How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy

Jose Aguasvivas [1,2*]

Manuel Carreiras [1,2,3]

Marc Brysbaert [4]

Paweł Mandera [4]

Emmanuel Keuleers [5]

Jon Andoni Duñabeitia [6,7]

[1] BCBL, Basque Center on Brain, Language and Cognition; San Sebastian, Spain

[2] Universidad del País Vasco/ Euskal Herriko Unibersitatea; San Sebastian, Spain

[3] Ikerbasque, Basque Foundation for Science; Bilbao, Spain

[4] Department of Experimental Psychology, Ghent University; Ghent, Belgium

[5] Department of Cognitive Science and Artificial Intelligence, Tillburg University; Tillburg, Netherlands

[6] Centro de Ciencia Cognitiva (C3), Universidad Nebrija; Madrid, Spain

[7] Department of Language and Culture, Faculty of Humanities, Social Sciences, and Education, The Arctic University of Norway; Tromsø, Norway

***Contact information**: Jose Aguasvivas j.aguasvivas@bcbl.eu

Basque Center on Cognition, Brain and Language; Paseo Mikeletegi 69, 2nd floor; 20009 Donostia/San Sebastian – Spain

Abstract

Vocabulary size seems to be affected by multiple factors, including those that belong to the properties of the words themselves and those that relate to the characteristics of the individuals assessing the words. In this study, we present results from a crowdsourced lexical decision megastudy in which more than 150,000 native speakers from around 20 Spanish-speaking countries performed a lexical decision task to 70 target word items selected from a list of about 45,000 Spanish words. We examined how demographic characteristics such as age, education level, and multilingualism affected participants' vocabulary size. Also, we explored how common factors related to words like frequency, length, and orthographic neighbourhood influenced the knowledge of a particular item. Results indicated important contributions of age to overall vocabulary size, with vocabulary size increasing in a logarithmic fashion with this factor. Furthermore, a contrast between monolingual and bilingual communities within Spain revealed no significant vocabulary size differences between the different communities. Additionally, we replicated the standard effects of the words' properties and their interactions, accurately accounting for the estimated knowledge of a particular word. These results highlight the value of crowdsourced approaches to uncover effects that are traditionally masked by small-sampled in-lab factorial experimental designs.

Keywords: Spanish lexical decision; crowdsourcing megastudy; vocabulary size; and ageing

How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy

The knowledge of a language's vocabulary is an essential aspect of language proficiency. This knowledge seems to be an important aspect of intelligence, with most general IQ scores including one or several distinct vocabulary measures (Bowles & Salthouse, 2008). However, the structure and size of vocabulary seem to differ considerably based on an individual's life experience, interests, skills, and age (Brysbaert, Stevens, Mandera, & Keuleers, 2016b; Keuleers, Stevens, Mandera, & Brysbaert, 2015; Kuperman & Van Dyke, 2013; Solomon & Howes, 1951). The heterogeneity of vocabulary across distinct contexts is the focus of the present paper. We build upon previous work to study the factors affecting the vocabulary size of Spanish speakers through a crowdsourced online lexical decision megastudy (Aguasvivas et al., 2018).

One simple way to measure vocabulary size is by presenting strings of letters and having the participant decide whether these represent an existent word (e.g., the Spanish word for book, *libro*) or not (e.g., the nonword *lirbo*). This procedure is commonly known as a lexical decision task (LDT; for an overview, see Kuperman & Van Dyke, 2013), and has been long used to study how different variables affect participant's lexical access and word recognition time (for an overview, see Balota, Yap, & Cortese, 2006). Thanks to the task, we know how word length, word frequency, concreteness, and orthographic neighbourhood size, among other properties, can affect the time required to recognise and retrieve a word from the lexicon (Andrews, 1997; Grainger, 1990).

Word properties are commonly obtained by analysing collections of naturally occurring written (or oral) language (Gierut & Dale, 2007). For example, to obtain a word's frequency, the appearance of that word within multiple sources is counted. Other properties, however, require participants to complete questionnaires asking about different subjective dimensions that cannot be automatically computed from corpora, and that may vary depending on participants' characteristics (e.g., valence, arousal, age of acquisition; Gierut & Dale, 2007). In this sense, Keuleers and Marelli (n.d.) distinguish between unelicited properties –those that can be obtained from linguistic resources using computational methods– and elicited properties that can be obtained directly from participants´ elicited behaviour.

Several lexical databases combining both elicited and unelicited word properties have been developed for various languages. In most cases, there exists more than one database per language. In Spanish, for instance, the most commonly used lexical databases include: BuscaPalabras based on books (Davis & Perea, 2005), ESPAL based on books, web sources, and movie subtitles (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013), and SUBTLEX-ESP based on movie subtitles (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011).

The source on which distributional measures for words are based can influence the expected results of LDT. For instance, the performance of younger adults is better predicted by frequencies obtained from internet sources (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010), while the frequencies of a corpus based on movie subtitles in the US, but not in the UK, better predicts the performance of US students (Brysbaert & New, 2009). There is not a unique corpus that can fully capture the heterogeneity of a language's vocabulary across different individuals. Due to this, Keuleers and Balota (2015) suggest using approaches where participants can assess word properties in conjunction with corpus information. Under this novel *crowdsourcing* approach, online platforms function as a vehicle for the assessment of properties from a vast number of raters.

The information about vocabulary knowledge can be further broadened using laboratory megastudies, that is, large-scale experiments involving hundreds or thousands of participants. There have been numerous efforts to create and analyse large word-processing datasets (for a list, see http://crr.ugent.be/programs-data/megastudy-data-available). Lexical decision megastudies have paved the way for measuring other factors influencing lexical access using more heterogeneous populations (Keuleers & Balota, 2015). Megastudies like this have been carried out in several languages, including American and British English (Balota et al., 2006; Keuleers, Lacey, Rastle, & Brysbaert, 2012), French (Ferrand et al., 2010), and Dutch (Brysbaert, Stevens, Mandera, & Keuleers, 2016a; Keuleers, Diependaele, & Brysbaert, 2010).

Perhaps the most relevant integration of crowdsourcing and a lexical decision megastudy is offered by Keuleers et al. (2015). By using an online platform, they tested around 300,000 native Dutch speakers on more than 53,000 words, presenting a randomly selected subset of 70 words per participant. Their findings not only confirmed previous statements that vocabulary increases as a function of age and education level (for a meta-analysis, see Verhaeghen, 2003), but also suggested that other variables, such as the number of foreign languages an individual knows, their L2 proficiency, and their geographic location (in this case Belgium or the Netherlands) were also factors affecting vocabulary size. Moreover, they introduced the concept of *word prevalence*, referring to the mean proportion of a population that knows a specific word (Keuleers et al., 2015). This variable served as a complement to word frequency and was an important predictor of reaction times in the other LDT studies (Brysbaert, Mandera, McCormick, & Keuleers, 2019; Brysbaert et al., 2016b).

Crowdsourced lexical decision megastudies have numerous advantages. First, they allow for massive data collection at a reduced cost by distributing the experiment through an online platform and providing alternative incentives to participants (e.g., sending scores via e-mail; see Dufau et al., 2011). Second, the effects of continuous variables (like frequency) can be treated as such without the need to categorise them (Keuleers et al., 2012). Third, the studies provide normative information on performance from a vast number of participants on many words (and nonwords). Fourth, virtual experiments can be run within the database to evaluate novel hypotheses or better control stimuli selection, and computational models of word recognition can be evaluated against the data (Stadthagen-Gonzalez, Imbault, Pérez Sánchez, & Brysbaert, 2017). Finally, the data from multiple megastudies can be combined to produce meta-megastudies, drawing inferences about language processing beyond the scope of a specific language (Myers, 2016).

Word accuracy as an indicator of vocabulary size

Vocabulary knowledge can be measured at different levels, ranging from being acquainted with a word's existence (word recognition) to comprehending its meaning and use in different contexts (semantic, morphological, and even syntactic processing). LDT and naming are tasks that tap into the former category, while picture naming tasks, overt definition or sentence completion tests fall into the latter. Despite this, the format in which a test measures vocabulary knowledge is thought to be interchangeable, given that they refer to the same underlying construct (Bowles & Salthouse, 2008). This assumption makes LDT, albeit incomplete in the broad sense of semantic access, a valid measure of word recognition and vocabulary size (Diependaele, Brysbaert, & Neri, 2012).

