Behavior Research Methods
2006, 38 (4), 610-615

E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque)

MANUEL PEREA
Universitat de València, València, Spain

MIRIAM URKIA
UZEI, Donostia, Spain

COLIN J. DAVIS
Royal Holloway University of London, Egham, England

AINHOA AGIRRE
UZEI, Donostia, Spain

and

EDURNE LASEKA and MANUEL CARREIRAS
Universidad de La Laguna, Tenerife, Spain

We describe a Windows program that enables users to obtain a broad range of statistics concerning the properties of word and nonword stimuli in an agglutinative language (Basque), including measures of word frequency (at the whole-word and lemma levels), bigram and biphone frequency, orthographic similarity, orthographic and phonological structure, and syllable-based measures. It is designed for use by researchers in psycholinguistics, particularly those concerned with recognition of isolated words and morphology. 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 www.uv.es/mperea/E-Hitz.zip.

Agglutinative languages (i.e., languages in which words are formed by joining morphemes together; e.g., Hungarian, Turkish, Basque) are an excellent testing ground for psycholinguistic research in some of the key issues in lexical access. At present, there are several useful databases for computing a number of relevant psycholinguistic statistics in nonagglutinative languages (for English, see Coltheart, 1981; Davis, 2005; for French, see New, Pallier, Brysbaert, & Ferrand, 2004; for Spanish, see Davis & Perea, 2005), but none for those in agglutinative languages. Here, we present both a word frequency list (for word forms and lemmas) and an application for computing a wide variety of psycholinguistic statistics (so that relevant variables can be manipulated and/or controlled) in an agglutinative language that is of special interest to researchers in psycholinguistics: Basque.

Basque is a pre-Indo-European language—with no demonstrable genetic relationship to any other living language—that is spoken at the western end of the Pyrenees, close to the Spanish–French border. Basque (*Euskara* in Basque) holds co-official language status (together with Spanish) in the Basque Autonomous Community and in some parts of Navarre. It is spoken by more than 700,000 people. Research in Basque is of special interest not just because of its very long history (its origins are unknown), but also because of its distinctive morphology and syntax.

The interest of Basque for morphological studies is based on the fact that Basque is an agglutinative language and, as such, has a high proportion of inflected words (e.g., *etxe* [house], *etxea* [the house], *etxeak* [the houses], *etxean* [in the house], *etxera* [to the house], *etxetik* [from the house]) and compound words (e.g., *etxezaina* [butler], *etxelaguna* [housemate], *etxetxoria* [sparrow], *etxejabe* [homeowner]). Indeed, compound words are often formed by several lexemes (e.g., the Basque word for "argument" [*eztabaida*] can be decomposed as *ez-da-bai-da*, which literally means "It is, it isn't"). The interest of Basque for syntax studies is due to the fact that Basque has not only rich inflectional morphology but also overt case markers, including up to three agreement markers on the inflected verb (marking subject, object, and dative agreement; e.g., *daramazkiot* ["I'm carrying (now) these (pl.)

The research reported in this article was partially supported by Grant SEJ2005-05205/EDU from the Spanish Ministry of Education and Science. Correspondence concerning this article should be addressed to M. Perea, Departament de Metodologia, Facultat de Psicologia, Av. Blasco Ibáñez, 21, 46010-València, Spain (e-mail: mperea@valencia.edu).

for her/him"]. Furthermore, Basque is a language with free word order, with a preverbal position for the focus of the sentence (e.g., nik *esan dut hori* ["it's *I* who said that"]; hori *esan dut nik* ["it's *that* what I said"]). Finally, Basque is an ergative language—that is, subjects of intransitive clauses and objects of transitive clauses are morphologically identical (e.g., *etxea* [subject] *erori da* ["the house has collapsed"]; *nik* [subject] *etxea* [object] *erosi dut* ["I bought the house"]), whereas subjects of transitive clauses are morphologically distinct from objects (see Laka & Korostola, 2001).

