Behav Res (2013) 45:765–771
DOI 10.3758/s13428-012-0296-8

NIM: A Web-based Swiss army knife to select stimuli for psycholinguistic studies

Marc Guasch · Roger Boada · Pilar Ferré · Rosa Sánchez-Casas

Published online: 28 December 2012

Abstract NIM is Web-based software developed to help experimenters with some of the usual tasks carried out in psycholinguistic studies. It allows the user to search for words according to several variables, such as length, matching substrings, lexical frequency, or part of speech, in English, Spanish, and Catalan. NIM also provides the user with the possibilities to obtain different word metrics, such as lexical frequency, length, and part of speech; to find intralanguage and cross-language lexical neighbors; and to get control words for critical stimuli. Regardless of the language used, the program also enables the user to get the orthographic similarity between word pairs and to identify repeated items in lists of experimental stimuli. NIM is free and is publicly available at http://psico.fcep.urv.cat/utilitats/nim/.

Keywords Psycholinguistics · Stimuli selection · Lexical frequency · Orthographic similarity · Lexical neighbors · Web-based program

Every single experiment in psycholinguistics begins by collecting an appropriate set of stimuli to test the theory behind it. Once collected, these stimuli have to be controlled at least for length and lexical frequency, since these two variables have been widely demonstrated to influence lexical processing (e.g., Balota & Chumbley, 1984; Forster & Chambers, 1973; Hauk & Pulvermüller, 2004; Hudson & Bergman, 1985). When starting from scratch, the task of stimulus compilation can be very tedious, and manual control for the linguistic variables can be error prone. To facilitate this task and reduce errors, we present NIM: a suite of tools with a Web-based interface, which ensures ease of access as well as platform-independence.

At this moment, the complete suite provides six tools. The first four are language-dependent and are available in English, Spanish, and Catalan: (1) a search engine (one for each language) that allows the user to find words that fulfill some criteria, such as containing a given number of letters, having a lexical frequency within a certain range, belonging to a fixed part of speech, or beginning or ending with or containing a given input string of letters; (2) a search engine (also one for each language) that, given a list of words, returns indices of lexical frequency and word length, as well as information about the parts of speech to which the words belong; (3) a tool that enables the user to find lexical neighbors both within and across languages; and (4) a tool to speed up the search for control words matched in length and lexical frequency with a given target in one of the available languages. The remaining two tools of NIM can be used with any language that is written with Roman script: (5) a calculator to compute both Van Orden's measure of orthographic similarity (Van Orden, 1987) and Levenshtein distance (Levenshtein, 1966) for large numbers of word pairs at the same time, and (6) a tool that allows researchers to detect repeated words in large sets of experimental stimuli.

In the next sections, we will describe each tool in detail. But before describing them, it will be necessary to know in depth the core of NIM: the frequency databases used. The available corpora-based lexical frequency databases in English, Spanish, and Catalan were not constructed according to the same procedures. In order to make the three databases as directly comparable as possible, we have made a special effort to unify the information by including in all cases the nonlemmatized frequencies, based on written-only

M. Guasch · R. Boada · P. Ferré · R. Sánchez-Casas
CRAMC (Research Center for Behavior Assessment), Department of Psychology, Universitat Rovira i Virgili, Tarragona, Spain

M. Guasch (✉)
Departament de Psicologia, FCEP, Universitat Rovira i Virgili, Carretera de Valls, s/n, Campus Sescelades, 43007 Tarragona, Spain
e-mail: marc.guasch@urv.cat

corpora; matching the part-of-speech classifications; and using the same criteria to trim the original data. Thus, apart from providing an integrated interface that makes the use of the different databases easier, NIM allows researchers to make searches not only within languages, but also between languages, and in so doing, they can be sure that the data provided for every one of the three languages are comparable to each other. Thus, NIM is a very useful tool not only for studies conducted within particular languages, but also for research in bilingualism.