When people are visually presented with a stream of letters and a forced-choice task, a word identification and retrieval process is engaged (Katz et al., 2012). Various factors can alter this process. We can categorise these factors into those reflecting individual experiences, such as age, education level, multilingualism, among others (extrinsic factors); or those belonging to the words themselves, including their frequency of occurrence, the number of orthographic neighbours, and others (intrinsic factors). These are variables that tend to be controlled for or factored in lexical decision studies, but using massive data collections allow us to test them continuously (Stadthagen-Gonzalez et al., 2017).

So far, no previous attempt has been made to produce a crowdsourced lexical decision megastudy in Spanish, the second most used native language after Chinese (Ethnologue, 2016). The current study presents a detailed analysis of data obtained from more than 20 Spanish-speaking countries across the globe (Aguasvivas et al., 2018; data freely available at https://figshare.com/projects/SPALEX/29722). Hence, the purpose of this study is to examine how intrinsic and extrinsic factors affect Spanish vocabulary size and word knowledge. For the rest of this Introduction, we focus on detailing how LDT relates to vocabulary knowledge, outlining a selection of factors influencing this knowledge.

Extrinsic factors affecting LDT

Age. With time, individuals can encounter and learn novel words in both their native and other languages. Studies measuring the effect of age on vocabulary knowledge tend to conclude that, independently of the format used (e.g., multiple choice, production, lexical decision), vocabulary increases drastically throughout early adulthood, then flatten in middle-age, only to then decline gradually or hold steady through late adulthood (Bowles & Salthouse, 2008; McCabe, Roediger, McDaniel, Balota, & Hambrick, 2010; Singer, Verhaeghen, Ghisletta, Lindenberger, & Baltes, 2003; Singh-Manoux et al., 2012). Recent LDT megastudies suggest that vocabulary keeps increasing with age, and does not decline as previously thought (at least not in the participants that take part in the test), suggesting that age is one of the most relevant predictors of vocabulary size (see Brysbaert et al., 2016a). Furthermore, the effect of intrinsic properties such as frequency and age of acquisition seems to be mediated by age, with a decrease in the size of the effect as age increases (Davies, Birchenough, Arnell, Grimmond, & Houlson, 2017). Also, lexical decision response time appears to remain largely unaffected by age (Schröter & Schroeder, 2017). While slowing response times in other tasks is often attributed to an ageing-related decline in information processing capacities, it can, in fact, reflect increased information processing demands (Ramscar, Hendrix, Love, & Baayen, 2014; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014; Ramscar, Sun, Hendrix, & Baayen, 2017).

Education. Although commonly used as a control variable in vocabulary knowledge research, education exposes individuals to novel vocabulary in both common and specialised knowledge domains (Keuleers et al., 2015). In this regard, Tainturier et al. (1992) noted that the frequency effect is reduced in individuals with more years of education than in those with fewer years of schooling. They attribute these results to people with more education having higher chances of being exposed to lower frequency words. Kuperman and Van Dyke (2013) pointed out that this interaction between frequency and skill relies on the use of corpus word frequencies, which are especially based in the low-frequency range. When subjective measures of word occurrence are used, the skill-frequency interaction disappears. Likewise, accuracy in LDT seems to be affected by education, as individuals with a high education level can recognise words and discard non-words more accurately than those with lower education level (Kosmidis, Tsapkini, & Folia, 2006).

Geographic location. It is known that language varies across social and regional contexts, which is the subject of study of sociolinguistics and dialectology (Eisenstein, O'Connor, Smith, & Xing, 2010). These variations also suggest that vocabulary, albeit similar in size, might be composed of different words depending on the location of the speaker, as is the case with Latin-American versus Castilian Spanish (Aguasvivas et al., 2018). By using geotagged material, inferences can be drawn on lexical, syntactic, and semantic variations not only across countries but also within regions of the same country (Kulkarni, Perozzi, & Skiena, 2016). This is particularly interesting for countries like Spain, in which linguistic policies acknowledge the country's multilingual and multicultural character, allowing some communities to increase the presence of languages other than Spanish in compulsory education (Huguet, 2007). Despite this, there is scarce tradition of research on the linguistic aptitudes of individuals within these regions (Huguet, Lapresta, & Madariaga, 2008). For this study, we are interested in knowing whether Spanish vocabulary size is similar within these regions as compared to regions where both the educational and social context is limited to Spanish. Furthermore, we are interested in comparing Spanish across multiple Spanish-speaking countries.

Multilingualism. Before megastudies were run, small-scale studies comparing bilinguals and monolinguals on linguistic tasks suggested that bilinguals showed decreased lexical retrieval capacity (Portocarrero, Burright, & Donovick, 2007), less verbal fluency (Bialystok, Craik, & Luk, 2008), and greater interference in lexical decisions (Gollan & Acenas, 2004). They all pointed to disadvantages that arose due to (a) individuals dividing their word usage between the languages they know, and (b) multilinguals being exposed less to a specific language than a monolingual person (Gollan, Montoya, Cera, & Sandoval, 2008). However, contrary to these early findings (and researcher intuitions), Keuleers et al. (2015) found not only that L1 vocabulary size was larger in bilinguals, but that L1 vocabulary size increased with the number of languages the participants reported to know. This is a critical finding that deserves close attention, and the use of a parallel megastudy approach in a different language will allow us to test its replicability. Overall, Keuleers et al.'s conclusion regarding multilingualism and vocabulary size is that vocabulary in a language might be aided by the knowledge of other languages, mainly because the knowledge of extra languages gives people more diverse contexts in which to learn words. Given that many of these words are cognates in several languages (have the same form and meaning), knowing words in a second language is likely to increase knowledge of the same words in the native language. For instance, knowing the Spanish word *siesta* increases the English vocabulary as well. This line of argumentation fits well with recent evidence demonstrating the role of cognate words in the process of language learning (e.g., Casaponsa, Antón, Pérez, & Duñabeitia, 2015).

Intrinsic factors affecting LDT

Although an exhaustive evaluation of every intrinsic factor affecting LDT is beyond the scope of this study, we attempt to analyse how some of the most prominent factors in the literature impact word knowledge in Spanish. In this sense, we consider word frequency, length, and orthographic neighbourhood as the main factors of interest.

Word frequency. The word frequency effect is one of the most robust and well-documented effects of the word recognition literature (Brysbaert, Mandera, & Keuleers, 2018). It refers to the decrease in the latency of response (or response time) for high-frequency words - those that appear very commonly in a language- in contrast to low-frequency words, which occur less in a language. Murray and Forster (2004) describe the frequency effect as one of the most decisive factors controlling the time required to recognise a word pattern, with almost all the other factors only influencing the performance for a certain range of frequencies. The rationale behind this effect is that continuous exposure to a word in different contexts leads to a strengthening of the activation and connections of its representation, and therefore a reduction of the time required to access it (Brysbaert et al., 2018).

While the frequency of occurrence of a word relates to the chances of an individual being exposed to it, individual experiences can alter the effect in LDTs. For instance, the frequency effect appears to vary depending on the reading skill and age of an individual. In the former case, the effect is weaker for skilled readers than for less skilled readers, although, if frequencies are obtained using subjective ratings as a substitute of corpus frequencies, the effect equates across groups (Kuperman & Van Dyke, 2013). Conversely, the effect of frequency decreases with the age of the participant, although older participants, in general, become slower. The result is that older participants are relatively slower in their responses to high-frequency words (Brysbaert et al., 2019; Davies et al., 2017). In all, although the frequency effect seems to be very robust, it is susceptible to individual differences, and the way the frequencies are obtained can also influence the magnitude of the effect (see Dimitropoulou et al., 2010).

For this study, we tackle the question of how word frequency relates to vocabulary knowledge. The frequency measure used in this study was extracted for each word from the EsPal database using the Zipf scale (Duchon et al., 2013), which is roughly equivalent to the base 10 logarithm of the frequency per billion words and ranges from 1 to 7 (for a detailed description of the scale, see van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The higher the value in Zipf scale, the more frequent a word is seen in the corpus.

Orthographic neighbourhood size. The time required to recognise a printed word also seems to depend on the degree of orthographic similarity it has to other words in the language (Diependaele et al., 2012). In the traditional definition (Coltheart, Davelaar, Jonasson, & Besner, 1977), a word's orthographic neighbourhood (N) is the number of words that have the same length as that word, but that differ in exactly one letter (e.g., *cake* – *lake*). A higher value for the orthographic neighbourhood implies that a word has more similarity to existing words. A more recent definition (Yarkoni, Balota, & Yap, 2008), operationalises orthographic neighbourhood density as the average Levenshtein distance (Levenshtein, 1966) between a word and its 20 nearest orthographic neighbours (OLD20). Higher values in this measure indicate a sparser neighbourhood, as the average distance between the target words and its neighbours is larger.