In summary, a frequency corpus in Basque (at the word-form and lemma levels) and its application may be useful to cognitive psychology researchers in their experiments in Basque, in both monolingual (see, e.g., Carreiras & Perea, 2005; Perea & Carreiras, 2006a) and bilingual (e.g., Perales & Cenoz, 2002) studies. Given the increasing interest in psycholinguistic research on agglutinative languages such as Basque, we believe that it is important to provide a single user-friendly application with a standardized vocabulary (at the word-form and lemma levels) so that researchers can easily obtain the relevant indices to manipulate and/or control the linguistic stimuli in their experiments. One such solution is *EuskalHitzak* ("BasqueWords").

EuskalHitzak (henceforth, *E-Hitz*) is the Basque version of the original N-Watch program for English stimuli (Davis, 2005). It takes into account the particular characteristics of the Basque orthographic system (e.g., the letter *ñ*) that cannot be used with N-Watch. Furthermore, unlike the original N-Watch application, and because of the specific characteristics of the Basque language, it contains one corpus based on word-form frequencies and another based on lemma frequencies. Furthermore, E-Hitz provides indices related to syllabic measures (i.e., token and type syllable frequencies) in orthographic and phonological terms.

Another relevant feature of E-Hitz is that it allows researchers to employ user-defined indices. This is especially useful because it may include new norms on different potentially relevant variables. As an example, in the present version of E-Hitz we have included the familiarity norms for compound words collected as part of an ongoing project, which can serve as a user-defined index.

Like the original N-Watch program, E-Hitz computes statistics such as neighborhood size, neighborhood frequency, transposed-letter neighbors (e.g., *gazta–gatza*), and related measures of orthographic similarity (including syllabic measures) online; furthermore, these outputs can be obtained for both word and nonword inputs, as will be explained below. In addition, this is the first database to provide information on phonological neighborhoods and phonological syllables with Basque stimuli. In other words, the program is useful to researchers in both written and spoken word processing domains.

In summmary, in addition to the comprehensive role of E-Hitz for researchers working with Basque stimuli, the program provides a number of orthographic and phonological indices of special interest to researchers, as we 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 8.2 MB of hard-disk space. It is available free of charge from www.uv.es/mperea/E-Hitz.zip.

The Reference Vocabulary

The most updated and comprehensive corpus in Basque is that developed by the UZEI (Basque Centre for Terminology & Lexicography). This project is based on data obtained from the Statistical Corpus of 20th Century Basque (available at www.euskaracorpusa.net/XXmendea), since actual data of current usage are absolutely necessary in order to reflect the current reality. Until now, the main application of this corpus has been associated with lexicography projects carried out by the Euskaltzaindia (Royal Academy of the Basque Language), especially the completion of *Hiztegi Batua* ("Unified Dictionary"). The corpus is based on an exhaustively classified inventory of Basque published works of the 20th century—all documents include period, dialect, and gender information—and was created from a statistical sample that reflected the universe proportionally (Urkia, 2002, 2005). For example, gender holds the same proportion in the corpus as in the published universe—that is to say, all topics are in the lexicon, as are oral transcriptions, and therefore many fields have been treated. As a result, the corpus comprises approximately 5 million words (*Maiztasun Hiztegia*, 2004).

Because current usage of Basque is needed for psycholinguistic experiments, periods have been delimited in the reference vocabulary in E-Hitz, and we have chosen only the last two (1969–1990 and 1991–1999). The reason for this is that the Basque Academy took the first steps toward standardizing Basque in 1968, although the rules did not make an appearance until the 1990s. Besides, the development of Basque dialects in recent years is reflected in the corpus. Furthermore, given that our focus was on the Basque language from the southern territories (the Basque Autonomous Community and Navarre—i.e., the provinces that form part of the state of Spain), and because of the peculiarities of Basque words spoken in the northern provinces (Lapurdi, Nafarroa Beherea, and Zuberoa—i.e., the provinces that form part of the state of France), only written texts from the southern territories were included in the reference vocabulary.

