Behavior Research Methods
2010, 42 (1), 118-125
doi:10.3758/BRM.42.1.118

SYLLABARIUM: An online application for deriving complete statistics for Basque and Spanish orthographic syllables

Jon Andoni Duñabeitia, Joana Cholin, and José Corral
Basque Center on Cognition, Brain, and Language, Donostia, Spain
and University of La Laguna, Tenerife, Spain

Manuel Perea
University of Valencia, Valencia, Spain

and

Manuel Carreiras
Basque Center on Cognition, Brain, and Language, Donostia, Spain
University of La Laguna, Tenerife, Spain
and University of the Basque Country, Bilbao, Spain

The present article introduces SYLLABARIUM, a new Web tool addressing the needs of linguists, psycholinguists, and cognitive scientists who work with Spanish and/or Basque and are interested in retrieving information about several syllable-related parameters. This new online syllabic database allows the user to generate complete lists of Spanish and Basque syllables with information about the syllable frequency. Among other measures, for a given orthographic syllable, SYLLABARIUM provides its number of occurrences (i.e., the type frequency), the summed lexical frequency of the words that contain this syllable (i.e., the token frequency), and the positional distribution of type and token frequencies. The cross-language feature of SYLLABARIUM is of special interest to researchers aiming to explore the influence of the syllable in bilingualism. The Web tool is available at www.bcbl.eu/syllabarium.

For more than a decade, a large body of empirical evidence has shown that, for the recognition of a written word, subword units are accessed at early stages of visual word processing, and the properties of these subword units have an effect on reading behavior. Researchers have repeatedly reported data showing how letters and phonemes (e.g., Pelli, Farell, & Moore, 2003; Perea, Duñabeitia, & Carreiras, 2008b; Rastle & Brysbaert, 2006), syllables (e.g., Conrad, Stenneken, & Jacobs, 2006; Perea & Carreiras, 1998), and morphemes (e.g., Duñabeitia, Perea, & Carreiras, 2007a, 2008; Rastle, Davis, & New, 2004) constitute the building blocks of word processing. However, how polysyllabic words are segmented into their syllables during reading is still an open question, and further research is needed in order to shed some light on this issue. The present study reports the process of creating a database of Spanish and Basque syllables: SYLLABARIUM, a Web tool for psycholinguistic experiments that includes features for material selection and syllable analyses.

Syllables As Processing Units

Much research has been done to explore the influence of the syllable in word processing, focusing on transparent languages like Spanish, in which orthographic representations (i.e., graphemes) map to phonological representations (i.e., phonemes) almost in a one-to-one manner (see Álvarez, Carreiras, & Perea, 2004; Carreiras, Álvarez, & de Vega, 1993). Two findings have been repeatedly reported in the literature: the syllable-congruency priming effect (Carreiras & Perea, 2002) and the inhibitory effect of the first syllable's positional frequency (Carreiras et al., 1993).

The term *syllable-congruency priming effect* refers to the fact that, when a word is preceded by a string containing the same orthographic or phonological syllable, this word is recognized faster and more accurately than when it is preceded by a string in which the initial syllable is not the same. Carreiras and Perea (2002) were the first authors to report a syllable-congruency priming effect, showing that a Spanish word like PASTOR (*shepherd*), syllabified as PAS.TOR, was recognized faster in a Spanish lexical decision task when it was preceded by a string like PAS*** than when it was preceded by a string like PA****. In the same experiment, Carreiras and Perea also showed that a word like PASIVO (*passive*), syllabified as PA.SI.VO, was recognized faster in the lexical decision task when it was preceded by the syllabic-congruent string PA***** than

J. A. Duñabeitia, j.dunabeitia@bcbl.eu

when it was preceded by the graphemic control PAS***. These results led the authors to conclude that the initial syllable constitutes an important processing unit in word recognition, above and beyond initial graphemes.

In subsequent studies, this effect has been further defined and replicated, obtaining similar results in the same and different languages (e.g., for Spanish, see Álvarez et al., 2004; Carreiras, Ferrand, Grainger, & Perea, 2005; for French, see Chetail & Mathey, 2009), with different techniques (for an experiment using event-related brain potential recordings, see Carreiras, Riba, Vergara, Heldmann, & Münte, 2009), and populations (for a study testing neurological patients with Alzheimer's disease and neurologically intact elderly controls, see Carreiras, Baquero, & Rodríguez, 2008).

The evidence showing inhibitory effects of the first syllable is also extensive. This effect was first reported by Carreiras et al. (1993), who showed that words starting with a high-frequency syllable—namely, with a syllable that also appears in many other words of the language—yield to a reading cost (longer recognition and reading latencies), as compared with words starting with a low-frequency syllable—namely, a syllable that appears only in a few other words. This effect is interpreted in terms of competing activation of syllabic neighbors, whereby low-frequency syllables activate fewer competing lexical representations than high-frequency syllables do, and, consequently, the time needed for verifying words with low-frequency syllables is less than the time needed for verifying words with high-frequency syllables. Thus, a word containing a high-frequency syllable is, by default, a word with many syllabic neighbors, since many other words also share that syllable. Correspondingly, a word containing a low-frequency syllable is a word with few syllabic neighbors.