Lexical frequency corpora

The Spanish database

The frequency database selected for the Spanish search engine was the Léxico Informatizado del Español (LEXESP; Sebastián-Gallés, Martí, Carreiras, & Cuetos, 2000). This database is, up to now, one of the most extensive frequency databases in Spanish, and one of the most used in psycholinguistic studies involving the Spanish language. The corpus is based on 5,629,279 Spanish tokens (i.e., total count of words in the sampled texts) that give rise to 166,494 types (i.e., unique words after eliminating redundancies). The source material sampled to build the corpus was entirely obtained from written sources. It includes, according to the authors, different kinds of literary and nonliterary sources: Forty percent of the tokens belong to narrative texts (i.e., novels), and another 40 % were selected from press (including newspapers, sport magazines, and news magazines). The remaining 20 % were equally distributed between popular science magazines and essays. The time window of the source materials ranges from 1978 to 1995. LEXESP includes software (CORCO) to browse through the indices in the database. One of these indices is lexical frequency, and for a given word it provides the absolute frequency and the part of speech of a given word coded in its own way.

To feed NIM with data, we extracted two measures from LEXESP: the absolute frequencies and parts of speech of words. From the absolute-frequency index, we computed the relative frequency per million of each word. This measure was obtained by dividing the absolute frequency of the word by the total corpus count, and multiplying the result by 1,000,000 to avoid extremely small numbers (e.g., the word *perro* ["dog"] has an absolute frequency of 339 and a frequency per million of [339/5,629,279] * 1,000,000 = 60.2209). We also counted the numbers of letters. Then we conducted some data trimming on the corpus to get rid of aberrant entries. We eliminated punctuation marks, dates, numbers or words with numbers within them, and words containing either nonlexical symbols or diacritical marks that are unacceptable in Spanish (e.g., a grave accent [`]). A total of 30,769 types were eliminated for these reasons, resulting in a reduction of 18.5 % of the words in the original corpus. Then we recoded the information about parts of speech in a comprehensive way, converting codes like "ncfs000" (used in LEXESP) to "noun." This resulted in words with repeated parts of speech (e.g., the word *luz* ["light"] is a common noun but also a proper noun, and it thus had the "noun" code repeated twice). After eliminating redundancies, parts of speech were finally reduced from the original 15 in LEXESP to the following eight: noun, adjective, adverb, verb, pronoun, conjunction, interjection, and the "other" category, which included the categories of LEXESP with few members (e.g., articles), as well as other types of words that do not constitute a lexical category per se (e.g., numerals, which are in fact proper names for numbers). Furthermore, these eight categories were the only ones shared by the three databases included in NIM. The final Spanish database contained 135,725 types from 1 to 26 letters in length. They had relative frequencies per million ranging from 0.18 to 47,025.7, and they belonged to at least one, and up to five, of the eight possible parts of speech.

The Catalan database

The corpus used for the Catalan search engine was the dictionary of frequencies included in the *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC; Rafel, 1998). This corpus is the largest and most important one existing in Catalan. It is based on 51,253,669 Catalan tokens distributed in 137,400 lemmas (i.e., the canonical form of a word). As described by the authors, the source materials used to elaborate the corpus included 49 % informative writings (from the sciences, arts, religion, philosophy, etc.), 44 % literary writings (i.e., mainly narrative texts, but also including drama, essays, and poetry), and 7 % other nonliterary writings, such as press or personal letters. The time window for inclusion in the corpus ranged from 1833 to 1988, although the vast majority of the sampled texts were from after 1914. In contrast to CORCO, in which data for single words can be directly obtained, the software included in the Catalan frequencies dictionary provides as the default option the absolute frequencies and parts of speech of lemmas. To use the same criterion to compute the lexical frequency as with LEXESP, our first step was to split the Catalan lemmas into their single forms and to obtain the lexical frequency for a given form by adding the different frequency values of the same form belonging to different lemmas. For instance, the word form *cases* ["houses"] belongs to the lemma *casa* ["house"], but it is also a conjugated form of the lemma *casar* ["to marry"]. In this case, we added the absolute frequencies of the noun and the conjugated verb. With regard to parts of speech, each

word form inherited the parts of speech of all of its parent lemmas (e.g., in the example of *cases*, this word was coded as a noun as well as a verb).