The 1969–1999 corpus has a total of 292,469 separate entries with their corresponding frequencies (around 3.5 million words).[1] However, a few of the entries are not common Basque words (i.e., they are marked as foreign) and some are not even normally pronounceable, since they were created with poetic license (e.g., *xxxxxxxiiiiito*). Given that one of our aims is to compute how many words known by a typical reader are orthographically similar to a given letter string, the presence of a number of highly infrequent words (or nonwords) in the reference vocabulary may be undesirable. To overcome this problem, we followed the method used for the N-Watch application (Davis, 2005) and initially filtered the corpus to exclude items that occurred only very rarely. To this end, the default vocabulary for the E-Hitz program was selected by excluding word forms with frequencies of less than 0.34 per million.

To further filter the whole-word frequency corpus, we cross-checked the word forms it contains against the word-

form entries in the *Elhuyar hiztegia* (2000) dictionary and the Xuxen spelling corrector (Aduriz et al., 1997).[2] For instance, word forms that did not have standard spellings (i.e., dialectal word forms) would be modified to their standard spellings (e.g., *maaia* was modified to *mahaia* [table]). Finally, given that most experiments in which verbal stimuli are used employ words that are 2–18 letters in length, only words in this range were included in the dictionary. The total number of entries included in the default vocabulary of the word-form corpus is 100,080.

For each entry in the word-form reference vocabulary, we computed three objective indices: orthographic syllabic structure, phonological syllabic structure, and pronunciation (using DISC codes). We used DISC coding because, unlike with CELEX, each phoneme is coded by a single character, and this facilitates the computation of phonology statistics. Furthermore, this is the coding used in prior work (Davis, 2005; Davis & Perea, 2005). As we describe below, E-Hitz provides a number of other indices, especially relating to orthographic/phonological neighborhoods. Finally, it should be noted that E-Hitz can use any user-defined vocabulary (the file containing the database is in .txt format), so that researchers can readily use a Basque vocabulary file other than the reference vocabulary that is provided with the program. For researchers interested in Basque, the UZEI unfiltered word-form corpus can be obtained from the authors and readily used as the reference vocabulary (e.g., in cases in which researchers are interested in the processing of unfamiliar, low-frequency words).

With respect to the lemma corpus, it is known that lemmatization is the process by which every word form of the corpus is assigned a standard lemma (i.e., a dictionary entry). This means that all declined or conjugated forms, as well as the dialectal variants, will have a common entry. For example, the word forms *etxe*, *etxea*, *etxia*, *etxetik*, and *etcheko* are included under the ETXE entry or lemma. This process was carried out using the automatic lemmatizer EUSLEM, thanks to the collaboration of UZEI and the IXA group of the University of the Basque Country (Aduriz et al., 1996; UZEI, 1996). The lemmatizer has been very useful for corpus lemmatization in recent texts, especially texts written in standard Basque. The lexicographers of UZEI checked manually the entire lemma corpus, taking special care to desambiguate and detect multiword expressions. Since the main focus of the corpus is the lexicon, some information was deleted from lemmas, so the corpus does not contain proper names, foreign words, or metalanguage, nor does it include inflected verbs or declined forms. Because the lemmas are categorized and written in the unified language, there was no need to filter the list. A total of 86,749 lemma entries are included in the default vocabulary of the lemma corpus. (Note that in the lemma corpus, hyphenated compound words—e.g., *irakasle-ohi*—would be written in the list without the hyphen—i.e., *irakasleohi*.)