The findings from the initial Spanish study by Carreiras and colleagues (1993) have been replicated in subsequent studies in the same and different languages (e.g., for Spanish, see Álvarez, Carreiras, & Taft, 2001; Conrad, Carreiras, Tamm, & Jacobs, 2009; Perea & Carreiras, 1998; for German, see Conrad & Jacobs, 2004; for French, see Mathey & Zagar, 2000) and with different techniques (for studies using event-related brain potentials, see Barber, Vergara, & Carreiras, 2004; Hutzler et al., 2004; for a study using functional MRI, see Carreiras, Mechelli, & Price, 2006). These effects are commonly found in word recognition tasks (e.g., lexical decision) and constitute the key evidence that the syllable is a processing unit in the domain of visual word recognition.

The consideration of the syllable as a relevant unit is widespread in other domains as well. Speech production is another field that has extensively studied the role and functionalities of the syllable. The involvement of syllabic constituents within speech errors is commonly agreed upon. It has often been demonstrated that segmental exchange errors generally obey syllable internal positions (i.e., onsets exchange with onsets, nuclei with nuclei, and codas with codas) (Berg, 1988; MacKay, 1970; Meyer, 1992; Nooteboom, 1969; Shattuck-Hufnagel, 1979, 1983, 1987; Stemberger, 1982; Vousden, Brown, & Harley, 2000). There is also evidence from metalinguistic tasks that suggests that syllables play a role during speech planning (e.g., Schiller, Meyer, & Levelt, 1997; Treiman, 1983; Treiman & Danis, 1988; for a review, see Bagemihl, 1995). However, the syllable-congruency priming effect, which is a standard finding in language comprehension research, seems to be absent in production. Many researchers have tried to identify syllables as functional production units by presenting syllable-congruent primes prior to a to-be-produced target syllable, as in the PA****–PASIVO. example Under the assumption that syllables constitute relevant units during speech planning, it was predicted that syllable-congruent primes would speed up production relative to a syllable-incongruent prime, such as PAS***–PASIVO. However, after some initial findings in French (Ferrand, Segui, & Grainger, 1996) and English (Ferrand, Segui, & Humphreys, 1997) that supported this assumption, numerous studies in various languages (for Dutch, see Baumann, 1995; Schiller, 1997, 1998; for English, see Schiller, 1999, 2000; Schiller & Costa, 2006; for French, see Brand, Rey, & Peereman, 2003; Evinck, 1997; for Spanish, see Schiller, Costa, & Colomé, 2002) could not find a syllable-congruency priming effect, but did discover a segmental length effect (see Schiller, 2004). Even under optimized conditions (i.e., longer and unmasked prime presentation), no syllable-congruency priming effect could be demonstrated for production (Schiller & Costa, 2006).

The production model by Levelt and colleagues (Levelt, Roelofs, & Meyer, 1999) explains the absence of syllable-congruency priming effects as follows: At the level at which the syllable prime taps into the speech planning process, there is no syllabic information available. The Levelt et al. model assumes that the stored word-forms that are retrieved from memory are not yet syllabified. During word-form retrieval, a string of segments is spelled out, but is unspecified for syllables' internal positions. Thus, the more segments that are preactivated by the prime, the more efficient this prime is, leading to the segmental length effect, irrespective of syllable congruency. The production model by Dell (1986, 1988), on the other hand, assumes syllabified word-forms. Support for this assumption stems from studies showing priming effects of the abstract syllable structure (the consonant–vowel structure) regardless of segmental content (Costa & Sebastián-Gallés, 1998; Sevald, Dell, & Cole, 1995).

Evidence that syllables emerge at a postlexical encoding level stems from a study showing that syllables cannot be primed but are prepared for during production planning (Cholin, Schiller, & Levelt, 2004). Moreover, it has been proposed that syllables, as phonetic motor programs, are stored within a separate syllable inventory that supplies speakers with ready-made whole syllable units during phonetic encoding (Cholin, Levelt, & Schiller, 2006; Crompton, 1981; Levelt et al., 1999; Levelt & Wheeldon, 1994). The retrieval of precompiled syllable programs allows for rapid and fluent articulation and reduces the computational load relative to a segment-by-segment online assembly.

Syllable frequency effects in language production provide strong evidence for the assumption of stored syllables because only stored entities are assumed to exhibit frequency effects. In a number of studies using different

tasks and different languages (Carreiras & Perea, 2004; Cholin et al., 2006; Levelt & Wheeldon, 1994; Laganaro & Alario, 2006) and different populations (for a patient study, see Aichert & Ziegler, 2004), syllable-frequency effects were obtained. Interestingly, the typical pattern of results in these studies is the opposite of that observed in the language comprehension domain (the above-mentioned inhibitory effect of the first syllable): High-frequency syllables were found to be produced faster than low-frequency syllables. This result has been interpreted as showing faster retrieval times for high-frequency syllables (analogous to the word-frequency effect) (see, e.g., Jescheniak & Levelt, 1994).