The literature shows mixed results about the effect of orthographic neighbourhood size on word recognition, with some studies indicating a facilitatory effect and others suggesting an inhibitory effect or no effect at all (for reviews, see Andrews, 1997; Carreiras, Perea, & Grainger, 1997). Despite this, much of the LDT literature agrees that words with more neighbours are identified more rapidly and accurately than words with fewer neighbours (Pollatsek, Perea, & Binder, 1999). This variable also seems to be influenced by age, with children responding more accurately to words with many neighbours than those with fewer neighbours (Duñabeitia & Vidal-Abarca, 2008).

Length. The number of characters in a word can greatly influence the time required to recognise it, as the individual requires more grapheme-phoneme conversions during reading. Most studies have traditionally controlled for this variable instead of including it, which has led to an overshadowing of its possible effect on word recognition time and accuracy (González-Nosti, Barbón, Rodríguez-Ferreiro, & Cuetos, 2014). In this aspect, Acha and Perea (2008) compared beginner (children), intermediate, and adult readers in a Spanish LDT showing that, while the length effect for words was robust in children and disappeared in adults, the effect of the length of non-words followed the opposite pattern. They suggested that in a fully developed lexical system, access to known word representation occurs automatically while accessing unknown words or non-words requires letter-by-letter decoding (Acha & Perea, 2008).

Method

Participants

We collected data from May 12th, 2014 to December 19th, 2017 (see Figure 1). Up to that point, 209,351 participants had finished 282,576 tests by completing one (80.0%), two (14.1%), three (3.3%), or more sessions (2.6%). Most of the data (68.9%) were acquired during the first month of the experiment when a radio advertising campaign was run to attract the public's attention. Participants also had the option of publishing their results via social networks, which attracted new participants in a snow-ball sampling fashion. Additionally, before the experiment, participants were able to voluntarily provide information about their sex, age, country of origin, education level, handedness, number of known foreign languages, best foreign language, and geolocation information. The raw version of this data for native Spanish speakers is presented in the SPALEX database made available in Aguasvivas et al. (2018) and it can be retrieved from https://figshare.com/projects/SPALEX/29722.

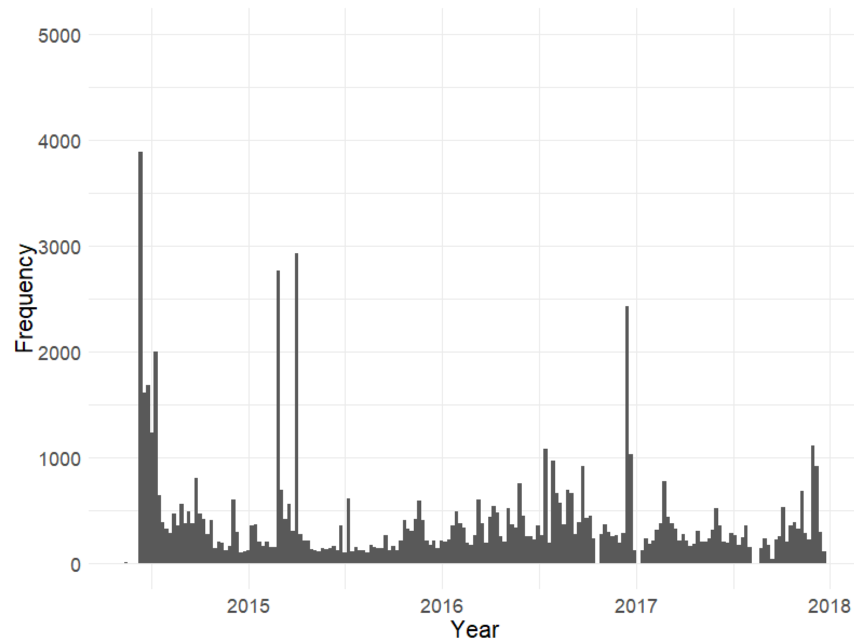

Figure 1. Frequency of participation per year. Each line represents a week. Participation in the year 2014 represented 73.77% of the data, while 2015 represented 9.20%, 2016 10.30%, and 2017 6.74% of the data. Gaps in the distribution of responses correspond with maintenance periods of the online platform.

Based on the country and native language information provided by the participants, we identified non-native speakers of Spanish (17.4% of the data) and discarded them for the current study, as the focus of this paper is on native Spanish speakers. After this, the sample was reduced to 169,628 participants from 19 Spanish-speaking countries who completed 227,665 experimental sessions in total. Out of these sessions, 34.9% were completed using a device other than a computer (mobile phone, tablet, etc.), indicating a high level of engagement of the participants through mobile platforms. We retained only the first session of each participant, reducing the amount of sessions to 169,628. Finally, we limited the age range of participants to keep it between 25 and 80 years, as an initial exploration of the histogram revealed scarce participation of individuals younger than 25 (0.6%) or above 80 (1.5%).

The final list included in the analysis consisted of 163,460 participants. Of these, 47.8% were females, while 0.9% of participants provided no gender information. Mean age was 45.8 (SD = 11.9). Regarding the country of origin, the majority of participants reported being born in Spain (49.3%), followed by Mexico (17.5%), Peru (10.5%), Argentina (6.1%), Colombia (5.9%), Chile (4.1%), and other countries from Latin-America (Bolivia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Honduras, Nicaragua, Panama, Paraguay, Dominican Republic, Uruguay, and Venezuela). This variable was recoded to separate native speakers from Latin-America and Spain. Education level was recoded into integer values (*secondary school,* the minimum mandatory education level = 2, *high school* = 3, *university degree* = 4, *master's degree* = 5, *PhD* = 6). Mean education level was 3.7 (SD=1.0), and only 1.2% of participants provided no education information. Handedness was also recoded into 1 (right-handed, 90.5% of the data) and 2 (left-handed, 8.5% of the data). We restricted the number of foreign languages to be between 0 and 8 (M=2.6, SD=1.40), as only less than 0.05% of participants reported knowing more than eight foreign languages. Participants reported 98 different best-known foreign languages, but we did not consider this variable for our analysis.

The geolocation was stored in the format of latitude and longitude and retrieved separately from the server. We only used the information from participants within Spain that were also present in our cleaned database. Using the reverse_geocoder module in Python (https://github.com/thampiman/reverse-geocoder), we obtained information about the city and region of these participants. This process was done offline, and further information such as postal code or street names were automatically discarded to protect the participant's identity.

Using only the geolocation information of participants within Spain, we identified those that were located in official bilingual communities (Basque Country, Catalonia, and Galicia). A group of participants living in official monolingual communities that matched the number of participants in the bilingual communities (Andalusia, Castile and Leon, Castile-La Mancha, Madrid, and Murcia) was also selected for comparison purposes. Furthermore, we limited the number of foreign languages reported by these participants to match monolingual and bilingual profiles. A total of 1,679 participants (885 bilinguals) were therefore extracted from the database and stored for a separate analysis.

Materials

Each experimental session consisted of 100 items presented randomly to each participant. The number of items per sessions was selected to ensure that the duration of each session would be approximately five minutes so that participants wouldn't be discouraged to participate. The items came from two pools of stimuli, namely words and nonwords. The words were selected from a pool of 45,389 Spanish words retrieved from the B-PAL (Davis & Perea, 2005) and the EsPal databases (Duchon et al., 2013) to account for both written and spoken corpora. The nonwords were obtained by feeding the word list to Wuggy (freely available at http://crr.ugent.be/programs-data/wuggy; see Keuleers & Brysbaert, 2010) to generate several potential nonword candidates for each word. From the resulting list, we selected a subset based on the candidate index produced by Wuggy. The final nonword list contained 56,855 items. Further information on the material, as well as on the task reliability, can be found in Aguasvivas et al. (2018).

Procedure

Participants were able to perform the task from their device by accessing the website of the experiment (http://vocabulario.bcbl.eu/). When first arriving on the website, participants saw a welcome screen with a button to begin the experiment. The instructions of the experiment were presented in Spanish and indicated to the participants that they would see 100 letter strings, with some of them representing real Spanish words and others representing made-up words. Their task was to indicate whether they knew the string or not by pressing either a 'YES' or 'NO' button on the phone/tablet or the 'F' and 'J' keys on their keyboard (see Figure 2). This part of the instructions was tailored depending on the device used. The task was not speeded nor did the instructions suggest that participants should respond as quickly as possible, so they could take all the time needed to respond to a word. Nevertheless, participants were warned that responding 'YES' to words that didn't exist in Spanish would result in a penalisation in their scores.