Specifying the Stimuli to Be Analyzed

The menus in the program are presented in Basque, except for a Help menu in which we provide instructions and descriptions of the output files in English. To simplify things for readers who do not speak Basque, in the present article we provide the menus in English translation/Basque original. The use of the program is essentially the same as that of the English version (Davis, 2005). The program's main window closely resembles a spreadsheet: 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 different ways to input stimuli to the program: (1) Type individual stimuli into the Edit line at the top of the screen, (2) use the File|Open/Fitxategia|Berria menu option to read in a text file (e.g., a list of stimuli, with one stimulus per line), and (3) paste a list of stimuli from the clipboard by using the Edit|Paste/Editatu|Itsatsi menu option, the right-click pop-up menu, or simply 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 pasted directly into the program.

Available Statistics

When the program starts, the only reported statistic is the UZEI frequency per million words. This is simply the value from the UZEI database divided by 3.6. In some cases, it may be appropriate to match items on log frequency. One of the program's output fields (*LOG10_FRQ*/*LOG10_MAIZ*) returns the (base 10) logarithm of a word's frequency (plus 1). Additional output fields can be selected by clicking the Analysis Options/Analisi Aukerak button. This brings up a list of available statistics. These statistics can be divided into the following four broad categories, in addition to word frequency: orthographic statistics, phonological statistics, neighborhood statistics, and assorted other statistics. In the following description, output fields are denoted in italicized capitals (e.g., *WORD_FRQ*/*H_MAIZT*).

Orthographic statistics. Most of the statistics in this category are bigram frequency measures, which are both position and length sensitive. The bigram frequencies were computed on the basis of the UZEI word frequency corpus. For example, the stimulus *atea* (the door) contains three bigrams: *at*, *te*, and *ea*. For the first of these, the corresponding bigram frequency is based on the number (and frequency) of four-letter words that begin with *at*—e.g., the type frequency for *at* is 15 (these types including *atal*, *ater*, *atez*, etc.) and the token frequency is the sum of the word frequencies for these 15 types (equal to 70.0). The token frequency of the *n*th bigram is obtained by selecting the field BF*n*/*n*BM; for example, selection of *BF1*/*1BM* for the stimulus *atea* gives a value of 91.25, representing the token frequency of the first bigram. E-Hitz can also use these bigram frequencies to compute a variety of summary measures for the entire string. The *BF_TK/BM_TK* field ouputs the average bigram token frequency across the entire letter string; for example, for *atea* the *BF_TK*/*BM_TK* value equals (132.4 + 288.4 + 1,199.2)/3 = 540. The *BF_TP*/*BM_TP* field outputs the average bigram type frequency across the entire letter string. For example, for

atea the *BF_TP*/*BM_TP* value equals (15 + 29 + 45)/3 = 29.67. Summed log bigram frequency (*SLBF*/*BMLB*) 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*/*BMLBB*) is simply *SLBF*/*BMLB* divided by the number of bigrams in the stimulus (i.e., the number of letters minus one). Finally, *LEN_L/LUZ_L* is the number of letters in the stimulus, and the *CV_O/KB_O* field provides a simple description of the letter string's orthographic consonant–vowel structure (e.g., *atea* has a VCVV structure).

Phonological statistics. Most of the phonological statistics output by the program are specific to words, or, more correctly, to those words for which a pronunciation is listed in the vocabulary file (unlisted stimuli return values of −1). The program's reference vocabulary of 100,080 words includes pronunciations for each word. Although there is no "standard" pronunciation in Basque, the phonological transcription and syllabic segmentation processes were performed on the basis of the recommendations of the Euskaltzaindia.[3] The vowel system is the same as that of Spanish, and the consonant system is similar to that of Spanish (except for a few digraphs that represent single sounds [*ts*, *tx*, and *tz*]; see Hualde, 1991).

These output fields include the word's pronunciation (*DISC_PRON*/*AHOSKERA_DISC*), its initial phoneme (*P1*/*F1*), the numbers of phonemes (*LEN_P/LUZ_F*) and syllables (*LEN_S/LUZ_S*) it contains, and whether or not it has any homophones (*HOM*). If the word has a homophone, then the spelling of this homophone is output (e.g., *hura* for the word *ura*); 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., *atea* is coded as A-tE-A). The *CV_P/KB_F* field provides a simple description of the letter string's phonological consonant–vowel structure. For example, both *atea* and *hatua* (/A-tU-A/) have a VCVV structure.