Programs for Deriving Statistics on Syllables

To date, the most extended program for deriving statistics for different syllabic measures in Spanish (the language in which the influence of the syllable has been most extensively studied) is the BuscaPalabras software (B-PAL; Davis & Perea, 2005). B-PAL provides valuable indexes for many different psycholinguistic variables for 31,491 Spanish words taken from the LEXESP corpus (Sebastián-Gallés, Martí, Carreiras, & Cuetos, 2000). Important for the purposes of the present study is the inclusion of a number of syllabic measures. Davis and Perea's software provides the user with the orthographic syllabification of an input word (e.g., the Spanish word CAMA, meaning *bed*, is outputted as CA.MA, denoting two CV syllables), the type and token frequencies of a word's syllables (e.g., 90 and 3,946 appearances per million words,[1] respectively, for the orthographic syllable CA in CAMA), and the frequency of the highest frequency syllabic neighbor of the input word (e.g., 784, which corresponds to the word CADA, translated as *each*).[2]

However, for exerting an exhaustive control on a word's syllables, researchers might also need a series of different indexes that cannot be obtained from B-PAL. One paradigmatic example of this is the search for position-dependent frequencies for a given syllable (e.g., the number of times that a given syllable appears in a given position), which can be performed in B-PAL for first, second, and third syllables only. According to Davis and Perea (2005), syllabic measures are

> 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). (p. 669)

This represents a clear limitation, especially if one considers that almost half of the Spanish words included in the lexicon used by B-PAL have more than three syllables. Of the 31,491 words that B-PAL deals with, 15,989 words (50.77% of the total) have three or fewer syllables, whereas 15,502 words (49.23%) have more than three syllables. Hence, if a user wants to obtain syllabic statistics for the syllable BLE in the Spanish word ADORABLE (A.DO .RA.BLE), this will not be possible using B-PAL. This is indeed critical, since some Spanish syllables tend to appear in word-final positions, as is the case for the syllable BLE, which appears 116 times in Positions 1–3 but appears more than three times as often (361 times) in subsequent syllabic positions. There is general consensus among researchers who work on visual word recognition that initial syllables show syllabic effects most clearly (e.g., Álvarez, Carreiras, & de Vega, 2000), but other syllabic positions receive increasing attention. One of the most noteworthy examples is the case of affixation: Several studies have examined the different processing procedures of syllables that are also derivational morphemes (e.g., the Spanish prefix RE in RE.FOR.MA, meaning *reform*, vs. the syllable RE in RE.GA .LO, meaning *present*; see Domínguez, Alija, Cuetos, & de Vega, 2006). In morphologically rich languages, like Spanish, suffixing is much more common than prefixing; thus, it should be possible to obtain frequency values for final syllable positions as well, especially for research focusing on the interplay of morpho-phonological processes. Syllabic representations associated with suffixes have been unattended but might have confounded some the observed results. Thus, the ability to disentangle these variables is a great advantage of SYLLABARIUM. Moreover, the B-PAL software provides the user with extensive information related to a word, but only when that given word has been inserted as an input, and researchers might also want to search for a pool of stimuli of certain characteristics, given only some restrictions in a search parameter (e.g., words with initial syllables corresponding to a type frequency between 50 and 100; words containing the syllable BLE). Also, researchers might want to check for the number of occurrences of a given Spanish syllable in all possible positions, and to our knowledge, this is not yet possible. As we present below, these are some of the functionalities of the Web-based application SYLLABARIUM, which provides the user with frequency counts for all of the existing syllables in Spanish, as well as those in Basque.

Overview of SYLLABARIUM

The selection of Spanish as a base language for SYLLABARIUM naturally follows from the idea that Spanish, as a prototype of a transparent language, represents an optimal language for testing syllabic effects. But why is Basque used as a base language for SYLLABARIUM?

Basque is a non-Indo-European language spoken by about 700,000 speakers in the Basque Country (comprising a geographical region at the southwest of France and the northeast of Spain). Basque has typological traits that are uncommon among European languages (e.g., subject–object–verb type, ergative, agglutinative) and is a highly transparent language. The reasons for also choosing a Basque lexicon for SYLLABARIUM are threefold. First, in recent times, the number of studies in Basque has grown exponentially, due to its high relevance for psycholinguistic research. The specific properties of Basque make it represent a great medium for exploring orthographic (e.g., Duñabeitia, Molinaro, Laka, Estévez, & Carreiras, 2009), morphological (e.g., Duñabeitia, Laka, Perea, & Carreiras, 2009; Duñabeitia, Perea, & Carreiras, 2007a, 2007b; Vergara-Martínez, Duñabeitia, Laka, & Carreiras, 2009), lexicosemantic (e.g., Perea, Duñabeitia, & Carreiras, 2008a), and syntactic processes (e.g., Díaz, Erdozia,

Mueller, Sebastián-Gallés, & Laka, 2006). Certainly, efforts should be made to provide investigators of Basque with tools that allow for an appropriate selection and control of experimental materials. Second, and closely linked to the previous issue, a parallel of the Spanish B-PAL software has recently been created for Basque: E-Hitz (Perea et al., 2006). The abovementioned limitations of B-PAL are similarly applicable to E-Hitz, whose structure and functionalities essentially mimic those of B-PAL. And third, considering that a huge number of Basque speakers are also Spanish speakers (i.e., Basque–Spanish-balanced bilinguals), the combination of these two languages in a single software for analyzing and selecting syllables provides psycholinguists with an invaluable tool for designing experiments that aim to clarify the role of the syllable in bilingualism.

