EHME: a New Word Database for Research in Basque Language

Joana Acha, Itziar Laka, Josu Landa, and Pello Salaburu
Universidad del País Vasco UPV/EHU (Spain)

The research reported in this article has been supported by grants EHU 12/02 from the Universidad del País Vasco UPV/EHU, IT665–13 from the Basque Government, FFI2012–31360 from the Spanish Ministerio de Economía y Competitividad, and FP7-SSH-2013–1 from the European Union.

Correspondence concerning this article should be addressed to Joana Acha. Department of Basic Cognitive processes. Universidad del País Vasco UPV/EHU. Tolosa Hiribidea. 20018. Donostia (Spain). E-mail: joana.acha@ehu.es

Abstract

This article presents EHME, the frequency dictionary of Basque structure, an online program that enables researchers in psycholinguistics to extract word and nonword stimuli, based on a broad range of statistics concerning the properties of Basque words. The database consists of 22.7 million tokens, and properties available include morphological structure frequency and word-similarity measures, apart from classical indexes: word frequency, orthographic structure, orthographic similarity, bigram and biphone frequency, and syllable-based measures. Measures are indexed at the lemma, morpheme and word level. We include reliability and validation analysis. The application is freely available, and enables the user to extract words based on concrete statistical criteria[1], as well as to obtain statistical characteristics form a list of words[2].

Received 20 November 2013; Revised 10 February 2014; Accepted 4 April 2014.

Keywords: psycholinguistics, corpus linguistics, word frequency, orthographic and phonological structure, word statistical measures

[1] http://www.ehu.es/ehg/ehme/datu2hitz.htm

[2] http://www.ehu.es/ehg/ehme/hitz2datu.htm

Research with linguistic stimuli requires tools for computing psycholinguistic statistics in order to select and manipulate the word parameters that the researcher has got in mind. At present, there are several databases for computing most relevant psycholinguistic statistics in alphabetic and non agglutinative languages (for English, see Davis, 2005; for Spanish, see Davis & Perea, 2005). However, most recently, languages with other typological properties have entered the arena of research in psycholinguistics, particularly in the field of word recognition and reading. This is the case of Basque.

Basque is a non-indoeuropean isolate language spoken by more than 700,000 people, which has an array of typological properties that have recently become the focus of interest for research on psycholinguistics. First, it is an agglutinative language (such as Finnish or Turkish) so that all inflectional morphemes are morphologically complex, corresponding to phrases or inflected verbs, comprising several morphological constituents (De Rijk, 2007; Hualde & Ortiz de Urbina, 2003; Laka, 1996). For example the lexeme "*etxe*" (house) can be attached to one morpheme (etxe-a [the house], or several morphemes (*etxe-a-ren* [of the house]) or also to another lexeme (*etxe-bide* [houseway, way to the house]) to form a compound. Second, Basque has rich morphology, that is, most words are composed by a lexeme and a limited set of inflectional or derivational morphemes, available at http://www.ehu.es/seg/morf/5/16 which operate in syntax and the lexicon, respectively (Azkarate, 1993). Thus, lexeme/morpheme manipulations can highlight questions about how derivational morphology that has an impact in vocabulary-formation (as in [*etxe-gile*, "house-builder"]) and inflectional morphology that has an impact on syntax (as in [*etxe-a-k*, the house transitive subject]). Third, Basque is an ergative language (Laka, 2006). This implies, roughly, that transitive subjects are marked differently from objects and intransitive subjects, which are marked alike. Ergativity is a rare typological property (25% of languages in the world) not found in Europe (Dixon, 1994). Basque displays great word order freedom, and word order variations convey differences in informational structure (new and old information). Fourth, Basque coexists with Spanish in the western side of the Basque speaking area, and with French on the eastern side. Spanish is similar to Basque in terms of orthographic transparency (almost direct grapheme-phoneme mapping) but it is a Romance language, not agglutinative or ergative. Hence, the nature of Basque makes it a suitable language to examine the role of lexical and morphological processes during word and sentence processing, particularly in cross-language studies. This is the reason why psycho/neurolinguisitic research on this language has increased significantly during the last decade (see Acha, Laka, & Perea, 2010; Carreiras, Duñabeitia, Vergara, de la Cruz-Pavia, & Laka, 2010; Erdozia, Laka, Mestres-

Misse, & Rodriguez-Fornells, 2009; Laka & Korostola, 2001; Zawiszenwski, Gutierrez, Fernandez, & Laka, 2011, among others).

As a consequence, a linguistic database for Basque (E-HITZ) was recently developed by Perea et al., (2006) based on the design of the above mentioned databases. This has been the most used and cited reference for Basque researchers during the last years. This corpus took into account the characteristics of the Basque orthographic system, including measures for lemmas and also for whole word forms. The measures provided included those most relevant for researchers in psycholinguistics such as word frequency, syllable frequency, word structure, word length, syllabification, bigram frequency and word neighborhood measures at two levels: orthography (measures based on orthographic computations) and phonology (measures based on phonological computations). E-HITZ is a complete and user-friendly application to extract word indices, and can be used from a free downloadable application from the author's webpage. However, the database has some limitations that needed to be overcome. The main one is that research on morphemic complexity requires exact estimations of compound, derived and inflected whole words, but also of lemmas and morphemes in isolation, and the currently available database does not supply with this information. The second one is that, taken that into account, neighborhood statistics have to be calculated for each neighborhood type (substitution, deletion, addition and transposition). Finally, E-HITZ offers the possibility to extract the statistics from a word set, but it does not permit to extract a word list from some previously settled criteria. Based on the limitations observed, and in order to provide researchers with a more comprehensive tool, we developed EHME.