E-Hitz 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 field *MLBPF*/*BPMLBB* gives the mean log frequency of the biphones in a word (e.g., the *MLBPF* for *atea* is 2.56).

Neighborhood statistics. There are a number of statistics in this category. The standard measure of orthographic neighborhood size is *N*, which is 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). A list of the orthographic neighbors of each stimulus can be viewed by switching to a different window (Window|Orth-Neighbor List/Leihoa|Orto-Auzokide Zerrenda); choose Window|Main Form/Leihoa|Leiho Nagusia to return to the main window.

Several fields provide information about the distribution of neighbors. Fields *N1–N5* display the number of neighbors at Positions 1–5, respectively (for four-letter stimuli, Positions 1–4 are covered). For example, *atea* has one neighbor at Position 1 (*otea*), four at Position 2 (*ajea*, *alea*, *area*, and *asea*), two at Position 3 (*atia* and *atxa*), and three at Position 4 (*ateo*, *ater*, and *atez*). *P* represents the number of positions at which legal neighbors can be formed (Pugh, Rexer, & Katz, 1994—e.g., $P = 4$ for the stimulus *atea*).

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 field *NF_SIG*. *NF_MAX* is the frequency of the neighbor with the highest frequency (e.g., *atea*'s highest frequency neighbor, *alea*, has a frequency of 21.7). *NF_MIN* is the frequency of the lowest frequency neighbor (e.g., *atea*'s lowest frequency neighbor, *atia*, has a frequency of 0.5). Finally, it has been suggested that the most critical neighbor frequency variable is relative frequency rather than absolute frequency (see, 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 than the input string, whereas *LFN* is is the number of neighbors of the input that have lower frequencies. For example, of the 10 neighbors of *atea*, all 10 are lower frequency neighbors. (Note that when the input is a nonword, $HFN = N$ and $LFN = 0$.)

Other measures of orthographic similarity. Investigations of orthographic similarity effects have focused mainly on neighbors formed by letter replacement. However, there is recent evidence that other forms of similarity relationship can also influence performance in standard reading tasks. In other words, perception of the word *gazta*, for example, is affected not only by the presence of orthographic neighbors such as *gazte*, but also by the presence of the orthographically similar word *gatza*. This type of similarity, in which two letter strings differ with respect to a single pair of adjacent letters, is known as transposed letter (TL) similarity. Recent work has shown that TL similarity affects performance in a variety of reading tasks, including lexical decision, naming, and semantic categorization (see, e.g., Andrews, 1996; Perea & Carreiras, 2006b; Perea & Lupker, 2003, 2004; Taft & van Graan, 1998). Indeed, there is recent evidence of robust TL similarity effects in Basque (Perea & Carreiras, 2006a). Selecting the *TL* field causes the program to check whether or not the input string is a member of a TL pair—that is, whether a word can be formed by transposing an adjacent pair of letters in the input string. If a TL competitor is found, its identity is reported in the *TL* field (e.g., given the input *gazta*, the output of the *TL* field is *gatza*). If the field *TL_FRQ*/*TL_MAIZTASUNA* is selected, then the frequency of the other member of the TL pair is reported. The field *TL_POS/TL_KOK* records the (initial) position of the letter transposition (e.g., for the word *gatza*, *TL_POS/TL_KOK* = 3).