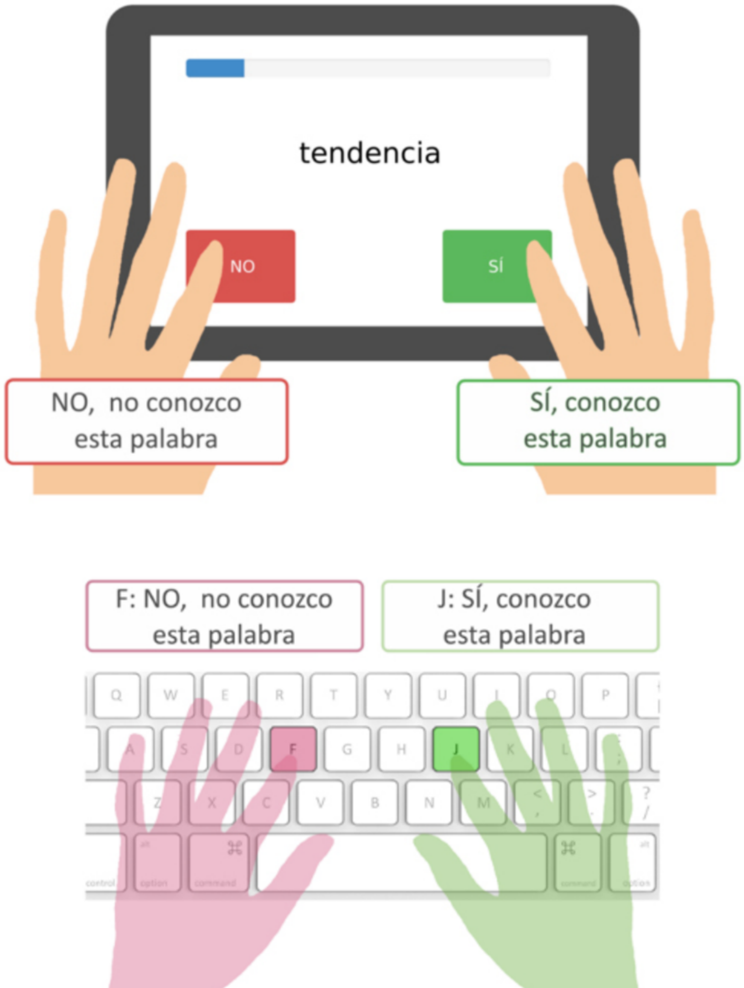

Figure 2. Experiment screen layout and key configurations for phone/tablets (top) and computers (bottom). The layout for the presentation of the word and progress bar was identical in all devices.

Before the beginning of the experimental session, each participant had the option to fill in the demographic questionnaire and provide their geolocation information voluntarily. Answering these questions was not required to proceed with the experiment, but participants not answering them were not included in the analyses. After the questionnaire screen, participants were instructed to place their fingers in the instructed position (buttons or keys) and press a button to begin the experiment. The stimuli were always presented in a vertically and horizontally centred position on the screen, and a blue progress bar on the top of the screen informed participants of their advancement through the experiment (see also Figure 2). Responses were automatically coded into correct and incorrect responses, and response time (RT) was recorded in milliseconds for each response. It is important to note that in Aguasvivas et al. (2018) we tested whether the 70/30 word to nonword ratio introduced bias in the accuracy scores by using the LD1NN algorithm (Keuleers & Brysbaert, 2011). The results indicated that if participants were to base their decisions only on the statistical characteristics of presented words and nonwords, they would be 2.6 times more likely to identify a stimulus as a word than as a nonword. Values from other studies range from 0.34 to 4.1, depending on how nonwords are created. We also tested the reliability of RT scores by using the split-half method, obtaining Spearman-Brown corrected reliability of 0.92 for words and 0.91 for nonwords.

When participants had responded to all stimuli, they were able to see their score, which was calculated by subtracting the percentage of incorrectly accepted nonwords from the percentage of correctly recognised words. This screen also allowed participants to examine their answers, redo the experiment, or share their answers via Facebook, Twitter, or email. When clicking on each word, participants could either see the definition (e.g., https://dle.rae.es/?id=9AwuYaT for the Spanish word *ciencia*, which means science) or report the word as non-existent in Spanish.

Results

We calculated a score for each participant by subtracting the percentage of false alarms (incorrectly accepted nonwords) from the percentage of hits (correctly accepted words). This score could range from -100 (all nonwords accepted, all words rejected) to 100 (all nonwords rejected, all words accepted). We identified participants with scores below or above 1.5 times the interquartile range as outliers and removed them from further analyses (2.4% of the data). After this, a list of 157,912 participants remained. Following Keuleers et al. (2015), we used the corrected score of each participant as a proxy for vocabulary size and average accuracy per word as a measure of word knowledge. These two variables are the main focus of this study. Figure 3 shows the mean accuracy and RTs for each bin of two trials. While accuracy seemed to stabilise after a few trials, RT diminished as the experiment progressed.

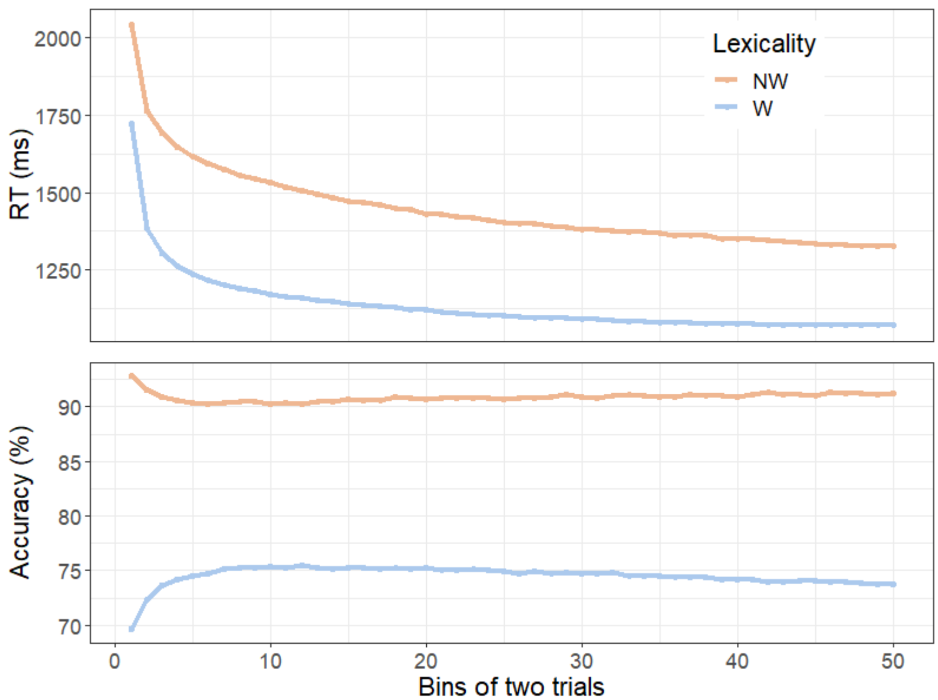

Figure 3. Average RT for correct responses (top) and average accuracy (bottom) per trial bin. Each bin represents two trials. RTs above and below 1.5 times the interquartile range were identified as outliers and removed from the calculation. RT = response time; NW = non-words; W = words.

Variables affecting vocabulary knowledge

Extrinsic effects. To test for the extrinsic effects on vocabulary size, we used a multiple regression that included the score of each participant as the outcome, and as predictors: age (log transformed) treated as a continuous variable, education level as a factor with five levels (secondary school, high school, major, master, and PhD), location as a factor with two levels (native speakers from Latin-America, and native speakers from Spain), number of foreign languages as a continuous variable, and gender as a factor with two levels (male and female).

Due to the amount of observations and terms in the regression, we opted to run a first model including all factors and their two- and three-way interactions. We then selected only those terms that accounted for more than 0.5% of the variance. After the first iteration, only the main effects remained. Table 1 shows the results of the final model for the score of the participants, which accounted for 28% of the variance in scores (R^2 = 0.278, F = 4851.914, $p <$ 0.001, 95% CI [0.27, 0.28]. While most of the factors were significant in the initial model, the surviving terms after applying the criteria were age (F = 34751.097, $p < 0.001$, $\eta^2 = 0.164$, 95% CI [0.161, 0.168]), geographic location (F =17142.431, $p < 0.001$, $\eta^2 = 0.081$, 95% CI [0.079, 0.083]), education level (F = 828.432, $p < 0.001$, $\eta^2 = 0.016$, 95% CI [0.015, 0.017]), reported number of foreign languages (F = 1103.272, $p < 0.001$, $\eta^2 = 0.005$, 95% CI [0.005, 0.006]), and gender (F = 929.117, $p < 0.001$, $\eta^2 = 0.004$, 95% CI [0.004, 0.005]).