EHME, Landa, Sarasola, & Salaburu, 2010) is a rich application of a Basque word frequency dictionary based on texts of the 21st century that provides the user with all relevant measures for language researchers. It is based on a corpus of 22,704,373 words, with 53,310 lemmas. It provides measures for lemmas, morphemes and whole words. Due to the transparency of the Basque orthography and the lack of context dependent letters, only orthographic indices have been calculated. The program can be used online and it is freely available at the web page http://www.ehu.es/ehg/ehme/ which belongs to the Basque Language Institute.

The reference vocabulary corpus

The corpus has been updated from the Ereduzko Prosa Gaur [Contemporary Reference Prose] (EPG) corpus (Sarasola, Salaburu, Landa, & Zabaleta, 2007) of the Basque Language Institute (www.ei.ehu.es) at the University of the Basque Country. This corpus has been created out of the reference vocabulary of 287 published books and press from 2000 to 2006 in the whole Basque speaking territory, including France and Spain. Sources involve a broad range of disciplines, from history, literature, to science or medicine.

From the whole pool, only common Basque words were included, that is to say, true Basque lemmas. Proper names and words from other languages –except cognates- were excluded, so that of the 25.1 million words in this corpus, 22.7 were finally included in this database. To compute frequency measures, all the words extracted have been taken into account. Frequency measures have been computed in three ways. The raw measure consists of the number of repetitions of each word across the texts (token). Taking this measure as reference, the most frequent word appears 987,639 times, and the less frequent once, mean frequency being 60. Also frequency per million and Log. frequency have been obtained by dividing the total number of times by 22,7 and applying the Log. formula to the frequency per million, respectively. The utility of these measures is further explained in the *Word measures* section.

Before the words were incorporated into the database they were filtered to have the standard form, so that there are 377,795 different words (type), the number of letters ranging from 1 to 30. First, all the text data were copied into a computer, and words separated by a dash were considered in the database as one entry. Then lemmas and inflections were selected. This process was carried out using the automatic lemmatizer *Kapsula.* This program analyzes word letters entered in rows, detecting repeated letter patterns and splitting recurrent probabilities among each other. As a result of this parsing procedure, the program counts repeated structures that match with a minimal unit in the row (katuari, katuare, katuzale, would all match the minimum recurrent unit katu) to count for lemma frequencies, and whole forms for word frequencies. The proportion of lemmas is 15%, and the proportion of whole morphemic words is 85% from the total amount of word types.

The database includes nouns, verbs and adverbs in all derivational and inflectional forms. Proportion of nouns, verbs, adjectives, and adverbs implies 34.1%, 31.8%, 10.1%, 6.1% of the whole database, respectively. Grammatical functions that in other languages are driven by prepositions are developed here by morphemes (17% of the database). Hence, both lemma and whole morphemic words (lemma+morpheme) need to be counted for. From this 22.7 million word pool, only 53,310 words are lemmas, the rest are morphemic complex words. This provides a clue about the morphemic complexity of Basque language.

Inflections and derivations were categorized in the standard variety of the Basque language, so that there was no need to filter them. Thus, when the program encounters a word, it extracts the corresponding lemma and the morphological information, case, number, inflection and so on. The database is presented in a Basque language interface with a menu for each of the main available statistics (see Figure 1).

Insert Figure 1 around here

There is a link to the menu "From data to words/*Datuetatik hitzetara*" and another one for the menu "From words to data/ *Hitzetatik datuetara*". The difference is that the last one includes a folder to enter a list of words, and after the selection of the required statistics an output file is created with the measures corresponding to each of the words entered. In the two menus, the experimenter has the possibility to organize the output ordered by values or by alphabet. This can be done by pressing the button on the right of the menu, after making the selection of the measures. The steps to go from data to words are the following: 1) Click on the left button of the criteria we want to work with, and enter the ranges of the measures for the words we want on the spaces that appear on the right, 2) On the right side of the screen, below the spaces, select the order type (by alphabet or by frequency) and press the button "search/*bilatu*". The steps to get data from a set of words are the following ones: a) Copy and paste a list on the "Word list/*Hitz zerrenda"* folder at the left, or go to the second folder "Upload file/*Fitxategia igo*" and upload a .txt file pressing "Choose file/*Hautatu fitxategia*", b) Select the criteria for the words entered and the order type. The program will provide us with a .txt file with the words and criteria we asked for.

Available statistics

When the program starts, the user will see on the top of the screen three main links that lead to the pages that report the relevant values for all the statistics to work with. In the page "Data/*Datuak"* we can get the raw data for each of the measures. In the page "From data to words/*Datuetatik hitzetara"* we will have the maximum and minimum values for each of the measures. There we can see all the available statistics in four folders. This is so because for each word four main indices were computed: word measures, neighborhood measures, syllabic measures and morphological measures. In the page "From words to data/*Hitzetatik datuetara*" we will see the folder in which we can enter the words to get the previously settled measures. The four main folders are displayed as follows.

Word Measures

All the statistics in this category are measures computed on the basis of the EPG corpus. The first measure is frequency of use ("Frequency/*Maiztasuna"*), Frequency has shown to be one of the major measures that modulates access to the lexicon (see Coltheart, Davelaar, Jonasson, & Besner, 1977) and the principal output field of most databases (see also E-HITZ, Perea et al., 2006). High frequency words are easier to recognize than low frequency words because high frequency words are more strongly represented in the lexicon than low frequency words. This measure has shown to be one of the most powerful lexical factors that influence word reading in the most paradigmatic tasks, lexical decision (Balota & Chumbley, 1984) and word naming (Hino & Lupker, 2000), and it is provided in three modalities: raw value, frequency per million and Log. frequency. The raw value holds every word token from the corpus (no of repetitions for each type). The frequency per million is obtained dividing the raw measure by 22.7. This is a more comfortable way to work with frequency values. Log. frequency is calculated to provide 5 intervals that represent an exponential increase of the frequency magnitudes, instead of a linear scale. It has proved to be a valuable measure to compare frequency. measures by ranges (see Brysbaert et al., 2011)