A further form of orthographic similarity that has recently been shown to influence reading performance is subset/superset similarity. For example, recent research has shown that the presence of embedded words (e.g., *arm* within the word *army*) interferes with both lexical decision (Davis, Perea, & Taft, 2005; Davis & Taft, 2005; de Moor & Brysbaert, 2000) and semantic categorization (Bowers, Davis, & Hanley, 2005). There is also evidence of an inhibitory effect of *addition neighbors*—that is, words that involve the addition of a letter (e.g., *gatza* and *gaitza*; Bowers et al., 2005; Schoonbaert & Grainger, 2004). Selecting the fields *SUB*/*AZPIM.* and *SUP*/*GAINM.* 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 fields *SUB_FRQ*/*AZPIM_MAIZTASUNA* and *SUP_FRQ*/*GAINM_MAIZTASUNA*. These neighbors are also displayed in the Neighbor List/Auzokideen Zerrenda window (by selecting Window|Orth-Neighbor List/Leihoal Orto-Auzokide Zerrenda), 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 (this metric is defined in Davis & Taft, 2005). For example, $N^* = 28$ for the stimulus *atea* because it has 10 substitution neighbors (i.e., $N = 10$) as well as 15 addition neighbors (e.g., *atera*) and 3 deletion neighbors (e.g., *ate*).

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 phonological neighborhoods (*PN*) field includes not only substitution neighbors but also deletion and addition neighbors, following the usual convention for computing phonological neighborhoods (see Vitevitch & Luce, 1999, for details). Thus, the phonological neighbors of *atea* include deletion neighbors such as *ata* and addition neighbors such as *artea* and *batea*, as well as substitution neighbors such as *atia* and *hatua* (note that the latter 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 reference vocabulary of the whole-word list (i.e., words whose phonological transcriptions are known). A list of a word's phonological neighbors can be seen by switching to a different window (Window|Phon-Neighbor List/Leihoa|Fon-Auzokide Zerrenda); choose Window|Main Form/Leihoa|Leiho Nagusia to return to the main window.

Syllabic Measures

Research on visual word recognition suggests that a word's syllabic neighbors are partially activated during identification of the target word (at least in transparent languages with clearly defined boundaries), possibly via a syllable level that mediates between the letter level and the word level (e.g., when the Spanish word *cabo* is presented, the high-frequency word *casa* is partially activated because of the shared initial syllable /ka/). One key finding is the syllable frequency effect: Words composed of high-frequency syllables are responded to more slowly than words composed of low-frequency syllables in lexical decision (Carreiras, Álvarez, & de Vega, 1993; see also Álvarez, Carreiras, & Taft, 2001; Carreiras & Perea, 2002, 2004; Perea & Carreiras, 1998). Syllabic effects are posited to be phonological rather than orthographic in nature (see Carreiras, Ferrand, Grainger, & Perea, 2005), and, hence, the program computes syllable frequency on the basis of orthographic and phonological codes. Basque is a transparent language with clear, well-defined syllable boundaries. As such, recent research suggests that the syllable is a relevant factor in visual word recognition in Basque (Carreiras & Perea, 2005).

E-Hitz offers a number of indices relating to syllable frequency, including type frequency, token frequency, and maximum syllabic neighbor frequency. Each of these measures is computed separately for the first, second, and third syllables; these 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, the type and token (orthographic) frequencies for the first syllable of *atea* are 983 and 11,890.89, respectively (i.e., there are 983 trisyllabic words beginning with the syllable /A/, and the summed frequency of these words is 11,890.89 per million); the maximum syllable frequency in this case is 707.19 (this is the frequency of the most common disyllabic word beginning with /A/, which is *agertzen* ["appearing"]).

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*/*Bete* button next to one of the User Field/Eremuak Definitu labels in the Analysis Options/Analisi Aukerak form and then selecting a text file. This text file 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. E-Hitz will return these values when the corresponding user field is selected. As an example of a user-defined index, we have included the familiarity ratings for compound words collected, in an ongoing project, as the file User1.txt, which is distributed with the program. The scale ranges from 1 to 7. For example, if this user field is selected, the input *udaleku* ["summer camp"] returns a value of 4.39.