Table 1. *Analysis of variance table showing effects of predictors on vocabulary size*

Term	df	SS	F	*p*	η^2	95% CI [LOW, HIGH]
Log(Age)	1	340.424	34751.097	<0.001	0.164	[0.161, 0.168]
Location	1	167.929	17142.431	<0.001	0.081	[0.079, 0.083]
Education	4	32.462	828.432	<0.001	0.016	[0.015, 0.017]
No. foreign lang.	1	10.808	1103.272	<0.001	0.005	[0.005, 0.006]
Gender	1	9.102	929.117	<0.001	0.004	[0.004, 0.005]
Residuals	154625	1514.719	-	-	-	-

Note. Score used as criterion. df = degrees of freedom; SS = sums of squares; η^2 = eta-squared; no. foreign lang. = number of foreign languages; 3278 observations deleted due to missingness. Values in square brackets indicate the bounds of the 95% confidence interval for eta-squared.

The effect of age on score reflects the fact that vocabulary size increases with age. This is illustrated in Figure 4, showing that the knowledge of Spanish vocabulary is about 55% (about 25,000 words in our test) between the ages of 25 and 30, and it increases up to 75% (around 34,000 words) by 75 to 80 years of age. This idea is consistent with previous studies in English (Brysbaert et al., 2016a). However, contrary to vocabulary declining in late adulthood, as previous studies suggest (McCabe et al., 2010), our results show that until 80 years of age, vocabulary keeps increasing, at least for the people who took part in our study.

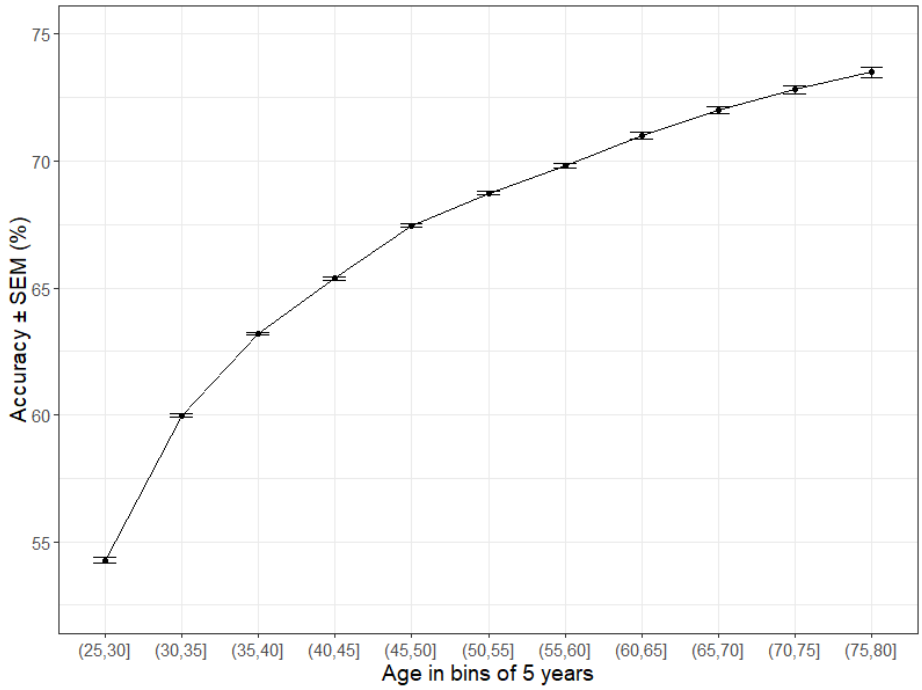

Figure 4. Score increases as a function of age. Age is plotted in bins of five years. Score is plotted in percentage. SEM = standard error of the mean.

Although we expected vocabulary size to be similar across different Spanish-speaking locations, differing only in words used, results show that on average, native speakers from Spain (M = 69.2, SD = 10.0) have a larger vocabulary size than native speakers from Latin-America (M = 61.5, SD = 11.7). The difference was of about 8% or around 3,500 words in our database. A likely factor in this difference is the fact that our word list did not contain typical Latin-American words. This fact was also evidenced in Aguasvivas et al. (2018; Figure 2), who observed there is a gap between Latin-American and Spanish speakers in the knowledge of about 30% of the words in this test.

Following previous findings, education level plays an important role in vocabulary size. Figure 5 shows the effect of education level on scores. For a student of secondary school, the mean score is 59% (SD = 12.2), which is more than half of the vocabulary in this test. Moreover, the score seems to increase linearly with the education level. For PhD students, the mean score is 71% (SD = 9.9). This implies a progressive increase of up to 12% or about 5500 words.

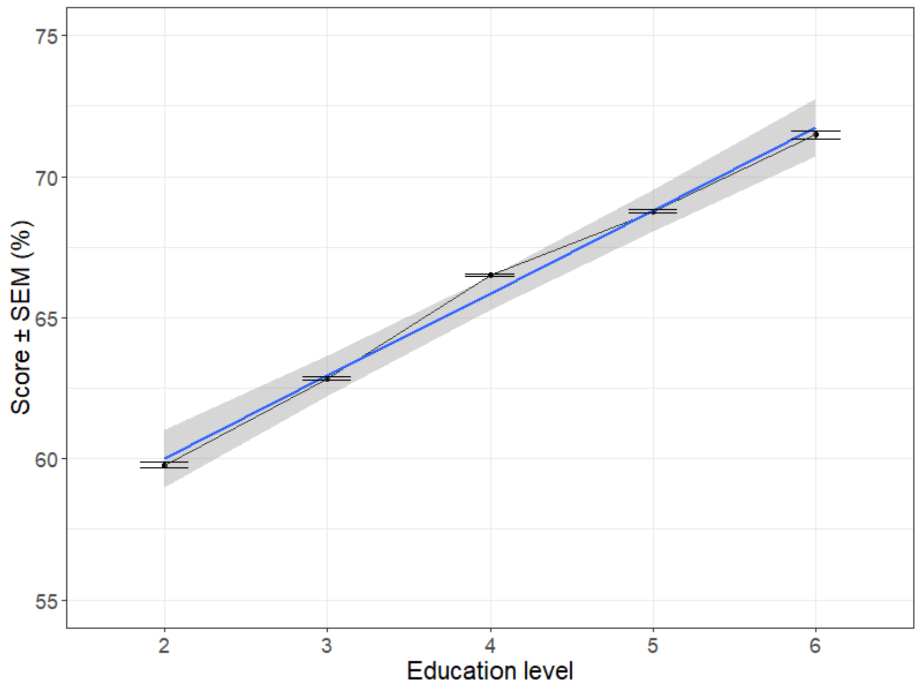

Figure 5. Score increases as a function of education level. SEM = standard error of the mean. Regression line is plotted in blue, with shading indicating standard error.

Contrary to the old studies suggesting a detrimental effect of foreign language knowledge on native language vocabulary size, our results seem to corroborate the idea of vocabulary size increasing with the knowledge of foreign languages (Keuleers et al., 2015). Figure 6 shows the effect of number of foreign languages on vocabulary size. The average difference between someone who knows 6 to 8 foreign languages and someone who knows 1 to 2 foreign languages is around 7%, which corresponds to a difference of around 3,000 words. Nonetheless, it is worth mentioning as a cautionary note that we did not take into account participants' proficiency in the languages as part of this survey.

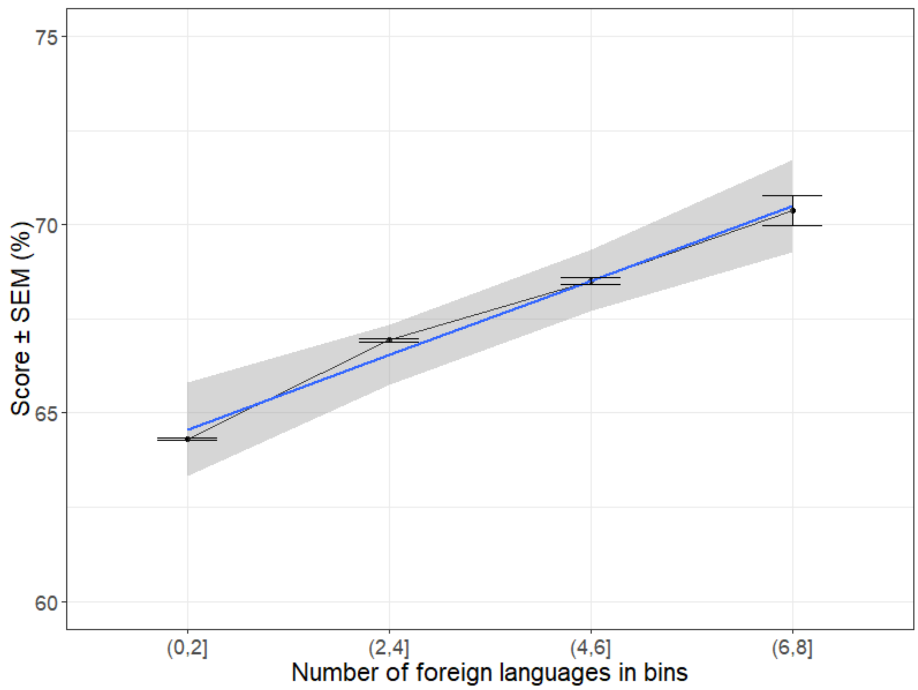

Figure 6. Effect of number of foreign languages on vocabulary size. Due to some levels showing very few observations, we opted to present the number of foreign languages known in bins of 2. SEM = standard error of the mean. Regression line is plotted in blue, with shading indicating standard error.