After collapsing word forms and computing their absolute frequencies, we obtained 431,820 types. As in the case of Spanish, we computed their relative frequencies per million, and we also counted their numbers of letters. We then carried out data trimming similar to what we had done with LEXESP, by eliminating words that contained either characters illegal in Catalan (e.g., [ñ]) or nonlexical symbols. A total of 23,005 types were eliminated for these reasons (5.3 % of the initial count). Regarding parts of speech, CTILC classifies lemmas in 13 categories. Eight of them are the same ones that had been kept in the Spanish database after trimming, so we changed their labels to match them with the names used in Spanish. The remaining five minor categories were merged into the more general "other" category, to keep uniformity with the rest of the databases. After all data processing, the final database contained 408,815 types from 1 to 25 letters in length, with relative frequencies per million ranging from 0.02 to 48,581.50, and belonging to at least one, and up to seven, of the eight possible parts of speech.

The English database

The frequency corpus used for the English search engine was the written frequency data from the British National Corpus (BNC; BNC Consortium, 2007). The materials used for constructing this database were obtained from different sources. Sixty percent of the sources were books, another 25 % were periodicals, between 5 % and 10 % were miscellaneous published materials (e.g., brochures, advertising texts, etc.), another 5 % to 10 % were unpublished written materials (e.g., personal letters or essays), and the rest (less than 5 %) came from written speeches, scripts, and so on. Regarding the kind of sources, 75 % of the materials were obtained from informative writings (from the sciences, arts, world affairs, etc.), and the remaining 25 % were literary and creative works. The sampled temporal period was between 1964 and 1993. The written part included in NIM was based on 98,119,624 English tokens (after removing from the database the frequencies of punctuation marks, which are 12.2 % of the total tokens), giving a total count of 303,829 types. We counted the numbers of letters and computed the relative frequencies per million by using the same procedure as with the other two corpora. To obtain the final database, we eliminated 46,325 aberrant entries in the data (15.2 % of the total), such as words containing numbers or nonlexical characters. Words with more than 26 letters were also omitted. The original parts of speech in BNC were 14. Eight of these were the same ones used in the other two corpora, with different labels. First of all, we changed these labels to match those used in Catalan and Spanish. Then, as we had done with the other databases, we collapsed words in the remaining minor categories into the "other" category. As a result of the data processing, the final English corpus contained 257,504 types from 1 to 26 letters in length, with relative frequencies per million ranging from 0.02 to 61,702.581, and belonging to at least one, and up to six, of the eight possible parts of speech.

The toolbox

In this section, we will describe in detail the six tools included in NIM.

Word search engine

The first tool allows the user to retrieve words from the databases. The English, Spanish, and Catalan corpora are accessed via separate screens. The words retrieved by NIM will satisfy the lexical parameters specified by the user. One of these parameters is length: NIM can search words either with an exact number of letters or with a minimum or maximum number of letters (also, a combination of the final two options is possible). When using any of the tools involving word length, one must take into account that the hyphen "-" is counted by NIM as a letter. The logic behind counting this way is that the hyphen takes up a space on the screen and, although it is not properly a letter, it is a character that has to be considered when controlling for word length. The same criterion is applied in Catalan regarding the interpunct ("·"), a small dot used to create the letter "l·l" (i.e., geminate l).

Another parameter refers to letter strings, enabling the user to search for words that start or end with a given sequence, or that contain a given sequence within them. Letter strings can be up to six letters long, and two wildcards are accepted in the input field: the character "_" means that its space can be replaced by any other character, whereas the character "%" means that it can be replaced by any sequence of letters. For instance, if we were to use in English the first wildcard, such as in "p_ay," we would retrieve words including "play" and "pray." By contrast, using the second wildcard (i.e., "cr%ed"), we would obtain results including "cried" or "creed," but also words like "created" or "crowded." The different possible input fields (i.e., words "beginning with," "containing," and "ending with"), combined with the use of the two wildcards, allow the user to search for virtually any possible pattern of letter strings among the existing words.