Another important measure is the word´s orthographic structure. Recent research has shown that the consonant vowel structure of the word has an impact on the early processes involved in word recognition and reading (Berent & Marom, 2005). Research with different techniques such as letter search, or masked priming have shown that very early on processing, the visual system is sensitive to the orthographic structure of the word, and that the activation of the orthographic tier drives the process of word recognition (see Buchwald & Rapp, 2006). This hypothesis has been supported by neurological evidence; aphasic patients commit letter omission ad migration errors preserving the consonant-vowel structure of words (Caramazza, 1990). In the program, the orthographic structure can be extracted based on letter or syllable parameters. One option is the selection of number of letters (*Letra kopurua*) or/and number of syllables (*Silaba kopurua*). The program offers then the possibility to determine the vowel consonant structure and the syllabic structure. If a specific structure is entered in the field (e.g., Capital-Vowel-Capital-Vowel, CVCV, in case we want four letter words) the program will search and count all the words that match this criteria. A % can be used in any position, for the program to be flexible in the search (CV%, will provide with all the words in the corpus that begin with this structure being flexible in the rest), or a _ sign if the flexibility applies just to a specific position (CV_, will search for all three letter words that begin with this structure). The same possibilities are offered for specific syllables (KA-TU, KA%, KA__). This is an important option if we take into account the impact of the syllabic units in word recognition (see *Syllabic measures* section). The next option refers to the letter repetition constraint in the word (1, letter repeated, 0, no letter repeated). For example, if 1 is entered in the field, the program will extract all the words with repeated letters in the corpus. At the bottom of the screen, there are two other alternative measure options (*Bestelakoak*). "Word info/*Hitza bera"* offers the possibility to extract words being flexible in one part or position (%z, searches for all the words that end with z, for example, whereas _z, searches for all the words of two letters that end with z). Take into account that both first and last letters in a word act as anchor points for orthographic coding and word identification (Whitney, 2001). The measure "Distinctive orthographic point/*Bereiztasun puntu ortografikoa"* refers to the position at the word that makes it discriminative from other words that share the same letters at the beginning (e.g., kat.u, kat.egoria, the discriminability point would be 3), which is a relevant factor that influences word reading (Miller, Juhasz, & Rayner, 2006). This value goes from 1 to 23.

Neighborhood Statistics

This field provides information about the type and distribution of neighbors. The first one is the standard measure of orthographic neighborhood size, *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 et al., 1977). This measure has proven to influence word recognition in terms of reading times, reading errors and eye movements (Perea & Pollatsek, 1998). Recent evidence has shown that not only substitution neighbors, but also other types of neighbors can have an impact on reading (Acha & Perea, 2008; Davis et al., 2009). Due to this, the concept of neighborhood has been extended to include other types of measures. All of them are included in this section. The first option refers to substitution neighbors ("A change in one letter/*Letra bat aldatuz*"), and offers the possibility to select two indices: "Number of neighbors/*Auzokideen kopurua"*, and "Number of higher frequency neighbors/*Maiztasun handiagoko auzokideen kopurua*". The same can be done with deletion neighbors, formed by deleting one letter in the word at any position ("One letter deletion/*Letra bat kenduz"*), addition neighbors, formed by adding one letter to the word at any position ("One letter addition/*Letra bat gehituz"*), transposition neighbors, formed by transposing two letters in the word ("Two letter transposition/*Bi letra transposatuz"*), or all ("*Denera"*). For each type of neighborhood measure the left menu informs us about the minimum and maximum value of N and the minimum and maximum value of the N frequency range (that is, the number of neighbors or a certain word, and the minimum and maximum frequency values extracted from the words that constitute the N pool). In order to know, not only the amount of neighbors classified by type but also the corresponding words, the user needs to enter the word list in the "From Word to data" sheet, get the complete data result (*Xehetasun guztiak*) and click on the arrow in the upper centre of the web, above the output list. The program will automatically create a WordPad document in which all the neighbor words are included.

Syllabic Measures

One of the basic units in word recognition and production apart from the letter is the syllable. This has become an important unit of research in syllabic languages, particularly those in which the percentage of multisyllabic words is high (the proportion of polysyllabic words is much higher in Basque and Spanish than in English for example, see Carreiras & Perea, 2002). Syllables are important units of activation in word recognition; particularly the first syllable of the word. Carreiras, Alvarez, and De Vega (1993; see also Perea & Carreiras, 1998) tested the role of the syllable as a sublexical unit in word recognition in Spanish, using the single presentation lexical decision task. They used words that began either with a high or a low frequency syllable. Words with a low frequency first syllable were identified faster than words with a high frequency first syllable. Carreiras and Perea (2002) found that frequent syllabic primes (alto-ALGA) inhibited the recognition of the target compared to control syllabic primes (esto-ALGA), but also that primes that shared the syllabic structure of the target (zo.ta-ZO.CO) produced facilitation with respect to primes that did not share it (ziel-ZO.CO). From these experiments one can conclude that syllable frequency -particularly the first syllable- is an important sub-lexical unit that operates at a pre-lexical level (Álvarez, Carreiras, & Taft, 2001; Carreiras & Perea, 2002, 2004). Due to this fact, statistics related to the word´s number of syllables and the word´s syllabic structure, are provided in the database. Some of the measures in the first folder described above, allow the researcher to obtain certain measures about the orthographic syllabification of words, but some other measure possibilities are offered in this folder. More specifically, the statistics in this category allow selecting words with a certain syllable, bigram, or trigram in the position required from a range of letters offered (1 to 14). To do so, the user can go to "Syllables and groupings/*Silabak eta multzoak"*, and select the left button of the measure wanted: syllable, bigram or trigram (*Silabak*, *Letra bikoteak*, *Letra hirukoteak*, respectively), and enter the letters required on the folder that appears on the right of the screen for this purpose –these sublexical properties or words also influence the speed of processing (see Grainger, 1990). Bigram and trigram raw frequencies are created by counting all bigram in all positional combinations in all tokens. Syllable raw frequency is created the same way applying a syllabic parsing procedure. Positional frequencies are related by counting the same combinations by type (katu, 1 count, kale, 1 count for first syllable position "ka"). To this purpose, the field "Placement/*Kokapena*" offers the option to select the number that refers to the position of the letters entered in the word (e.g., *ka* in the 1st position). If there is flexibility about the position of the selected syllable, bigram or trigram across the word, the option "Anyone/*edozein"* should be selected.