Saving the Output

There are two ways to extract the output from the program: (1) Use the File|Save/Fitxategia|Gorde menu option to save the output to a text file (one stimulus per line, with tabs separating the output fields) and (2) copy selected output to the clipboard (using the Edit|Copy/Editatu|Kopiatu menu option, the right-click pop-up menu, or just the shortcut Ctrl-C). Once again, the latter option is useful when one is working with a spreadsheet program such as MS-Excel or OpenOffice; the required fields can be se-

lected, copied to the clipboard, and pasted directly into an open spreadsheet. To select all the input rows and output columns containing data, the option Copy All/Kopiatu Dena can be selected from the right-click pop-up menu.

REFERENCES

Aduriz, I., Aldezabal, J. M., Alegria, I., Artola, X., Ezeiza, N., & Urizar, R. (1996). EUSLEM: A lemmatiser/tagger for Basque. *Proceedings of EURALEX'96*, **1**, 17-26.

Aduriz, I., Alegria, I., Artola, X., Ezeiza, N., Sarasola, K., & Urkia, M. (1997). A spelling corrector for Basque based on morphology. *Literary & Linguistic Computing*, **12**, 1.

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.

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

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., & 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.

Carreiras, M., & Perea, M. (2005, June). *Morphological or syllabic segmentation in Basque? Masked priming effects in an agglutinating morphology*. Paper presented at the 2005 Morphology Meeting, Cambridge.

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.

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., & 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.

Davis, C. J., Perea, M., & Taft, M. (2005, April). *The inhibitory effects of deletion neighbors in English and Spanish*. Paper presented at the VII Simposio de Psicolingüística, Valencia, Spain.

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.

de Moor, W., & Brysbaert, M. (2000). Neighborhood-frequency effects when primes and targets are of different lengths. *Psychological Research*, **63**, 159-162.

Elhuyar hiztegia: Euskara–gaztelania/castellano–vasco [*Elhuyar dictionary: Basque–Spanish/Spanish–Basque*] (2000). Usurbil, Spain: Elhuyar.

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.

Hualde, J. I. (1991). *Basque phonology*. New York: Routledge.

Laka, I., & Korostola, L. E. (2001). Aphasia manifestations in Basque. *Journal of Neurolinguistics*, **14**, 133-157.

Maiztasun hiztegia [*Word frequency dictionary*] (2004). Donostia, Spain: UZEI.

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

Perales, J., & Cenoz, J. (2002). The effect of individual and contextual factors in adult second-language acquisition in the Basque country. *Language, Culture, & Curriculum*, **15**, 1-15.

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. (2006a). Do transposed-letter effects occur across lexeme boundaries? *Psychonomic Bulletin & Review*, **13**, 418-422.

Perea, M., & Carreiras, M. (2006b). Do transposed-letter similarity effects occur at a prelexical phonological level? *Quarterly Journal of Experimental Psychology*, **59**, 1600-1613.

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.

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

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., Cuetos, F., Martí, M. A., & 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.

Urkia, M. (2002). XX. mendeko euskara-corpusa. [Twentieth-century Basque corpus] *Hizkuntza-corpusak: Oraina eta geroa. Jardunaldiak*. Donostia, Spain: UZEI.

Urkia, M. (2005). Euskararen maiztasun hiztegia [Basque word frequency]. Donostia, Spain: UZEI.

UZEI (1996). EUSLEM: Euskarako lematizatzaile/etiketatzaile [EUSLEM: Lemmatizer/tagger for Basque]. Donostia, Spain: UZEI.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language*, **40**, 374-408.

NOTES

1. It is worth noting that the Spanish database (LEXESP; Sebastián-Gallés, Cuetos, Martí, & Carreiras, 2000) has a corpus of 5 million words and 81,475 word-form entries, whereas the French database (Lexique—New et al., 2004) has a corpus of around 300 million words and 129,000 word-form entries.

2. The Xuxen spelling corrector is available at www.euskadi.net/euskara_soft/.

3. The recommendations of the Euskaltzaindia can be found at www.euskaltzaindia.net/arauak/dok/ProNor0013.htm.

(Manuscript received June 19, 2005;
revision accepted for publication September 14, 2005.)