The cross-language feature of SYLLABARIUM is of special interest to researchers aiming to explore the influence of the syllable in bilinguals' word processing. Bilingual word recognition is affected by cross-linguistic orthographic and phonological overlap, as shown by a number of studies (e.g., Bijeljac-Babic, Biardeau, & Grainger, 1997; van Heuven, Dijkstra, & Grainger, 1998). There is consensus that the phonological representations in the two languages of a bilingual individual are simultaneously activated when a word in one of these languages is being read (e.g., Van Wijnendaele & Brysbaert, 2002), and this assumption has been accordingly integrated in recent models of bilinguals' word recognition (see Dijkstra & van Heuven, 2002). Thus, it can be expected that bilinguals with a certain level of proficiency in their second language, when presented with a word in one of the languages, will activate syllabic neighbors in that language (i.e., words in the target language that contain the same syllable), as well as syllabic neighbors in the nontarget language. In the case of Basque–Spanish bilinguals, this issue is highly relevant because the number of shared syllables is very high (more than 700 syllables). Hence, even though some of the Spanish orthographic syllables do not exist in Basque (e.g., the syllable CHO in CHO.CO.LA.TE), and despite the fact that some of the Basque orthographic syllables do not exist in Spanish (e.g., the syllable TXA in TXA.KUR, which is the Basque word for *dog*), many syllables are present in both languages (e.g., the syllable BA, which appears in the Spanish and Basque words for *whale*, BA.LLE.NA and BA.LE.A, respectively). Another important benefit of SYLLABARIUM is that it provides researchers with cross-linguistic statistics for syllables that exist in both Basque and Spanish.

The Basque and Spanish Lexical Databases

In order to create the syllable database for Basque and Spanish, we used the two most common lexical databases in these languages. The base lexicons used by the Spanish B-PAL (Davis & Perea, 2005) and the Basque E-Hitz (Perea et al., 2006) were selected for creating the base corpora of SYLLABARIUM. The Spanish lexicon comprises 31,491 words in their lemmatized form, and the Basque corpus comprises 18,511 words; both databases include words of less than 13 letters. The mean frequency of the words in the Spanish database is 12 (±2.18; range: 0–27,352). The mean length of these words is 8.06 letters (±2.12; range: 3–12). The mean frequency of the words in the Basque database is 24 (±3.68; range: 1–44,713). The mean length of these words is 7.85 letters (±2.14; range: 3–12). The words from the two corpora were taken in their syllabified form in order to compute all the measures corresponding to these words' syllables.

As explained by Davis and Perea (2005) and Perea et al. (2006), Spanish and Basque have straightforward syllabification rules. As in many other syllable-timed languages, Basque and Spanish have very transparent syllabic boundaries. In general terms, typical Spanish onsets allow a maximum of two consonants, and Spanish nuclei include a vowel followed and/or preceded by a semivowel, whereas Spanish codas allow a maximum of two consonants (Harris, 1969). Basque has a very similar syllabic structure, with the exception of more complex codas, since consonant clusters can occur with up to three consonants (e.g., the monosyllabic Basque word BELTZ, translated as *black*; see Hualde, Elordieta, & Elordieta, 1995). The stress pattern in Basque, however, is different from that in Spanish. Spanish tends to accentuate the penultimate syllable, whereas Basque usually locates the accent on the second syllable (from the onset) and the final syllable (to a lesser extent). (Note that these are general rules that may also vary within languages and across dialects.) The mean number of syllables of the words is 3.49 (±0.99; range: 1–7) in the Spanish database and 3.53 (±1.06; range: 1–7) in the Basque database. It should be noted that, in the Spanish and Basque databases, only words with fewer than 13 letters are included, and, within this limit, no words with more than 7 syllables were found.

The Basque and Spanish Syllable Counts

All of the different orthographic[3] syllables from the two lexicons were initially found and counted. Whereas 1,751 different syllables were obtained from the Spanish database, 1,481 different syllables were obtained from the Basque database. Of these syllables, 762 were present in both databases (e.g., the syllable BA), 989 were present only in the Spanish database (e.g., CHO), and 719 syllables (e.g., TXO) existed only in the Basque database.

For each of the syllables, different measures were computed. First, the type frequency was obtained by counting the number of occurrences of each orthographic syllable in each of the lexicons. The mean type frequency of the Spanish syllables was 63 (±240; range: 1–3,403), and the mean type frequency of the Basque syllables was 44 (±170; range: 1–2,249). Second, the token frequency was obtained for each syllable, corresponding to the summed lexical frequency of all of the words containing that syllable. The mean token frequency of the Spanish syllables was 499 (±1,970; range: 0–32,936), and the mean token frequency of the Basque syllables was 767 (±3,989; range: 1–100,120). Third, the mean lexical frequencies (i.e., the number of occurrences of the words that include those syllables appearing in the Spanish and Basque base corpora) and standard deviations of the Basque and Spanish words containing each of those syllables were also obtained. Fourth, for each of the syllables, the highest

frequency syllabic neighbor (i.e., the highest frequency word containing a given syllable) was obtained, together with this word's frequency. And fifth, the positional type and token frequency of each syllable from Positions 1 to 7 were computed (note that, as stated above, none of the words in the databases had more than 7 syllables). In this way, for each syllable and each language, we obtained the number of occurrences in the first, second, third, fourth, fifth, sixth, and seventh positions (positional type frequency) and the summed lexical frequency of the words, including each syllable in each of the positions (positional token frequency).[4] An additional measure corresponding to the number of appearances of letter clusters (e.g., the frequency of co-occurrence of the letters) was also included. This was done by counting the number of times that the letters that form syllables appeared in the respective lexicons, independently of whether those letters formed syllables (e.g., the bigram BA is a syllable in the Spanish word BA.ÑO, *bathroom*, but is not a syllable in BAR.CO, *ship*). Table 1 includes general information about mean type and token syllabic frequency (including standard deviations and ranges) for the whole set of Basque and Spanish syllables.