Morphological Measures

Research on morphological complexity has revealed that the morphological properties of the language have an impact on the way words are processed, both in terms of internalization of regularities. Regular structures in the language, such as morphemes, are stored and retrieved easily during language acquisition (Treiman & Zukowski, 1991) and activated later on as autonomous units in word production and recognition (Holopainen, Ahonen, & Lyytinen, 2002). In fact, there is converging evidence about morphemes being regular units automatically identified in morphological complex languages, similarly to syllables (see Acha et al., 2010; Taft, 2004). Although most research has focused on the impact of whole word frequency in word processing (Giraudo & Grainger, 2000). Due to this fact, morphological measures have become an interesting unit for research, and a necessary measure to take into account when it comes to research on morphologically complex words. In this field the option on the top allows to select the lemma indices first ("Frequency of lema/*Lemaren maiztasuna*"). Here lemma refers to the root that can be attached to any morpheme, let´s say the base word. There is a possibility to select the three frequency measures here. As mentioned before, the program is designed to parse the lemma from the morpheme and count the token for the base word, calculating other measures afterwards. The option below is designed to settle a range of morphemes attached to the lemmas selected. This way, the program has fields to click in different grammatical categories: noun, adjective, verb, adverb, locative, counter, pronoun, determiner of question ("Morphology-Grammatical category/*Morfología-Kategoría gramatikala"*). This selection will lead to get specific words: lemmas to which only certain type of morphemes have been attached and its frequencies. If the aim is to obtain all the morphemes that can be attached to a lemma, one can skip this folder. This way, the program will search for all the morphemes and morphemic possibilities for the lemma/s entered. The last option "Others/*Bestelakoak"* was designed to offer the option to be flexible in the type of lemma. Making a click on the button "Lema/*Lema bera"* allows entering either a % or a _ (see Figure 1). These options are designed for an exhaustive search of certain morphemic words that contain certain letters, being flexible in either a part of the lemma or a certain position of the lemma, respectively. The main difference between the option "Lema info/*Lema bera*" in this folder, and the option "Frequency/*Maiztasuna*" in the word folder is that the "Lema" option is designed to obtain and manipulate frequency statistics for lemmas and morphemes, whereas the "Word" option searches for, and provides with whole word frequencies only.

Definition of Fields

The database is designed to enter fields in an additive way. The user can go to each of the folders and make a click on the measures on the left, so that the spaces to enter the ranges for each measure appear on the right. The user is free to select one or all of the measures in all folders. On the right, the spaces to determine ranges will appear one below the other one, following the selection order. In the end, the user will have a column on the right, with all the measures selected, and their respective ranges. After doing so, the using can press "Find/*Bilatu*", and a .txt file will show up, with a box in which all the words fitting the selection criteria appear in the column of the left, and with the concrete value of each measure required on the following columns to the right.

Output

There are two ways to extract information in the database. The user can enter the criteria for each measure as we mentioned previously in the "From data to words/*Datuetatik hitzetara*" link, and finally click "Find/*Bilatu*" to get the output file, which can be saved either as a .txt file of as an .xls file. However there is the possibility to do the same in the "From words to data/*Hitzetatik datuetara*" link, so that a list of words is either uploaded or pasted directly from a .txt file, and after making a click in "Find/*Bilatu*", a new window pops up with the previously required statistics presented by column.

Index Comparisons and Validity

A way to test any tool is to correlate it with the measures of another similar tool. In this case, we had the E-HITZ (Perea et al., 2006) a recent and commonly used database in psycholinguistic research on Basque language. First, we examined reliability by correlating both lexical and sub-lexical measures. Both databases showed very high correlations for both Log. frequency, *r*(5721) = 0.97, *p* = .001, and Neighborhood size (N) measures, *r*(5721) = 0,89, *p* = .001. Correlations were equally high for First syllable frequency, *r*(258) = 0.97, *p* = . 001, and Mean bigram frequency, *r*(46) = 0.97, *p* = .001.

We also examined the validity of the corpus comparing the effects of two lexical measures (Word frequency and N) from the EHME and E-HITZ databases in a lexical decision task. The main reason of doing so is that many researchers and grad students rely and have used E-HITZ to find frequency and N measures until now. The aim of the behavioral study was to examine whether the measures in EHME and E-HITZ were equally predictive of the obtained reaction times.

Method

Participants.

Thirty participants at the University of the Basque Country took part voluntarily in the experiment. All participants reported being native speakers of Basque with normal or corrected-to-normal vision.

Materials.

