–tɔ/)

and *gasto* (/gɑs–tɔ/), as well as substitution neighbors such as *gano* (/gɑ–nɔ/) and *gallo* (/gɑ–ʟɔ/; note that this would not count as an orthographic neighbor). Another difference from the orthographic neighborhood measures is that the phonological neighborhood statistics are available only for words that are listed in the program's vocabulary (i.e., words for which the phonological transcription is known). A list of a word's phonological neighbors can be seen by switching to a different window (Window|Phon Neighbour List); choose Window|Main Form to return to the main window.

Syllabic Measures

In recent years, there has been some shift in the studies of word recognition, in the sense that there is growing interest in experiments in which multisyllabic words are used, as opposed to short monosyllabic words. As Rastle and Coltheart (2000) have pointed out, any comprehensive model of lexical access needs to confront the problems that arise when multisyllabic words are considered. This is even more relevant in Spanish, in which the percentage of multisyllabic words is much higher than in English (see Carreiras & Perea, 2002). Furthermore, the syllable is considered a perceptual unit for the process of word identification in Spanish. More specifically, research on visual word recognition in Spanish suggests that a word's syllabic neighbors are partially activated during identification of the target word, possibly via a syllable level that mediates between the letter level and the word level (e.g., when *cabo* is presented, the high-frequency word *casa* is partially activated, because of the shared initial syllable, /ka/). One key finding for the advocates of this account is the syllable frequency effect: Words composed of two high-frequency syllables are responded to more slowly than words composed of two low-frequency syllables in lexical decision (Carreiras et al., 1993; see also Álvarez, Carreiras, & Taft, 2001; Carreiras & Perea, 2002, 2004; Perea & Carreiras, 1998). Perea and Carreiras (1998) found that the main determinant of the syllable frequency effect in lexical decision was the activation of high-frequency *syllabic* neighbors. Syllabic neighbors would be words that share one syllable in two-syllable words, especially the initial syllable, given the special role of the initial letters in visual word recognition. Note that syllabic effects are posited to be phonological, rather than orthographic, in nature (see Álvarez, Carreiras, & Perea, 2004), and hence, the program computes syllable frequency on the basis of phonological codes. Nonetheless, given that prior research in Spanish has used the *orthographic syllable*, rather than the *phonological syllable* (e.g., Carreiras et al., 1993; Carreiras & Perea, 2002, 2004; Perea & Carreiras, 1998), the program also computes syllable frequency on the basis of the orthographic syllables. Note that, because of the characteristics of Spanish, both values tend to be rather similar, except for a few letters (e.g., *b*/*v*, *g*/*c*, and *h*). The orthographic syllabification of a word can be found by selecting the ORTH_SYLL output field; for example, for the word *hasta*, the ORTH_SYLL field gives the output *has·ta*.

B-Pal offers a number of indices relating to syllable frequency, including type frequency, token frequency, and maximum syllabic neighbor frequency. This is computed for both phonological syllables and orthographic syllables. Each of these measures is computed separately for the first, second, and third syllables, and measures are both position and length sensitive (e.g., the syllable frequencies returned for the first syllable of a two-syllable word are based only on the initial syllables of disyllabic words). For example, when the phonological syllables are used, the type and token frequencies for the first syllable of *hasta* are 14 and 1,192.14, respectively (i.e., there are 14 disyllabic words beginning with the syllable /ɑs/, and the summed frequency of these words is 1,192.14); the maximum syllable frequency in this case is 1,148.57 (this is the frequency of the most common disyllabic word beginning with /ɑs/, which is *hasta*). When the orthographic syllables are used, the type and token frequencies for the first syllable of *hasta* are 1 and 1,148.57, respectively (i.e., *hasta* is the only disyllabic word beginning with the orthographic syllable *has*).

Subjective Ratings

Research has shown that a number of subjective ratings are excellent predictors of the latency to recognize and respond to words. In particular, subjective measures of concreteness, familiarity, and imageability are known to be good predictors of word recognition (e.g., Balota, Pilotti, & Cortese, 2001; Gernsbacher, 1984; James, 1975; Whaley, 1978). The LEXESP database provides norms for each of these measures for 6,233 words. These norms can be obtained in the B-Pal program by selecting the CONC, FAM, and IMG fields. These scores are on a scale from 1 to 7, where higher scores indicate greater *concreteness*/*familiarity*/*imageability*. We also include two ratings (*stimulus valence* and *arousal level*) that are of great interest for researchers focused on emotional processing (e.g., Hermans, De Houwer, & Eelen, 2001). Ratings of valence and arousal level are available for a set of 466 words (these were collected by Redondo et al., 2005). These ratings can be obtained in the B-Pal program by selecting the VAL and ARO fields. These scores are on a scale from 1 to 9, where higher scores indicate greater *valence*/*arousal*.

User-Defined Fields

Users are able to add up to three fields of their own. Each of these fields can be added by clicking the Load button next to one of the User Field labels in the Analysis Options form and selecting a text file. This text file should have the following format: It should contain a variable label on a line by itself at the top of the file, and each subsequent row should contain one word, followed by the corresponding variable value, separated by a tab. B-Pal will then return these values when the corresponding user field is selected. As an example of a user-defined index, we have included the (subjective) age-of-acquisition ratings collected by Cuetos et al. (1999) in the User1.txt file that is distributed with the program. The scale ranges from

1 to 11 (1 = *before 2 years old*, 2 = *two years old*, 3 = *three years old*, and so on up to 11 = *eleven years old or older*). For example, if this user field is selected, the input *gato* returns a value of 3.33.