NIM also allows the user to search for words by frequency, choosing relative frequency per million as the index of reference because it is more useful than absolute frequency, and so that data from the different languages can be compared. As in

the case of length, one can search for words with either a minimum or a maximum value of frequency, as well as combinations of both criteria. It is also possible to look for words with an exact value, obtaining as a result words with a frequency range between the integer part of the entered value and the next integer minus 0.01.

The last search criterion is part of speech: The user can look for words belonging to one or more of the eight possible parts of speech (i.e., noun, adjective, adverb, verb, pronoun, conjunction, interjection, and other). In the case of selecting more than one part of speech, the user can specify whether the selection is disjunctive or conjunctive. If the selection is disjunctive, one has to use the option labeled "or," and will obtain words that belong to at least one of the selected parts of speech. In contrast, a conjunction is tagged with the word "and," and the results will contain only words belonging, at least, to all of the parts of speech selected. There is yet one other option, which is to select the parts of speech that the user wants to exclude from the search. By combining inclusions and exclusions, the user can search, for instance, for words that can be used as nouns and adjectives, but not as verbs.

Once the search criteria have been defined, the user can go to the results page. This page displays information about the search done, as well as the words that meet the specified conditions. For technical reasons, the output is limited to 500 words. This means, for example, that if we look in the Spanish corpus for words containing [ñ], we will see that there are 2,079 results, but NIM will only show the first 500 according to alphabetical order. To get the rest of the words, the user would have to delimit the search criteria (e.g., adding as a search condition an exact value of word length or a delimited range of it). The columns in the output display include an ID number, the word itself, the relative frequency per million, the absolute frequency, the length, and the parts of speech to which the word can belong. Furthermore, in NIM all tables of results can be sorted by any of the columns, by clicking on the column header. In addition, all of the result pages have a button that exports the table directly to a Microsoft Excel spreadsheet to make data manipulation easier. The third constant in all of the result pages is that NIM includes a Print button that allows the user to output the results in a black-and-white layout without images.

Word value search engine

Once the experimenter has selected the stimulus list, it is necessary to check whether the different experimental conditions are matched in length and frequency. This can be easily done by using the second tool provided by NIM: a search engine that outputs the frequency and length values of the words inserted in a big text area, where the words can be either directly written or pasted from a clipboard.

One of the strengths of NIM is the possibility of entering raw data from any source, because before accessing the database the program trims the input, deleting a huge range of nonlexical symbols, as well as numbers. For example, it is possible to copy into the clipboard a whole Excel spreadsheet. The user does not have to worry about selecting the word list in a particular way: he or she can just copy different columns (even with numerical data), and NIM will trim the input, selecting just the words. The search in the database is case insensitive, but it is sensitive to diacritical marks. For example, in the Catalan corpus, the search for the word *os* ("bone") is not the same as the search for the word *ós* ("bear").

Concerning the output display, the first column contains an ID number. This ID number allows the user to sort the words in the order in which they were introduced into the input screen. As with the word search tool, the other columns in the results table are the words themselves, the relative frequencies per million, absolute frequency, length, and part of speech. In addition, this table shows the resulting values from a calculation of the base-10 logarithm of each word's relative frequency, plus one. This transformation is useful because, in some cases (Baayen & Lieber, 1997), the distribution of logarithmic frequencies satisfies the assumption of normality necessary for some statistical tests (e.g., Student's *t* test) better than does the distribution of the relative frequencies themselves. Finally, we added one to the relative frequency before computing the logarithm in order to avoid negative numbers.