Consider, for instance, the Spanish and Basque syllable AL as an example. The type frequency of AL is 313 in the Spanish database and 321 in the Basque database. The token (summed) frequency is 3,247 in the Spanish database and 5,430 in the Basque database. The mean frequency of the Spanish words containing the syllable AL is 10 (±74), and the mean frequency of the Basque words containing the syllable AL is 17 (±110). The Spanish highest frequency syllabic neighbor of the syllable AL is ALGO (meaning *something*), and its lexical frequency is 742. The Basque highest frequency syllabic neighbor of the syllable AL is ALDE (meaning *side*), and its lexical frequency is 1,116. The positional syllabic frequencies of the Spanish syllable AL (namely, the way in which the 313 appearances of the syllable AL are distributed across positions) are as follows: 288 in Position 1 (token frequency: 2,916); 11 in Position 2 (token frequency: 264); 12 in Position 3 (token frequency: 66); 1 in Position 4 (token frequency: 0); 1 in Position 5 (token frequency: 0); and 0 in Positions 6 and 7 (token frequency: 0). The positional syllabic frequencies of the Basque syllable AL (the positional distribution of the 321 AL syllable appearances in Basque) are as follows: 180 in Position 1 (token frequency: 4,536); 3 in Position 2 (token frequency: 4); 74 in Position 3 (token frequency: 571); 47 in Position 4 (token frequency: 264); 16 in Position 5 (token frequency: 53); 1 in Position 6 (token frequency: 1); and 0 in Position 7 (token frequency: 0).

Creation of the Online Application

The Web site that hosts SYLLABARIUM was developed using PHP 5.0 as the server-side programming language. Information about syllables and words is stored on a relational database hosted in a MySQL 5.0 server. The code served to client browsers complies with the World Wide Web Consortium's XHTML 1.0 transitional recommendations (2002b) and CSS2 recommendations (2002a), so compatibility with current and future Web browsers is better guaranteed. These databases are open to future changes, such as the inclusion of different languages, new information, or restructuring of the present counts. The online application can be reached at www.bcbl.eu/syllabarium.

Input and Output Forms

When one initially uses SYLLABARIUM, a definition of the search mode is required. There are three basic search options available (see Figure 1), depending on the language(s) in which the user wants to perform the search: (1) Basque, (2) Spanish, and (2) both. The user can then provide the program with a single or multiple target syllables in either language by simply typing it into the designated text box ["Syllable(s)"]. When the user clicks on the *Submit* button, the program will search for the matching string in the selected language's syllable database. When the match is found, an output screen (see Figure 2) shows

Table 1
Mean Values, Standard Deviations, and Ranges of the Type and Token Frequencies (General and Position-Specific) Obtained for the Whole Set of Spanish and Basque Syllables

		Spanish				Basque			
	Range	*M*	*SD*	Minimum	Maximum	*M*	*SD*	Minimum	Maximum
Frequency	Type	63	240	1	3,403	44	170	1	2,249
Summed frequency	Token	449	1,970	0	32,936	767	3,989	1	100,120
Position 1	Type	18	86	0	2,389	12	59	0	1,212
Position 1	Token	221	1,216	0	28,967	310	2,833	0	95,955
Position 2	Type	18	62	0	827	12	40	0	507
Position 2	Token	161	676	0	10,362	286	1,760	0	52,105
Position 3	Type	15	67	0	1,038	10	50	0	726
Position 3	Token	78	397	0	6,558	122	734	0	14,456
Position 4	Type	9	53	0	910	6	39	0	753
Position 4	Token	32	234	0	4,781	46	341	0	6,544
Position 5	Type	3	25	0	542	2	19	0	347
Position 5	Token	8	103	0	3,504	10	98	0	2,078
Position 6	Type	0	4	0	151	0	4	0	126
Position 6	Token	1	19	0	795	1	15	0	380
Position 7	Type	0	0	0	3	0	0	0	9
Position 7	Token	0	0	0	3	0	0	0	13

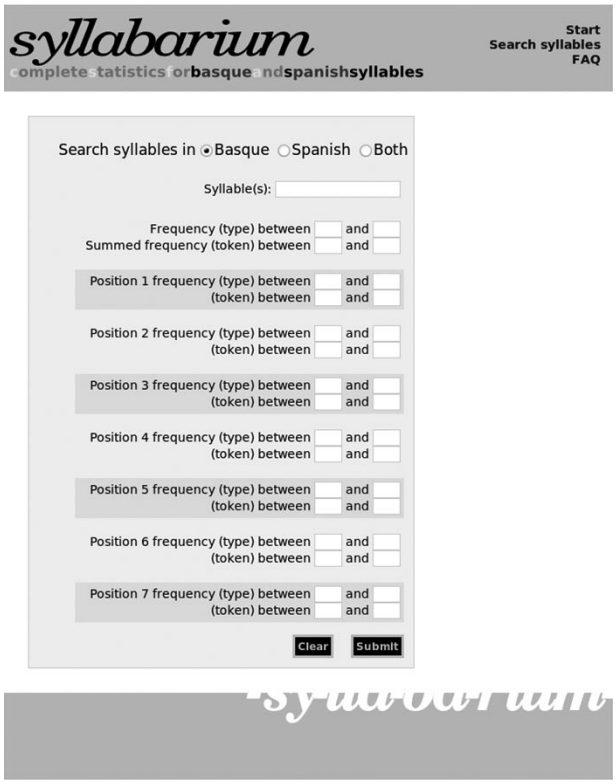

Figure 1. SYLLABARIUM's search screen, including the different parameter-delimiting options.

the type frequency of the target syllable, the summed frequency of all of the words containing that syllable, the letter string corresponding to the highest frequency syllabic neighbor and its frequency, the mean lexical frequency and standard deviation of all of the words containing the searched syllable, the type and token frequencies of the words containing the given syllable in Positions 1 to 7, and the number of appearances of the letters in the lexicon (an orthographic measure that does not rely on syllables but on the co-occurrence of the letters).