Finally, there seem to be small differences in vocabulary size according to the gender of the participants. These differences suggest that male participants score on average, about 2% higher than female participants. Although the difference was present throughout all ages, an informal exploration revealed that it was slightly larger for respondents older than 35. Nevertheless, it is important to note that these differences only represent a very small effect size barely surviving our criterion of 0.5% of variance explained, and considering the potential misconceptions that could arise from a lengthy discussion of this difference, we decided to withhold hypothetical interpretations in this regard.

Intrinsic effects. To test how intrinsic factors affected vocabulary knowledge in the LDT task, we performed a regression analysis using the average accuracy per word as the outcome variable, and frequency, orthographic neighbourhood size (old20), and word length as predictors. To obtain the average accuracy per word, we first excluded non-words from our database. Then we removed involuntary responses with RTs of less than 20ms (less than 0.01% of the data), and we trimmed the data removing RTs with response times above and below 3.0 box lengths to remove extremely slow or fast responses (3.55% of the data). Finally, we averaged the accuracy per word and discarded the words with less than 30 observations (0.49% of the words). In doing so, we retained information for 44,843 words, for which we ran a regression analysis with the predictors mentioned above.

As done in the analysis of the vocabulary size, we applied the criterion of 0.5% variance explained to successively eliminate two- and three-way interactions. Table 2 shows the estimates for the final model, which explained almost 50% of the variance ($R^2 = 0.49$, F = 8432.185, $p < 0.001$, 95% CI [0.48, 0.49]). In this model, frequency ($\beta = 1.06$, $p < 0.001$, 95% CI [1.03, 1.09]), length ($\beta = 1.22$, $p < 0.001$, 95% CI [1.19, 1.25]), and orthographic neighbourhood measured by old20 ($\beta = -0.80$, $p < 0.001$, 95% CI [-0.83, -0.78]) significantly predicted average accuracy. Furthermore, frequency showed a significant interaction with both length ($\beta = -1.28$, $p < 0.001$, 95% CI [-1.33, -1.23]), and old20 ($\beta = 0.82$, $p < 0.001$, 95% CI [0.77, 0.86]). Overall, the longer and more frequent a word is, the easier it is to recognize it. However, the fewer neighbours it has, the harder it is to recognize.

Table 2. *Regression results using average accuracy as the criterion*

Predictor	*b*	*b* 95% CI [LOW, HIGH]	*beta*	*beta* 95% CI [LOW, HIGH]	sr^2	sr^2 95% CI [LOW, HIGH]	*r*	Fit
(Intercept)	-0.23**	[-0.24, -0.21]						
Zipf	0.26**	[0.26, 0.27]	1.06	[1.03, 1.09]	0.07	[0.07, 0.07]	0.59**	
Length	0.16**	[0.16, 0.17]	1.22	[1.19, 1.25]	0.07	[0.07, 0.08]	0.15**	
Old20	-0.26**	[-0.27, -0.25]	-0.80	[-0.83, -0.78]	0.04	[0.03, 0.04]	-0.01*	
Zipf * Length	-0.04**	[-0.04, -0.04]	-1.28	[-1.33, -1.23]	0.03	[0.03, 0.03]		
Zipf * Old20	0.07**	[0.06, 0.07]	0.82	[0.77, 0.86]	0.02	[0.01, 0.02]		
								R^2 = 0.485**
								95% CI[.48,.49]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. Zipf indicates zipf transformed frequency. Old20 indicates orthographic neighborhood. * indicates $p < .05$. ** indicates $p < .01$.

Figure 7 shows the interaction between word length and frequency. For high-frequency words, length seems to become almost irrelevant in correctly recognising the word. On the other hand, word length seems to aid word recognition for lower frequency words. This interaction has been previously reported in multiple studies using different paradigms (LDT, naming, eye-tracking), suggesting an interplay between frequency and length in word processing (for a review, see Barton, Hanif, Eklinder Björnström, & Hills, 2014). Figure 8 shows the interaction between orthographic Levenshtein distance and frequency on word accuracy. Again, for high-frequency words, neighbourhood size does not seem to play a major role, but for low-frequency words, the more distant the word is from its neighbours (i.e., smaller orthographic neighbourhood), the higher the accuracy.

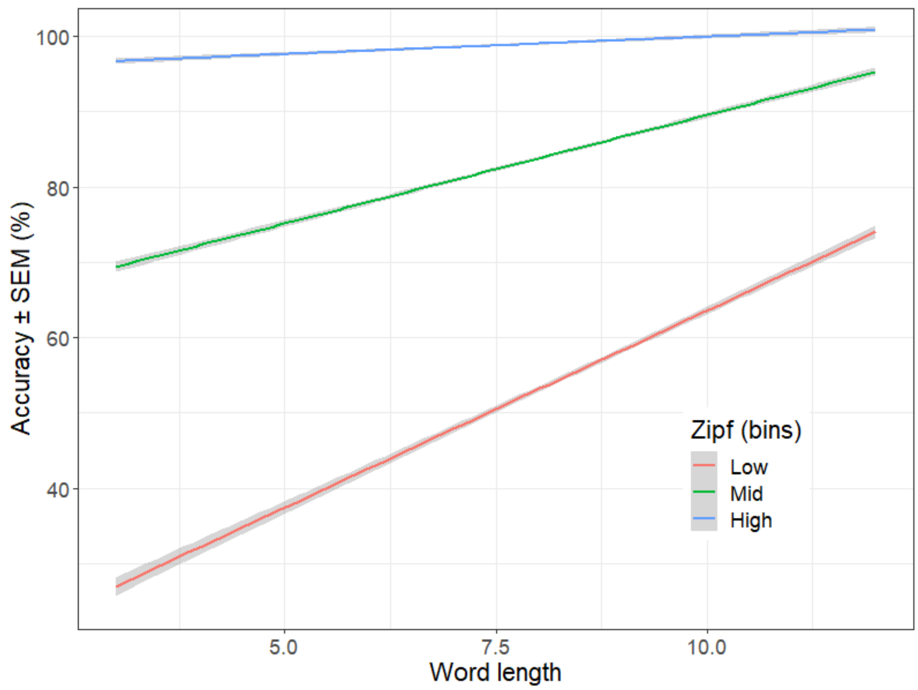

Figure 7. Interaction of word length and frequency on accuracy. Regression line lines are plotted in different colours according to the bin of frequency, shading indicates standard error. SEM = standard error of the mean

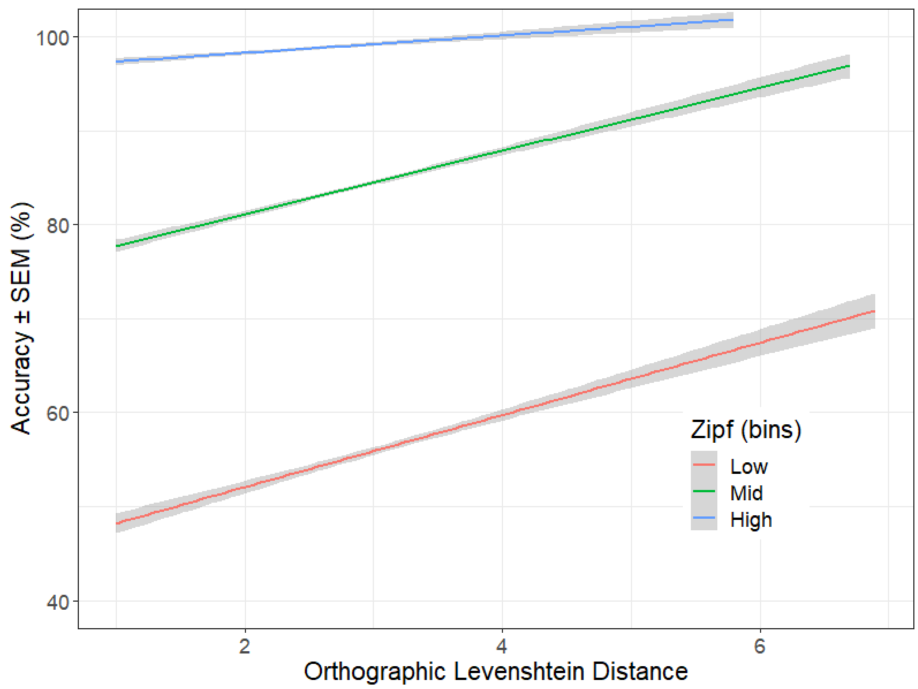

Figure 8. Interaction of orthographic Levenshtein distance and frequency on accuracy. Regression line lines are plotted in different colours according to the frequency bin, shading indicates standard error. SEM = standard error of the mean.