A set of 60 six-letter Basque words was selected for the experiment. In this set we selected 15 low frequency words (mean Log. frequency 0.8 in both databases) and 15 high frequency words (mean Log. frequency 2.1 in both databases), in addition to 15 low Neighborhood size words, and 15 high Neighborhood size words (these 15 words also had at least one higher frequency neighbor, Mean HFN = 2 and 1 for EHME and E-HITZ, respectively). Words were all paired in length and bigram frequency, and were represented with different Word frequency and Neighborhood values in E-HITZ and EHME. The differences between measures of the two databases were not significant for Log. frequency, *t*(29) = 0.047, *p* = .96, MSE = 0.028; but they were for the N measure, *t*(29) = 5.44, *p* < .001, *MSE* = 0.75. With respect to the HFN measure, no significant difference was found between the E-HITZ and the EHME list, *t*(29) = 1,94, *p* = .07, *MSE* = 0.29. The respective Log. frequency and N measures for each word are exposed in Table 1.

Insert Table 1 around here

For the purposes of the lexical decision task, we created 60 nonwords by replacing two to four letters of the target words. For example, from the high frequency word *aterki* the nonword *iferki* was created, from the low frequency word *jantzi* the nonword *fartzi* was created. We also controlled for the Neighborhood size of nonwords across conditions (*M* = 0.2 and 0.4 for nonwords paired with low and high frequency words, respectively; and *M* = 0.4 and 0.5 for nonwords paired with low and high N words, respectively).

Procedure

Participants were tested individually in a quiet room. The experiment was run using DMDX (Forster & Forster, 2003). Reaction times were measured form target onset until the participant's response. On each trial, a cross signal was presented for 500 ms in the centre of the screen. Next, a lowercase target was displayed and remained on the screen until the response. Participants were instructed to press one of two buttons on the keyboard to indicate whether the uppercase letter string was a legitimate Spanish word or not ("m" for yes and "z" for no). Participants were instructed to make this decision as quickly and as accurately as possible. Each participant received a different order of trials. Each participant received a total of 20 practice trials (with the same manipulations as in the experimental trials) prior to the experimental trials. Each session lasted approximately 15 min.

Results

Reaction times of 30 adult grade students of the University of the Basque Country (mean age 20) showed a significant effect of Word frequency, *t*(29) = 10.57, *p* = .003, (672 ms and 749 ms, for high and low frequency words, respectively) and Neighborhood size, *t*(29) = 4.57, *p* = .040 (831 ms and 744 ms for words with low N and for words with high N, respectively). Measures in the two databases showed similar and significant correlations with reaction times, for Log. frequency, *r*(28) = 0.52, *p* = .003, and *r*(28) = .57, *p* = .001; though not for N, *r*(28) = 0.080, *p* = .54, and *r*(28) = 0.019, *p* = .54, in EHME and EHITZ, respectively.

Finally, we conducted a multiple regression analysis to examine to what extent Word Frequency and Neighborhood size were predictive of the obtained reaction times in each database. To that aim, we entered the Log. frequency and N values of each of the 60 words used in the experiment as predictors, and we did the regression first with the EHME and then with the E-HITZ values. The regression analyses showed that both databases could predict the pattern of reaction times similarly, Frequency being the only reliable predictor. However, due to the big range of frequencies in the EHME a greater pool of words and subjects should be required for more adjusted fits in the regression analysis. Van Heuven, Mandera, Keuleers, and Brysbaert (2013) have very recently proposed a solution to find realistic and comparable Frequency measures, particularly when it comes to compare corpus of different sizes. This solution is the Zipf-scale, a Log. frequency scale that provides with values from 1 to 7 and allows selecting word from low to high frequency ranges in an intuitive and easy way (4 would be the point dividing low and high frequency words). To obtain a more exact picture of our databases predictability, we applied the formula provided by the authors to the Log. 10 freq per million. We conducted the same regression analysis entering the Zipf-scale frequency and N values as predictors. As expected the predictive value of the Zipf frequency was much greater than the Frequency per million for both databases. This shows that the Zipf value is a more adjusted and reliable frequency scale. Again, this was not so for the N measure in any of the databases. So far, it is not surprising to find inconsistent results with the N measure in the lexical decision task (see Acha & Perea, 2008) due to the fact that other lexical factors (such as neighborhood frequency) can have an impact on reaction times. The parameter estimates and distribution of the data in the regression models are shown in Table 2 and Figure 2, respectively.

Table 2 around here

Figure 2 around here

As we expected, measures in both databases show a high correlation and similarly account for two main behavioral effects highly replicated in the literature. Yet, the new database currently provides additional possibilities to manipulate these measures taking into account both recent requirements of psycholinguistic researchers in Basque, and the distinctive nature of this language.

Conclusion

This new Basque database provides with reliable frequency measures for whole morphologically complex words, as well as for lemmas and morphemes in isolation. In addition it offers information about other sensitive measures that influence word processing, such as neighborhood (N) and neighborhood frequency. An advantage of having frequency measures from a wide pool of words ensures a reliable control of lexical factors in psycholinguistic experiments, where this measure is usually manipulated or partialled out. In addition, the same criteria can be controlled both for lemmas and for morphemes, something essential to research on morphemic complex languages such as Basque. Indeed, recent experiments highlight the role of the frequency and length of morphemes in the process of the internalization of morphemic words (Taft, 2004). Thus, this information is essential for those researchers who examine the role of morphemes in word recognition and reading. Another important issue is that the experimenter can not only extract the desired measures from a list of words entered in the database, as in previously designed databases, but also get a list of words that fit some criteria once these are entered in the fields designed for this purpose. In sum, the database overcomes the limitations observed in previous databases, and provides experimenters with a complete and reliable tool for linguistic and psycholinguistic research on Basque language.