The user can also provide the program with one or multiple search parameters in which a minimum and a maximum frequency can be specified, thereby restricting the search. This option is of special interest to the user who, rather than wanting to obtain frequency values for a syllable, wants to select a subset of syllables for experimental purposes. The parameters that can be used for performing a restricted search are the type and/or token frequencies and the positional type and/or token frequencies. The output screen will show the same information shown for the syllable search, the only difference being that this will be presented for all of the syllables that match the search restrictions. A definition of each of the parameters is provided in a text file that contains basic information about the program and a set of frequently asked questions.

Word Retrieval

An additional feature of SYLLABARIUM is the *Export words* option. Once a search has been performed, the user can download all of the words that match the requested information. To this end, clicking on the corresponding button ("Export"; see Figure 2) causes a pop-up window to appear, allowing the user to save a plain text file in which all of the words containing the resulting syllable(s) are listed. Furthermore, the output corresponding to each word will be accompanied by the lexical frequency, the number of letters and syllables, and the number of orthographic neighbors (extracted from the B-PAL and E-Hitz lexical databases). Interestingly, the user can also delimit the number of letters and syllables of the words that will be exported (note that this is a useful delimitation for research on syllabic processing, which typically employs bisyllabic words). This feature of word exporting is of special interest to researchers who want to obtain a set of words that match one or many criteria for creating a stimuli list for an experiment.

Conclusion

The goal of the present article is to introduce SYLLABARIUM, an online application that offers several critical values for orthographic syllables in Spanish and Basque. This Web tool offers the ability to retrieve the frequency of occurrence for a syllable, as well as many other frequency-based measures (e.g., type and token frequencies, syllabic neighbors, positional frequencies). SYLLABARIUM also allows for broader searches, providing the ability to retrieve a list of the syllables (and the words containing those syllables) that match a series of restricted parameters as defined by the users. Due to its cross-linguistic feature, we believe that this Web tool will be particularly useful for researchers interested in bilingual word and syllabic processing.

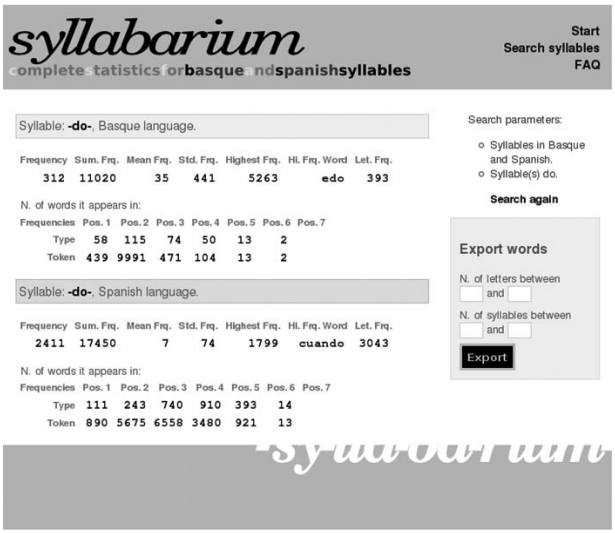

Figure 2. Example of SYLLABARIUM's output screen for the syllable "DO" in the combined language mode (Basque and Spanish). The output screen includes information about the type frequency of the syllable (Frequency), the token frequency (Sum. Frq.), the mean lexical frequency of the words containing the syllable (Mean Frq.), the standard deviation (Std. Frq.), the highest lexical frequency syllabic neighbor (Hi. Frq. Word) with its lexical frequency (Highest Frq.), and the letter co-occurrence (Let. Frq.).

AUTHOR NOTE

This research was partially supported by Grants SEJ2006-09238/PSIC, PSI2008-04069/PSIC, and CONSOLIDER-INGENIO2010 CSD2008-00048 from the Spanish government; Grant BFI05.310 from the Basque government; and Grant MTKD-CT-2005-029639 from the European Commission. The authors express their gratitude to Marc Brysbaert and to two anonymous reviewers for their comments on an earlier draft. Correspondence concerning this article should be addressed to J. A. Duñabeitia, Basque Center on Cognition, Brain, and Language, Paseo Mikeletegi, 53 20009–Donostia, Spain (e-mail: j.dunabeitia@bcbl.eu).

REFERENCES

Aichert, I., & Ziegler, W. (2004). Syllable frequency and syllable structure in apraxia of speech. *Brain & Language*, **88**, 148-159.

Álvarez, C., Carreiras, M., & de Vega, M. (2000). Syllable-frequency effect in visual word recognition: Evidence of a sequential-type processing. *Psicológica*, **21**, 341-374.

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

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

Bagemihl, B. (1995). Language games and related areas. In J. A. Goldsmith (Ed.), *Handbook of phonological theory* (pp. 697-712). Cambridge, MA: Blackwell.

Barber, H., Vergara, M., & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: Evidence from ERPs. *NeuroReport*, **15**, 545-548.

Baumann, M. (1995). *The production of syllables in connected speech*. Unpublished doctoral dissertation, Nijmegen University.

Berg, T. (1988). *Die Abbildung des Sprachproduktionsprozesses in einem Aktivationsflußmodell: Untersuchungen an englischen und deutschen Versprechern* [The representation of the speech production process in a spreading activation model: Studies of German and English speech errors]. Tübingen: Niemeyer.