Lexical neighbor search engine

A lexical neighbor is often defined as a word that differs from another word by one letter. The neighborhood size of a word (i.e., the number of lexical neighbors of that word) affects processing in visual word recognition, as has been demonstrated with different tasks and experimental paradigms (e.g., Carreiras, Perea, & Grainger, 1997; Van Heuven & Dijkstra, 1998). Thus, it is useful to have a tool to count how many lexical neighbors (and which ones) a word has. This is exactly the function of this tool in NIM. In order to obtain information about the lexical neighbors of a particular word, the user has to type the target word in the data input screen of the selected corpus. It is worth noting that when the user selects the corpus of reference, he or she can look for orthographic neighbors not only within a language (e.g., by typing a Spanish word and selecting the Spanish database), but also across languages. For instance, it is possible to enter a word in Spanish, but to look through the Catalan corpus for words that differ from it by one letter.

The output screen displays all of the combinations retrieved from the database that differ by one letter from the target word. The number of possible combinations can be calculated by multiplying the number of letters of the input word by the

number of possible substituting characters minus one. The numbers of characters that are possible to substitute in each letter position are not the same in the three languages involved, as this depends on the legal characters in the particular language. In the English version, the neighbors are created by exchanging each letter position of a word with one of the 26 following characters: [a b c d e f g h i j k l m n o p q r s t u v w x y z]. In the Spanish version, the list of valid characters is the same as that for English, plus seven additional items: [á é í ó ú ü ñ]. Finally, the Catalan version contains an extra ten valid characters beyond those from English: [à é è í ï ò ó ú ü ç]. In addition, it should be kept in mind that the hyphen ("-") has not been considered as a valid character in English, Catalan, or Spanish. Regarding the Catalan corpus, the same is true for the interpunct ("·"), because it always appears within the letter "l·l," composed of three characters. Finally, in addition to the resulting lexical neighbors present in the selected database, the output screen also displays the same indices returned by the word value search engine tool.

Control-word search engine

The aim of this tool is to speed up the process of searching for words with the same frequency and length as a given one. The user has only to enter a word, and NIM will suggest a list of 20 possible candidates with the same length and the closest frequency values to that of the target word.

To use this tool, one must introduce the word to be controlled and the language of choice in the input screen. Language selection for the input word is a necessary step, to allow the program to look for the word frequency in that language. Thus, it is possible to look for control words not only within a language, but also across languages (i.e., to find words in another language with the same length and frequency).

In the output screen, the input word is displayed in the middle of the results table. It is surrounded by ten words above and ten more words below it, which are those with the nearest frequency values above and below that of the target, respectively. If there are more words of a certain frequency than slots allocated to show the results, only a portion of these words will be presented.

Orthographic similarity calculator

This tool enables the user to calculate the degree of orthographic similarity between pairs of words. Before describing the tool, we will comment in detail on the indices on which NIM bases their calculations.

In studies of bilingualism, a relevant line of research focuses on cognate words, which are translation equivalents with a similar form (e.g., Midgley, Holcomb, & Grainger, 2011; Sánchez-Casas, Davis, & García-Albea, 1992; Sánchez-Casas & García-Albea, 2005). During the last decade, a growing number of studies have used the orthographic similarity metric (OS) developed by Van Orden (1987) as a continuous index of the cognate status of words (e.g., Dimitropoulou, Duñabeitia, & Carreiras, 2011; Duyck, Van Assche, Drieghe, & Hartsuiker, 2007; Schwartz, Kroll, & Diaz, 2007; Van Assche, Duyck, Hartsuiker, & Diependaele, 2009). This measure is an adaptation of the graphic similarity (GS) index described by Weber (1970) to compute how similar a substitution error in oral reading was to the correct response. The original Weber formula weighted seven subindices that had been chosen by the author according to her intuitions about the use of several cues in the identification of words. The formula was as follows:

$$\mathrm{GS} = 10\left(\frac{50\mathrm{F} + 30\mathrm{V} + 10\mathrm{C}}{\mathrm{A}} + 5\mathrm{T} + 27\mathrm{B} + 18\mathrm{E}\right).$$