References

Acha, J., Laka, I., & Perea, M. (2010). Reading development in agglutinative languages: Evidence with beginning, intermediate and adult Basque readers. *Journal of Experimental Child Psychology, 105*, 359–375. http://dx.doi.org/10.1016/j.jecp.2009.10.008

Acha, J., & Perea, M. (2008). The effect of neighborhood frequency in reading: Evidence with transposed-letter neighbors. *Cognition, 108,* 290–300. http://dx.doi.org/10.1016/j.cognition.2008.02.006

Alvarez, C. J., Carreiras, M., & Taft, M. (2001). Syllables and morphemes: Contrasting frequency effects in Spanish. *Journal of Experimental Psychology*: Learning, *Memory and Cognition, 27*, 545–555. http://dx.doi.org/10.1037//0278-7393.27.2.545

Azkarate, M. (1993). Basque compound nouns and generative morphology: Some data. In Ortiz de Urbina, J., & Hualde, J. I., (Eds.), *Generative studies in Basque linguisstics.* Amsterdam, Philadelphia: John Benjamins.

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. http://dx.doi.org/10.1037//0096-1523.10.3.340

Berent, I., & Marom, M. (2005). The skeletal structure of printed words: Evidence from the Stroop task. *Journal of Experimental Psychology: Human Perception & Performance, 31*, 328–338. http://dx.doi.org/10.1037/0096-1523.31.2.328

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, A., & Böhl, A. (2011). The word frequency effect. *Experimental Psychology, 58,* 412–424. http://dx.doi.org/10.1027/1618-3169/a000123

Buchwald, A., & Rapp, B. (2006). Consonants and vowels in orthographic representation. *Cognitive Neuropsychology, 23*, 308–337. http://dx.doi.org/10.1080/02643290442000527

Caramazza, A. (1990). The structure of graphemic representations. *Cognition, 37*, 243–297. http://dx.doi.org/10.1016/0010-0277(90)90047-N

Carreiras, M., Alvarez, C. J., & de Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language, 32,* 766–780. http://dx.doi.org/10.1006/jmla.1993.1038

Carreiras, M., Duñabeitia J. A., Vergara M., de la Cruz-Pavia I., & Laka I. (2010). Subject relative clauses are not universally easier to process: Evidence from Basque. *Cognition, 115,* 79–92. http://dx.doi.org/10.1016/j.cognition.2009.11.012

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. http://dx.doi.org/10.1037//0096-1523.28.5.1228

Carreiras, M., & Perea, M. (2004). Naming pseudowords in Spanish: Effects of syllable frequency. *Brain & Language, 90*, 393–400. http://dx.doi.org/10.1016/j.bandl.2003.12.003

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, NY: Academic Press.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37*, 65–70. http://dx.doi.org/10.3758/BF03206399

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. http://dx.doi.org/10.3758/BF03192738

Davis, C. J., Perea, M., & Acha, J. (2009). Re(de)fining the orthographic neighbourhood: The role of addition and deletion neighbors in lexical decision and reading. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 1550–1570. http://dx.doi.org//10.1037/a0014253

De Rijk, R. (2007). *Standard Basque, a progressive grammar.* Cambridge, MA: MIT Press.

Dixon, R. M. W. (1994). *Ergativity, Cambridge studies in linguistics 69.* Cambrige, UK: Cambridge University Press.

Erdozia, K., Laka, I., Mestres-Misse, A., & Rodriguez-Fornells, A. (2009). Syntactic complexity and ambiguity resolution in a free word-order language: Behavioral and electrophysiological evidences from Basque. *Brain and Language, 109*, 1–17. http://dx.doi.org/10.1016/j.bandl.2008.12.003

Forster, K.I., & Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 16 –124.

Giraudo, H., & Grainger, J. (2000). Effects of prime word frequency and cumulative root frequency in masked morphological priming. *Language and Cognitive Processes, 15*, 421–444. http://dx.doi.org/10.1080/01690960050119652

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29*, 228–244. http://dx.doi.org/10.1016/0749-596X(90)90074-A

Hino, Y., & Lupker, S. J. (2000). Effects of Word frequency and spelling to sound Regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 166–183. http://dx.doi.org/10.1037//0096-1523.26.1.166

Holopainen, L., Ahonen, T., & Lyytinen, H. (2002). The role of reading by analogy in first grade Finnish readers. *Scandinavian Journal of Educational Research, 46,* 83–98. http://dx.doi.org/10.1080/00313830120115624

Hualde, J. I., & Ortiz de Urbina, J. (Eds.) (2003). *A Grammar of Basque.* New York, NY: Mouton de Gruyter. ISBN: 3 11 017683.

Laka, I. (1996). *A brief grammar of Euskara, the Basque language*. Vitoria-Gasteiz, Spain: Universidad del País Vasco/Euskal Herriko Unibertsitatea. Retrieved from http://www.ehu.es/grammar.

Laka, I. (2006). "Deriving split-ergativity in the progressive: The case of Basque". In Alana Johns, Diane Massam & Juvenal Ndayuragije (Eds.) *Ergativity: Emerging Issues* (pp. 173–195). Dordrecht, Berlin: Springer.

Laka, I., & Korostola, L. E. (2001). Aphasia manifestations in Basque. *Journal of Neurolinguistics, 14*, 133–157. http://dx.doi.org/10.1016/S0911-6044(01)00012-4

Miller, B., Juhasz, B. J., & Rayner, K. (2006). The orthographic uniqueness point and eye movements during reading. *British Journal of Psychology*, *97*, 191–216. http://dx.doi.org/10.1348/000712605X66845

Perea, M., & Carreiras, M. (1998). Effects of syllable frequency and syllable neighborhood frequency in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 134–144. http://dx.doi.org/10.1037//0096-1523.24.1.134

Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 767–779. http://dx.doi.org/10.1037//0096-1523.24.3.767

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. http://dx.doi.org/10.3758/BF03193893

Landa, J., Sarasola, I., & Salaburu, P. (2010). *Euskal Hiztegiaren Maiztasun Egitura (EHME)*. *Euskal Herriko Unibertsitatea* [Dictionary of frequency structures in Basque. University of the Basque Country]. Bilbao, Spain: Euskara Institutoa.