Bijeljac-Babic, R., Biardeau, A., & Grainger, J. (1997). Masked orthographic priming in bilingual word recognition. *Memory & Cognition*, **25**, 447-457.

Brand, M., Rey, A., & Peereman, R. (2003). Where is the syllable priming effect in visual word recognition? *Journal of Memory & Language*, **48**, 435-443.

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

Carreiras, M., Baquero, S., & Rodríguez, E. (2008). Syllabic processing in visual word recognition in Alzheimer patients, elderly people, and young adults. *Aphasiology*, **22**, 1176-1190.

Carreiras, M., Ferrand, L., Grainger, J., & Perea, M. (2005). Sequential effects of phonological priming in visual word recognition. *Psychological Science*, **16**, 585-589.

Carreiras, M., Mechelli, A., & Price, C. J. (2006). Effect of word and syllable frequency on activation during lexical decision and reading aloud. *Human Brain Mapping*, **27**, 963-972

Carreiras, M., & Perea, M. (2002). Masked priming effects with syllabic neighbors in a 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.

Carreiras, M., Riba, J., Vergara, M., Heldmann, M., & Münte, T. (2009). Syllable congruency and word frequency effects on brain activation. *Human Brain Mapping*, **30**, 3079-3088.

Chetail, F., & Mathey, S. (2009). Syllabic priming in lexical decision and naming tasks: The syllable congruency effect re-examined in French. *Canadian Journal of Experimental Psychology,* **63**, 40-48.

Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition*, **99**, 205-235.

Cholin, J., Schiller, N. O., & Levelt, W. J. M. (2004). The preparation of syllables in speech production. *Journal of Memory & Language*, **50**, 47-61.

Conrad, M., Carreiras, M., & Jacobs, A. M. (2008). Contrasting effects of token and type syllable frequency in lexical decision. *Language & Cognitive Processes*, **23**, 296-326.

Conrad, M., Carreiras, M., Tamm, S., & Jacobs, A. M. (2009). Syllables and bigrams: Orthographic redundancy and syllabic units affect visual word recognition at different processing levels. *Journal of Experimental Psychology: Human Perception & Performance*, **35**, 461-479.

Conrad, M., & Jacobs, A. M. (2004). Replicating syllable frequency effects in Spanish in German: One more challenge to computational models of visual word recognition. *Language & Cognitive Processes*, **19**, 369-390.

Conrad, M., Stenneken, P., & Jacobs, A. M. (2006). Associated or dissociated effects of syllable frequency in lexical decision and naming. *Psychonomic Bulletin & Review*, **13**, 339-345.

Costa, A., & Sebastián-Gallés, N. (1998). Abstract phonological structure in language production: Evidence from Spanish. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 886-903.

Crompton, A. (1981). Syllables and segments in speech production. *Linguistics*, **19**, 663-716.

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**, 665-671.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, **93**, 283-321.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory & Language*, **27**, 124-142.

Díaz, B., Erdozia, K., Mueller, J. L., Sebastián-Gallés, N., & Laka, I. (2006). Individual differences in syntactic processing of a second language: Electrophysiological evidence. *Journal of Psychophysiology*, **20**, 228.

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language & Cognition*, **5**, 175-197.

Domínguez, A., Alija, M., Cuetos, F., & de Vega, M. (2006). Event related potentials reveal differences between morphological (prefixes) and phonological (syllables) processing of words. *Neuroscience Letters*, **408**, 10-15.

Duñabeitia, J. A., Laka, I., Perea, M., & Carreiras, M. (2009). Is Milkman a superhero like Batman? Constituent morphological priming in compound words. *European Journal of Cognitive Psychology*, **21**, 615-640.

Duñabeitia, J. A., Molinaro, N., Laka, I., Estévez, A., & Carreiras, M. (2009). N250 effects for letter transpositions depend on lexicality: "Casual" or "causal"? *NeuroReport*, **20**, 381-387.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2007a). Do transposed-letter similarity effects occur at a morpheme level? Evidence for morpho-orthographic decomposition. *Cognition*, **105**, 691-703.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2007b). The role of the frequency of constituents in compound words: Evidence from Basque and Spanish. *Psychonomic Bulletin & Review*, **14**, 1171-1176.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2008). Does darkness lead to happiness? Masked suffix priming effects. *Language & Cognitive Processes*, **23**, 1002-1020.

Evinck, S. (1997). *Production de la parole en français: Investigation des unités impliquées dans l'encodage phonologique des mots* [Speech production in French: Investigation of the units implied during the phonological encoding of words]. Unpublished doctoral dissertation, Bruxelles University.

Ferrand, L., Segui, J., & Grainger, J. (1996). Masked priming of word and picture naming: The role of syllabic units. *Journal of Memory & Language*, **35**, 708-723.

Ferrand, L., Segui, J., & Humphreys, G. W. (1997). The syllable's role in word naming. *Memory & Cognition*, **25**, 458-470.

Harris, J. (1969). *Spanish phonology*. Cambridge: MIT Press.

Hualde, J. I., Elordieta, G., & Elordieta, A. (1995). *The Basque dialect of Lekeitio*. Bilbao & Donostia/San Sebastián: Servicio Editorial de la Universidad del País Vasco/Diputación Foral de Gipuzkoa.

Hutzler, F., Bergmann, J., Conrad, M., Kronbichler, M., Stenneken, P., & Jacobs, A. M. (2004). Inhibitory effects of first syllable-

frequency in lexical decision: An event-related potential study. *Neuroscience Letters*, **372**, 179-184.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 824-843.