Vocabulary size in bilingual and monolingual communities within Spain

Participants who voluntarily provided their geolocation information and lived in one of designated regions in Spain (N = 1,679) were split into monolinguals and bilinguals depending on whether they fulfilled three conditions: (a) their country of origin was Spain, (b) the region were they were located was either a mainly monolingual community (Andalusia, Castile and Leon, Castile-La Mancha, Madrid, and Murcia) or a bilingual community (Basque Country, Catalonia, and Galicia), and (c) they reported knowing Spanish as their only language in the monolingual group, and knowing only the two co-official languages of the bilingual communities in the bilingual group (e.g., Basque and Spanish in Basque Country). The final monolingual group consisted of 794 participants, and the bilingual group included 885 participants.

The scores for both groups were subjected to a Bayesian t-test using the BEST package in R (Kruschke, 2013). We opted for a Bayesian framework because it provided a robust test of the differences between the groups, while also being able to test for the null hypothesis of no differences. We used the defaults of the BEST package, which assumes a *t* distribution as the descriptive model of the data and uses a non-informative prior that is updated with each observation to compute the posterior distributions for the means and standard deviations of both groups, as well as a parameter for normality (5 parameters in total) that are sampled using a Markov Chain Monte Carlo (MCMC) process (Kruschke, 2013). Figure 9 shows the results of the analysis, indicating that vocabulary size in monolingual communities (M = 69.6, SD = 10.2) did not differ significantly from that in bilingual communities (M = 69.5, SD = 10.1). The Bayes factor for this analysis indicated strong support for the null hypothesis of no differences between the groups (BF_{10} = 0.056). Additionally, the frequentist counterpart showed a similar result (t = 0.220, p = 0.826).

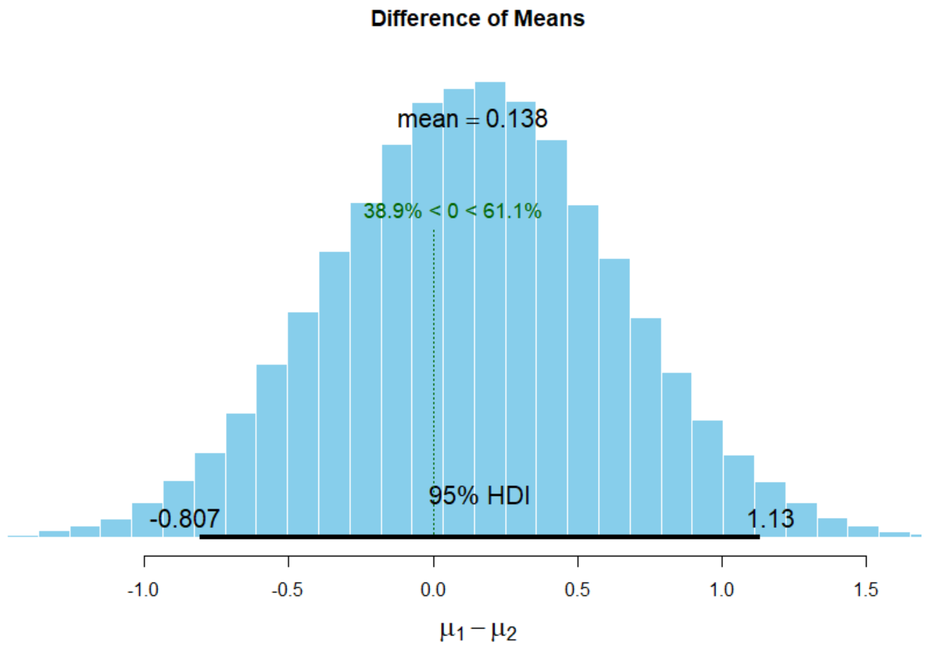

Figure 9. Difference of posterior means for monolinguals (μ_1) and bilinguals (μ_2).

Discussion

The present study aimed to examine Spanish vocabulary knowledge in a heterogeneous sample of native speakers collected through a massive online LDT. We discuss the results by focusing on the individual factors of the readers that directly affect visual word processing, after briefly summarising the impact of the words' properties in lexical access and vocabulary knowledge.

As expected, the frequency with which individuals are exposed to specific words influences how accurately they recognise them. Furthermore, we found an interaction between frequency and length and frequency and orthographic neighbourhood size on word accuracy. Overall, while high-frequency words are correctly recognised irrespective of their length, for low-frequency words, the longer they are, the more accurate participants are at recognising them.

The case is similar for the interaction between frequency and orthographic neighbourhood. For high-frequency words, the density of the word's neighbourhood does not seem to affect its recognition, but for low-frequency words, the less dense the neighbourhood, the more accurate participants are at recognising it. A possible reason is that participants feel uncertain about the spelling of low-frequency words with many neighbours and do not want to make a mistake by pressing yes to a misspelled word. Overall, the results corroborate previous conceptions of the mental lexicon that state that the ease of retrieval is mediated by the frequency with which individuals encounter words, and also by the length and orthographic neighbours of the word (for a review, see Barton et al., 2014). These results fit well with earlier studies from small and large scale studies in different languages (Balota et al., 2004; Brysbaert et al., 2019; González-Nosti et al., 2014).

How do individual differences determine vocabulary size?

Age. Age effects on vocabulary measures have traditionally reported a decrease in performance for middle- and older-aged individuals (McCabe et al., 2010). Our approach allowed us to test vocabulary across a wide range of ages and words, and the results, in conjunction with Keuleers et al. (2015), suggest that vocabulary knowledge keeps increasing with age in a seemingly logarithmic fashion. This logarithmic trend has also been corroborated in previous simulation studies (Ramscar, Hendrix, Shaoul, et al., 2014). The simple explanation is that, with time, individuals have more probability of encountering and learning novel words. While it is true that some of the previous studies have reported a decline with age in vocabulary knowledge, it is worth noting that they often have used productive vocabulary measures (e.g., Boston Naming Test; see MacKay, Connor, & Storandt, 2005; Simos, Kasselimis, & Mouzaki, 2011).

Why do we see these discrepancies? A first explanation might be that the mechanisms required for word recognition do not seem to be affected by age as those required for word production. This would be an interesting topic for further exploration. Nevertheless, an alternative is that most psychometric tests assume that vocabulary is age-invariant, and thus try to extrapolate vocabulary size from a limited set of words in the language, leading to an overall underrepresentation of the effect of age on vocabulary size (Ramscar, Hendrix, Shaoul, et al., 2014). Thus, by using the megastudy approach, we avoid most of the limitations by using a large set of words and assessing vocabulary size across a heterogeneous population.

Geographic location. Although we expected that different regions speaking the same language might exhibit lexical variations without reflecting differences in overall vocabulary size (Eisenstein et al., 2010), our results showed that native Spanish-speakers from Spain have a larger vocabulary size than native Spanish-speakers from Latin-America. While pinpointing the exact countries with smaller vocabulary sizes is beyond the scope of this study, we can attribute these differences to two reasons. First, despite the groups being similar in size, natives from Spain reported significantly higher education level, number of foreign languages, and age, which are all variables that also contributed to vocabulary size. Nevertheless, we did not find any significant interaction with these factors. Second, the words selected for the current test were obtained from written materials from Spain, which included less typical words from Latin-America, thus disfavouring participants from this region in contrast to those from peninsular

Spain. This fact has already been highlighted previously, detailing some of the examples in which there are differences between the variants of Spanish (Aguasvivas et al., 2018).