Given that P is the first word of a pair and R the second word, the subindices in Weber's formula were interpreted as follows:

- F the number of pairs of adjacent letters in the same order shared by P and R
- V the number of pairs of adjacent letters in reverse order shared by P and R
- C the number of single letters shared by P and R
- A average number of letters in P and R
- T ratio of number of letters in the shorter word to the number in the longer one
- B 1 if the first letter in R is the same as the first letter in P, or otherwise 0
- E 1 if the last letter in R is the same as the last letter in P, or otherwise 0.

Considering the formula, the values of GS between two words could range from a number near 0 (when two words have nothing in common and very different lengths), to a number near 1,400 (for an extremely long word with repeated letter pairs that is compared to itself). This method to compute the similarity between words satisfies the commutative property: Any word A is as similar to word B as word B is to word A. However, the application of the formula can lead to an illogical result, that not all pairs of identical words will have the same similarity value. For example, the GS of a word like "blue" with itself is 975, whereas the GS of "white" with itself is 1,000. This is because the formula is sensitive to word length. In order to overcome this limitation, Van Orden (1987) proposed another measure based on Weber's computation of GS. Van Orden's OS measure is computed as follows (with P being the first word of a pair, and R the second):

$$OS = \frac{GS(P,R)}{GS(R,R)}.$$

The main advantage of this formula is that the resulting values are distributed between 0 and 1, the latter being the value of maximum similarity, regardless of a word's length. The disadvantage of this calculation is that it lacks the commutative property of similarity. For instance, given that the GS between "blue" and "white" is equal to 242.22, the OS between "blue" and "white" would be .24, whereas it would be .25 when the words' order was reversed.

More recently, Schepens, Dijkstra, and Grootjen (2012) have proposed another measure of orthographic similarity. This measure is based on Levenshtein distance (Levenshtein, 1966). This metric is a well-known measure in computer science, used to compute the difference between any two strings. It is based on the minimum number of edit operations necessary to transform one string into another, considering as edit operations insertions, deletions, and substitutions of one character. On the basis of this measure, Schepens et al. devised an index called *normalized Levenshtein distance* (NLD). This index is calculated by applying the following formula:

$$\mathrm{NLD} = 1 - \frac{\text{Levenshtein distance}(\text{string1}, \text{string2})}{\text{Max. string length}(\text{string1}, \text{string2})} .$$

The resulting value always ranges between 0 and 1, it satisfies the commutative property, and all pairs of identical words have the same distance value (i.e., 1), regardless of word length. According to Schepens et al. (2012), the NLD is a reliable index of orthographic similarity between word pairs, as it correlates strongly with experimentally collected similarity ratings.

Focusing again on NIM, this program can be used to compute both Weber's GS and Van Orden's OS (as well as the seven subindices used for the calculations), and also to compute the Levenshtein distance and Schepens et al.'s (2012) NLD. To obtain all of these measures, the user only has to type or to copy and paste the word pairs into a text area. After the usual character trimming performed with all of the tools, NIM will take the words in pairs according to their insertion order to compute the measures. The calculator is case insensitive, but it is sensitive to diacritical marks. For example, the comparison between *ángel* and *àngel* (both words meaning "angel," in Spanish and Catalan, respectively) will give an OS/NLD value lower than 1. Finally, as with the previous tools, the table of results allows the user to sort the data by any of the measures. This option can be very useful to detect identical cognates or to establish a division between cognate and noncognate translations in large lists of stimuli. The table is also directly exportable to Microsoft Excel.

Detection of duplicate words

Like the previous tool, this last functionality of NIM does not depend on the use of a particular language and can be applied to any language based on the Roman alphabet. It consists of a script to detect duplicate words in large sets of stimuli. In some experimental designs involving lexical stimuli, it is important not to repeat any word or character string. This tool allows researchers to save time when searching for duplicate entries in large lists of words. Moreover, it can also be used to reveal which words are used most recurrently in texts of up to 40,000 characters (limited by technical issues).