Laganaro, M., & Alario, F.-X. (2006). On the locus of the syllable frequency effect in speech production. *Journal of Memory & Language*, **55**, 178-196.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral & Brain Sciences*, **22**, 1-75.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, **50**, 239-269.

MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, **8**, 323-350.

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.

Meyer, A. S. (1992). Investigation of phonological encoding through speech error analyses: Achievements, limitations, and alternatives. *Cognition*, **42**, 181-211.

Nooteboom, S. G. (1969). The tongue slips into patterns. In A. G. Sciarone, A. J. von Essen, & A. A. van Raad (Eds.), *Nomen: Leyden studies in linguistics and phonetics* (pp. 114-132). The Hague: Mouton.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature*, **423**, 752-756.

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., Duñabeitia, J. A., & Carreiras, M. (2008a). Masked associative/semantic priming effects across languages with highly proficient bilinguals. *Journal of Memory & Language*, **58**, 916-930.

Perea, M., Duñabeitia, J. A., & Carreiras, M. (2008b). R34D1NG W0RD5 W1TH NUMB3R5. *Journal of Experimental Psychology: Human Perception & Performance*, **34**, 237-241.

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**, 610-615.

Rastle, K., & Brysbaert, M. (2006). Masked phonological priming effects in English: Are they real? Do they matter? *Cognitive Psychology*, **53**, 97-145.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, **11**, 1090-1098.

Schiller, N. O. (1997). *The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography*. Unpublished doctoral dissertation, Nijmegen University.

Schiller, N. O. (1998). The effect of visually masked syllable primes on the naming latencies of words and pictures. *Journal of Memory & Language*, **39**, 484-507.

Schiller, N. O. (1999). Masked syllable priming of English nouns. *Brain & Language*, **68**, 300-305.

Schiller, N. O. (2000). Single word production in English: The role of subsyllabic units during phonological encoding. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 512-528.

Schiller, N. O. (2004). The onset effect in word naming. *Journal of Memory & Language*, **50**, 477-490.

Schiller, N. O., & Costa, A. (2006). Activation of segments, not syllables, during phonological encoding in speech production. *The Mental Lexicon*, **1**, 231-250.

Schiller, N. O., Costa, A., & Colomé, A. (2002). Phonological encoding of single words: In search of the lost syllable. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology: Vol. 7. Phonology and phonetics* (pp. 35-59). Berlin: Mouton de Gruyter.

Schiller, N. O., Meyer, A. S., & Levelt, W. J. M. (1997). The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants. *Language & Speech*, **40**, 103-140.

Sebastián-Gallés, N., Martí, M. A., Carreiras, M., & Cuetos, F. (2000). LEXESP: Una base de datos informatizada del español [LEXESP: A computerized database of Spanish]. Barcelona: University of Barcelona.

Sevald, C. A., Dell, G. S., & Cole, J. S. (1995). Syllable structure in speech production: Are syllables chunks or schemas? *Journal of Memory & Language*, **34**, 807-820.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295-342). Hillsdale, NJ: Erlbaum.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 109-136). New York: Springer.

Shattuck-Hufnagel, S. (1987). The role of word-onset consonants in speech production planning: New evidence from speech error patterns. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language: Neuropsychology and neurolingistics* (pp. 17-51). Hillsdale, NJ: Erlbaum.

Stemberger, J. P. (1982). The nature of segments in the lexicon: Evidence from speech errors. *Lingua*, **56**, 235-259.

Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, **15**, 49-74.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory & Language*, **27**, 87-104.

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

Van Wijnendaele, I., & Brysbaert, M. (2002). Visual word recognition in bilinguals: Phonological priming from the second to the first language. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 616-627.

Vergara-Martínez, M., Duñabeitia, J. A., Laka, I., & Carreiras, M. (2009). ERP correlates of inhibitory and facilitative effects of constituent frequency in compound word reading. *Brain Research*, **1257**, 53-64.

Vousden, J. I., Brown, G. D. A., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology*, **41**, 101-175.

World Wide Web Consortium (2002a). Cascading Style Sheets Specification (Level 2) [Computer language manual]. Retrieved April 11, 2008, from www.w3.org/TR/CSS2.

World Wide Web Consortium (2002b). XHTML 1.0: The Extensible HyperText Markup Language (2nd ed.) [Computer language manual]. Retrieved April 11, 2008, from www.w3.org/TR/xhtml1.

NOTES

1. Henceforth, the presented and discussed type frequencies refer to the number of appearances per million words.

2. Note, however, that the B-PAL program only indicates the highest frequency syllabic neighbor's frequency, omitting the string that corresponds to that given frequency value. Therefore, and considering that researchers might also want to get that word, this additional information feature has been included in SYLLABARIUM.

3. In the present version of SYLLABARIUM, only statistics referring to orthographic syllables have been included. However, it should be noted that several studies have shown that syllable effects in visual word recognition are phonological in nature, rather than orthographic (see, e.g., Álvarez et al., 2004). Future versions of SYLLABARIUM will include complete statistics for Basque and Spanish phonological syllables.

4. Positional token frequency was included because recent research has shown that token syllabic counts (rather than type counts) are the best predictors of syllabic effects in language processing (for a review, see Conrad, Carreiras, & Jacobs, 2008).

(Manuscript received March 6, 2009;
revision accepted for publication July 16, 2009.)