Sarasola, I., Salaburu, P., Landa, J., & Zabaleta, J. (2007). *Ereduzko Prosa Gaur (EPG)*. *Euskal Herriko Unibertsitatea* [Current prototypical prose. University of the Basque Country]. Bilbao, Spain: Euskara Institutoa.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology, 57*, 745–765. http://dx.doi.org/10.1080/02724980343000477

Treiman, R., & Zukowski, A. (1991). Levels of phonological awareness. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy. A tribute to Isabelle Y. Liberman* (pp. 67–83). Hillsdale, NJ: Erlbaum.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2013). Subtlex-UK: A new and improved word frequency database for British English. The *Quarterly Journal of Experimental Psychology (just published),* 1–36.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin and Review, 8,* 221–243. http://dx.doi.org/10.3758/BF03196158

Zawiszewski, A., Gutierrez, E., Fernandez, B., & Laka, I. (2011). Language distance and non-native syntactic processing: Evidence from event-related potentials. *Bilingualism: Language and Cognition, 14,* 400–411. http://dx.doi.org/10.1017/S1366728910000350

Table 1

EHME and EHITZ log 10 frequency and N values for words used in the lexical decision task

Frequency

EHME	E-HITZ	Word
0.62	0.60	aterki
0.63	0.60	bekoki
0.54	0.6	estura
0.69	0.6	zutoin
,0.66	0.66	izozki
0.70	0.77	artile
0.77	0.77	txango
0.78	0.83	jostun
0.89	0.86	katilu
0.93	0.86	otordu
1.02	1.00	usadio
1.11	1.05	belaun
1.04	1.06	orratz
0.63	1.12	buztin
1.08	1.18	behatz
2.51	2.58	liburu
2.1	2.25	ikasle
2.14	2.22	idazle
2.07	2.14	bihotz
2.13	2.11	jainko
1.82	2.01	esaldi
2.07	2.03	jantzi
2.09	2.00	osasun
2.07	1.94	urrats
2.04	1.96	bidaia
1.94	1.90	lekuko
2.43	1.93	iragan
1.83	1.93	otoitz
2.03	1.94	arreta
2.80	2.62	aukera

Neighbordhood

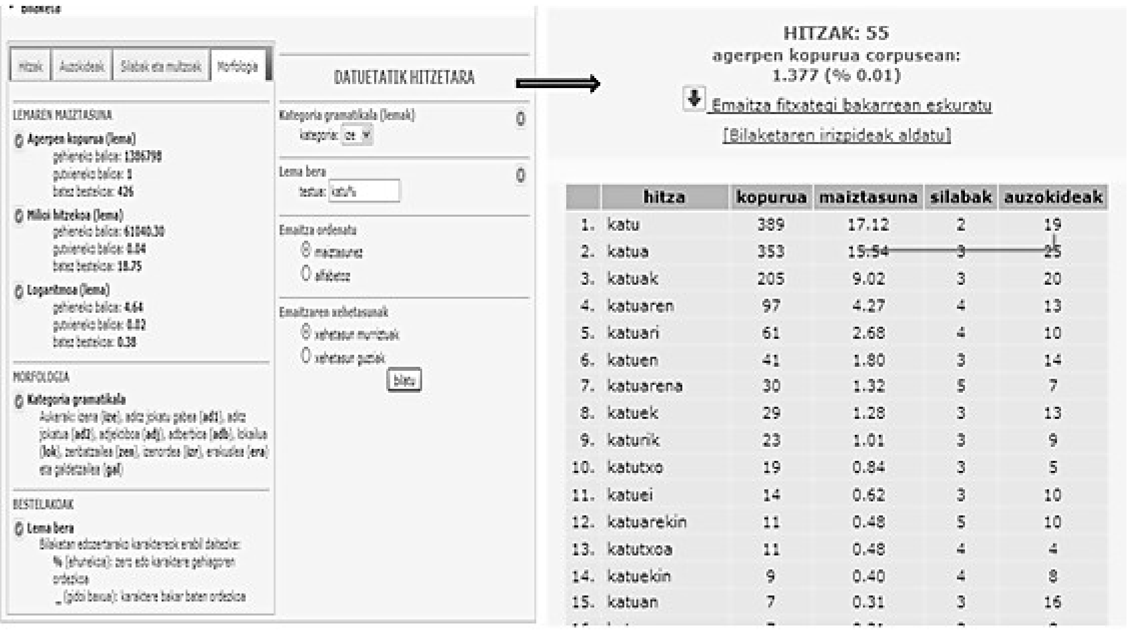
EHME	E-HITZ	Word
1	0	akeita
1	0	kresal
0	0	doilor
1	0	eurite
1	0	ihintz
1	0	zurgin
1	0	abuztu
1	0	atxilo
1	0	karobi
1	0	musker

1	0	ekidin
1	0	hiztun
1	0	txukun
2	1	pitxer
2	1	jangai
10	7	zentzu
20	7	arraio
17	6	arreta
13	6	arrano
15	6	galtza
9	6	zarata
9	5	dantza
10	6	sartze
7	7	zarama
12	7	bekatu
16	8	pareta
21	9	kantan
21	9	batera
22	11	erratu
11	7	berriz

Table 2

Regression parameter estimates using Word frequency and N from EHME and EHITZ databases on lexical decision times

	EHME			E-HITZ		
	β typified	*t*	*p*	β typified	*t*	*p*
Freq. per million	–.29	–2.23	.020	–.38	–3.07	.003
N	.005	0.03	.97	.112	0.88	.38
Zipf Frequency	–.65	–5.97	.000	–.68	–.677	.000
N	.120	1.10	.27	.185	1.85	.09