Saving the Output

There are two ways to extract the output from the program: (1) by using the File|Save menu option to save the output to a text file (one stimulus per line, with tabs separating the output fields), or (2) by copying selected output to the clipboard (by using the Edit|Copy menu option, the right-click pop-up menu, or just the shortcut Ctrl-C). Once again, the latter option is useful when working with a spreadsheet program such as Microsoft Excel or OpenOffice; the required fields can be selected, copied to the clipboard, and then pasted directly into an open spreadsheet. To select all the input rows and output columns containing data, the Copy All option can be selected from the right-click pop-up menu.

REFERENCES

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

Álvarez, C. J., Carreiras, M., & Perea, M. (2004). Are syllables phonological units in visual word recognition? *Language & Cognitive Processes*, **19**, 427-452.

Álvarez, C. J., Carreiras, M., & Taft, M. (2001). Syllables and morphemes: Contrasting frequency effects in Spanish. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 545-555.

Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory & Language*, **35**, 775-800.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, **4**, 439-461.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, **29**, 639-647.

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory & Language*, **52**, 131-143.

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

Carreiras, M., & Perea, M. (2002). Masked priming effects with syllabic neighbors in the lexical decision task. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 1228-1242.

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

Chambers, S. M. (1979). Letter and order information in lexical access. *Journal of Verbal Learning & Verbal Behavior*, **18**, 225-241.

Coltheart, M. (1981). The MRC Psycholinguistic Database, *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). New York: Academic Press.

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.

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

Davis, C. J., & Taft, M. (2005). More words in the neighborhood: Interference in lexical decision due to deletion neighbors. *Psychonomic Bulletin & Review*, **12**, 904-910.

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.

Grainger, J., O'Regan, J. K., Jacobs, A. M., & Segui, J. (1989). On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Perception & Psychophysics*, **45**, 189-195.

Hermans, D., De Houwer, J., & Eelen, P. (2001). A time course analysis of the affective priming effect. *Cognition & Emotion*, **15**, 143-165.

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

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

Mathey, S., & Zagar, D. (2000). The neighborhood distribution effect in visual word recognition: Words with single and twin neighbors. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 184-205.

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

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

Perea, M., & Carreiras, M. (in press). Do transposed-letter similarity effects occur at a phonological level? *Quarterly Journal of Experimental Psychology*.

Perea, M., & Lupker, S. J. (2003). Does *jugde* activate COURT? Transposed-letter similarity effects in masked associative priming. *Memory & Cognition*, **31**, 829-841.

Perea, M., & Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of Memory & Language*, **51**, 231-246.

Perea, M., & Rosa, E. (2000). The effects of orthographic neighborhood in reading and laboratory word identification tasks: A review. *Psicológica*, **21**, 327-340.

Pérez, M. A. (2004). *Influencia del orden de adquisición del léxico en el reconocimiento de palabras*. Unpublished doctoral dissertation, Universidad de Murcia.

Pérez, M. A., Alameda, J. R., & Cuetos, F. (2003). Frecuencia, longitud y vecindad ortográfica de las palabras de 3 a 16 letras del Diccionario de la Lengua Española (RAE, 1992). *Revista Electrónica de Metodología Aplicada*, **8**, 1-10.

Pugh, K., Rexer, K., & Katz, L. (1994). Evidence of flexible coding in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **20**, 807-825.

Rastle, K., & Coltheart, M. (2000). Lexical and nonlexical print-to-sound translation of disyllabic words and nonwords. *Journal of Memory & Language*, **42**, 342-364.

Real Academia de la Lengua (1995). *Diccionario de la lengua Española*. Madrid: Espasa Calpe.

Redondo, J., Fraga, I., Comesaña, M., & Perea, M. (2005). Estudio normativo del valor afectivo de 478 palabras españolas. *Psicológica*, **26**, 317-326.

Santiago, J., Justicia, F., Palma, A., Huertas, D., & Gutiérrez, N. (1996). LEX I and II: Two databases of surface word forms for psycholinguistic research in Spanish. *Behavior Research Methods, Instruments, & Computers*, **28**, 418-426.

Schoonbaert, S., & Grainger, J. (2004). Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language & Cognitive Processes*, **19**, 333-367.

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

Taft, M., & van Graan, F. (1998). Lack of phonological mediation in a semantic categorization task. *Journal of Memory & Language*, **38**, 203-224.

van Heuven, W. J. B., & Dijkstra, T. (2005). *Extended neighborhood effects in visual word recognition*. Poster presented at the XIV meeting of the European Society for Cognitive Psychology, Leiden.

WHALEY, C. P. (1978). Word–nonword classification time. *Journal of Verbal Learning & Verbal Behavior*, **17**, 143-154.

NOTES

1. The frequencies from the Alameda and Cuetos (1995) corpus are available online at the following address: www.uhu.es/jose.alameda/archivos/diccio.zip.

2. The LEXESP database is available on CD-ROM. It can be purchased from the Web site of the Universitat de Barcelona: www.ub.es/edicions/libros/v14.htm.

3. We should note that Santiago, Justicia, Palma, Huertas, and Gutiérrez (1996) compiled a Spanish database with phonological information (number of syllables, syllables, syllable CV structure, and subsyllabic units); however, this database was restricted to written production in children, and hence the number of lexical entries was severely limited (fewer than 6,000 entries).

4. The program for generating these values can be obtained from the first author.

5. The word *zulo* will enter the RAE dictionary in its next edition. However, to be consistent with our criteria, we have not included that word in the default dictionary.

(Manuscript received September 6, 2004;
revision accepted for publication December 22, 2004.)