The use of this tool is similar to that of the previously presented tools, since the input can be directly copied from other programs. The output screen displays the number of entered words as well as the number of repetitions. It also shows a list of all repeated words, together with the number of times that each word appears in the input text.

Conclusions and future development

Researchers, technicians, laboratory assistants, and graduate and undergraduate students doing research on psycholinguistics often spend a long time looking for a good set of materials. In our opinion, NIM provides them with a useful set of tools specifically designed to cover a large spectrum of practical problems.

Furthermore, NIM is a PHP program with a Web-based interface. This characteristic ensures ease of use and cross-platform compatibility. This also means that its structure is highly scalable and that the possibilities of adding new features are large. For instance, the orthographic similarity calculator can be adapted to accommodate new indices of similarity. In a similar way, both the word search engine and the word value search engine can be set up to work with any language's corpus that feeds the database. This enlargement would also allow the user to search for lexical neighbors or control words across languages other than English, Spanish, or Catalan. The more tools that NIM has, the faster and easier will be the job of the psycholinguist, increasing both reliability and productivity.

Author note The present work has been supported by the Spanish Ministry of Science and Innovation (PSI2009-12616 and Plan E) and by the Autonomous Government of Catalonia (2009SGR-00401). The authors thank Núria Sebastián-Gallés and the Institute of Catalan Studies for their permission to use their respective lexical frequency corpora. The authors also thank Adam Kilgarriff for providing the BNC data, and Antonio Masip and Enric Sunyer for their help.

References

Baayen, R. H., & Lieber, R. (1997). Word frequency distributions and lexical semantics. *Computers and the Humanities, 30,* 281–291.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology. Human Perception and Performance, 10,* 340–357. doi:10.1037/0096-1523.10.3.340

BNC Consortium. (2007). The British National Corpus, Version 3 (BNC XML Edition). Retrieved from www.natcorp.ox.ac.uk/

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 857–871. doi:10.1037/0278-7393.23.4.857

Dimitropoulou, M., Duñabeitia, J. A., & Carreiras, M. (2011). Phonology by itself: Masked phonological priming effects with and without orthographic overlap. *Journal of Cognitive Psychology, 23,* 185–203. doi:10.1080/20445911.2011.477811

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 663–679. doi:10.1037/0278-7393.33.4.663

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

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology, 115,* 1090–1103.

Hudson, P. T., & Bergman, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language, 24,* 46–58.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics – Doklady, 10,* 707–710.

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2011). Effects of cognate status on word comprehension in second language learners: An ERP investigation. *Journal of Cognitive Neuroscience, 23,* 1634–1647.

Rafel, J. (1998). *Diccionari de freqüències*. Barcelona, Spain: Institut d'Estudis Catalans.

Sánchez-Casas, R., Davis, C. W., & García-Albea, J. E. (1992). Bilingual lexical processing: Exploring the cognate/non-cognate distinction. *European Journal of Cognitive Psychology, 4,* 293–310.

Sánchez-Casas, R., & García-Albea, J. E. (2005). The representation of cognate and noncognate words on bilingual memory: Can cognate status be characterized as a special kind of morphological relation? In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 226–250). New York, NY: Oxford University Press.

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition, 15,* 157–166.

Schwartz, A. I., Kroll, J. F., & Diaz, M. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Language & Cognitive Processes, 22,* 106–129.

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

Van Assche, E., Duyck, W., Hartsuiker, R. J., & Diependaele, K. (2009). Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological Science, 20,* 923–927.

Van Heuven, W. J. B., & Dijkstra, T. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language, 39,* 458–483.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory and Cognition, 15,* 181–198. doi:10.3758/BF03197716

Weber, R.-M. (1970). A linguistic analysis of first-grade reading errors. *Reading Research Quarterly, 5,* 427–451.