Hitzak | Auzokideak | Silabak eta multzoak | Morfologia

LEMAREN MAIZTASUNA

○ **Agerpen kopurua (lema)**
gehieneko balioa: **1386798**
gutxieneko balioa: **1**
batez bestekoa: **426**

○ **Milioi hitzekoa (lema)**
gehieneko balioa: **61040.30**
gutxieneko balioa: **0.04**
batez bestekoa: **18.75**

○ **Logaritmoa (lema)**
gehieneko balioa: **4.64**
gutxieneko balioa: **0.02**
batez bestekoa: **0.38**

MORFOLOGIA

○ **Kategoria gramatikala**
Aukerak: izena (**ize**), aditz jokatu gabea (**ad1**), aditz jokatua (**ad2**), adjektiboa (**adj**), adberbioa (**adb**), lokailua (**lok**), zenbatzailea (**zen**), izenordea (**izr**), erakuslea (**era**) eta galdetzailea (**gal**)

BESTELAKOAK

○ **Lema bera**
Bilaketan edozertarako karaktereak erabil daitezke:
% (ehunekoa): zero edo karaktere gehiagoren ordezkoa
_ (gidoi baxua): karaktere bakar baten ordezkoa

DATUETATIK HITZETARA

Kategoria gramatikala (lemak)
kategoria: ize

Lema bera
testua: katu%

Emaitza ordenatu
● maiztasunez
○ alfabetoz

Emaitzaren xehetasunak
● xehetasun murriztuak
○ xehetasun guztiak

bilatu

HITZAK: 55
agerpen kopurua corpusean:
1.377 (% 0.01)
Emaitza fitxategi bakarrean eskuratu
[Bilaketaren irizpideak aldatu]

	hitza	kopurua	maiztasuna	silabak	auzokideak
1.	katu	389	17.12	2	19
2.	katua	353	15.54	3	25
3.	katuak	205	9.02	3	20
4.	katuaren	97	4.27	4	13
5.	katuari	61	2.68	4	10
6.	katuen	41	1.80	3	14
7.	katuarena	30	1.32	5	7
8.	katuek	29	1.28	3	13
9.	katurik	23	1.01	3	9
10.	katutxo	19	0.84	3	5
11.	katuei	14	0.62	3	10
12.	katuarekin	11	0.48	5	10
13.	katutxoa	11	0.48	4	4
14.	katuekin	9	0.40	4	8
15.	katuan	7	0.31	3	16

Figure 1. Example of morpheme field menu and output list for some of the criteria

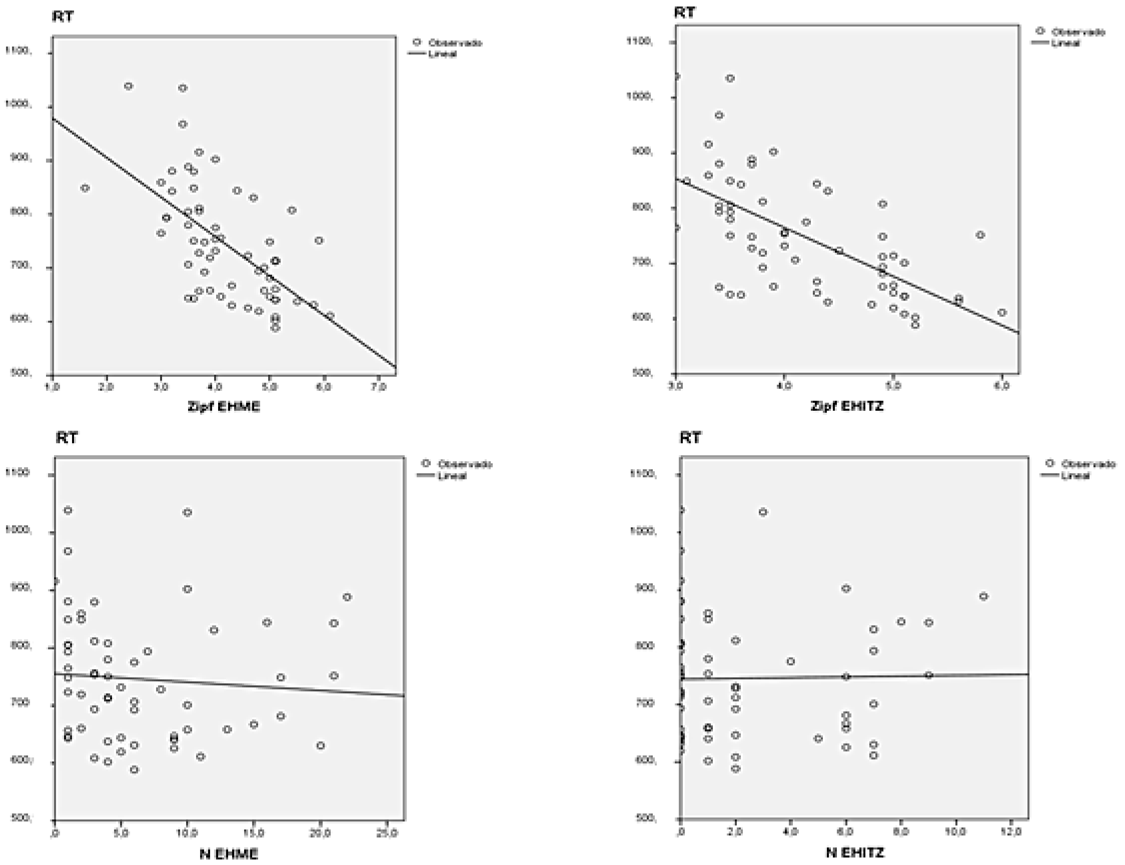


Figure 2. Distribution of EHME and E-HITZ reaction times in the linear regression model