Education. The robustness of the effect of education level on lexical or semantic access is perhaps one of the reasons why most studies try to control for this variable (Simos et al., 2011). Our results confirm that vocabulary size increases with education. This is to be expected given that higher education level also allows the opportunity to acquire lower frequency words (Tainturier et al., 1992). These results exemplify two important points. The first is the contextual opportunity that higher education offers individuals (Jones, Dye, & Johns, 2017). The likelihood of encountering new words depends highly on the context in which they appear. For instance, corpora analyses show that only the most frequent words appear across all texts, but more than 99% of the vocabulary is conditional on contextual factors (Jones et al., 2017). In this case, while the vocabulary size of an individual with a degree in physics and another one with a degree in psychology might contain a lot of overlapping words, a big part of the words they know will be highly dependent on the degree of their choosing, even though the overall vocabulary size appears to be similar (see also Ramscar, Hendrix, Love, et al., 2014). However, both of these individuals will have an increased vocabulary size when compared to individuals with a high-school education level. A larger variety of contexts in which one lives results in a larger number of words known.

The second point relates to conscientiousness. Individuals with higher education level might be more aware and careful of their responses, trying to reduce guessing in these types of tasks, which in turn can lead to fewer false alarms, and overall increased performance (Biderman, Nguyen, & Sebren, 2008), especially in an untimed LDT. A brief examination of the data indicates a small but negative correlation between education level and the rate of false alarms in our test, but also a positive correlation with a raw score for words, supporting both of the previously posed arguments.

Multilingualism. The common conception of the effect of multilingualism on vocabulary size is that multilingual individuals are less exposed to words in any of the languages they know (Gollan et al., 2008). If so, the natural prediction is that multilinguals will show decreased vocabulary size as compared to a native speaker of the language (Gollan & Acenas, 2004; Gollan et al., 2008). Previous research with monolingual and bilingual adults and children shows that there is a consistent difference in both productive and receptive vocabulary that does not vary with the language pair of the bilinguals (Bialystok & Luk, 2012; Bialystok, Luk, Peets, & Yang, 2010; De Houwer, Bornstein, & Putnick, 2012). Despite this, our results indicate that the knowledge of multiple languages increases Spanish vocabulary size rather than decreasing it. Keuleers et al. (2015) offer a possible explanation for this, suggesting that, because some languages share a big percentage of their vocabulary, the lack of exposure to L1 vocabulary might be compensated indirectly by learning novel vocabulary in a different language. In the case of Spanish and due to its close relation to other romance languages like French, Portuguese, and Italian, indirect vocabulary acquisition might explain increased vocabulary knowledge. Here again, a likely mechanism is that knowledge of various languages increases the variety of contexts in which people learn specific vocabularies.

When contrasting different regions within Spain based on their multilingual status, our results indicate moderate evidence towards the null hypothesis, suggesting that there are no reliable differences in vocabulary size between these regions, regardless of the number of languages used at the official level. Bilingual educational policies have been in place for more than 20 years in autonomous communities like Catalonia and the Basque Country, and yet a common criticism has been that students in these communities would not perform on par with students from monolingual communities when their level of Spanish is assessed (Huguet, 2007). While we acknowledge that our assessment of vocabulary size does not encompass other forms of linguistic competence, such as production or comprehension, we did not observe differences between monolingual and bilingual communities in vocabulary size.

Due to the similarity of the methods, our data and results are directly comparable with those of Keuleers et al., (2015) in several respects. First, despite being different languages and samples, our findings support the idea of a vocabulary size increase (not plateauing) with age. Second, we corroborated the effects of education and number of known foreign languages. Additionally, the present study also delves into other factors affecting word knowledge by replicating some of the most prominent effects in the lexical decision literature. In this sense, we examined not only extrinsic, but also intrinsic factors affecting vocabulary size and knowledge, providing additional support to well established psycholinguistic findings. Finally, our results also provide compelling data in favour of bilingual education, showing the lack of differences in vocabulary knowledge between monolingual and bilingual speakers within Spain.

Conclusion

The current study offers valuable data regarding individual word processing in Spanish on the largest data collection conducted so far in this language. We tested a large number of participants of varying origins and with different sociodemographic backgrounds, and a considerable amount of words that nicely capture the intricacies of the Spanish language. Thanks to the use of crowdsourcing techniques and following the way started by Keuleers et al. (2015), we were able to effectively replicate basic effects associated with the intrinsic characteristics of the words in the language, such as the word length and frequency effects, and the classic length by frequency interaction that has been repeatedly documented in the literature. But over and above validating these effects in a large-scale data collection, this study offered the possibility to explore the potential impact of some of the characteristics of the respondents in vocabulary knowledge. By following such an approach, we found a reliable and seemingly independent contribution of age, number of languages known, and education level, among others, to lexical knowledge as measured by a lexical decision task. Results demonstrated that vocabulary knowledge increases with age, yielding the conclusion that increased age is by no means detrimental to word recognition. Hence, in light of these results, it remains to be seen whether the differences observed in production tasks in the elderly could be related to issues that do not necessarily tap into lexical knowledge but on recollection or articulation concerns. More importantly, the data demonstrate that there's a linear increase in vocabulary knowledge as a function of both the number of languages known and the education level. Additionally, our approach showed that vocabulary size did not differ in monolingual and bilingual communities within Spain, an aspect of considerable importance for linguistic policies within these regions. Other than highlighting the value of crowdsourcing based megastudies to uncover critical effects that could be masked otherwise, these results highlight the benefits derived of multilingualism and education for lexical richness, and consequently, for language wealth.

Acknowledgements: This research is supported by the Basque Government through the BERC 2018-2021 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation SEV-2015-0490. This study was also partially supported by grants PGC2018-097145-B-I00, RED2018-102615-T and RTI2018-093547-B-I00 from the Spanish State Research Agency. Work by JA was supported by “la Caixa” Foundation and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 713673, and fellowship code LCF/BQ/IN17/116200154004.

Open Practices Statement

The data for this experiment is available at https://figshare.com/projects/SPALEX/29722. This experiment was not preregistered.

References

Acha, J., & Perea, M. (2008). The effects of length and transposed-letter similarity in lexical decision: Evidence with beginning, intermediate, and adult readers. *British Journal of Psychology*. https://doi.org/10.1348/000712607X224478

Aguasvivas, J. A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2018). SPALEX: A Spanish Lexical decision database from a Massive Online Data Collection. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.02156

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieal: Resolving neighbourhood conflicts. *Psychonomic Bulletin and Review2*, *4*(4), 439–461.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual Word Recognition: The Journey From Features to Meaning (A Travel Update). In *Handbook of Psycholinguistics*. https://doi.org/10.1016/B978-012369374-7/50010-9

Barton, J. J. S., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology*. https://doi.org/10.1080/02643294.2014.895314

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive Control and Lexical Access in Younger and Older Bilinguals. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/0278-7393.34.4.859

Bialystok, E., & Luk, G. (2012). Receptive vocabulary differences in monolingual and bilingual adults. *Bilingualism*. https://doi.org/10.1017/S136672891100040X

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*. https://doi.org/10.1017/s1366728909990423

Biderman, M. D., Nguyen, N. T., & Sebren, J. (2008). Time-on-task mediates the conscientiousness-performance relationship. *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2007.10.022

Bowles, R. P., & Salthouse, T. A. (2008). Vocabulary Test Format and Differential Relations to Age. *Psychology and Aging*. https://doi.org/10.1037/0882-7974.23.2.366

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721417727521

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016a). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.01116

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016b). The impact of word prevalence on lexical decision times: Evidence from the Dutch lexicon project 2. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/xhp0000159

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

Casaponsa, A., Antón, E., Pérez, A., & Duñabeitia, J. A. (2015). Foreign language comprehension achievement: Insights from the cognate facilitation effect. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.00588

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In *Attention and Performance VI*. https://doi.org/10.1006/brln.2001.2475

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles | SUBTLEX-ESP: Frecuencias de las palabras españolas basadas en los subtítulos de las películas. *Psicologica*.

Davies, R. A. I., Birchenough, J. M. H., Arnell, R., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000366

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*. https://doi.org/10.3758/BF03192738

De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2012). A bilingual-monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. *Applied Psycholinguistics*. https://doi.org/10.1017/S0142716412000744

Diependaele, K., Brysbaert, M., & Neri, P. (2012). How noisy is lexical decision? *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2012.00348

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2010.00218

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*. https://doi.org/10.3758/s13428-013-0326-1

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., … Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0024974

Duñabeitia, J. A., & Vidal-Abarca, E. (2008). Children like dense neighborhoods: orthographic neighborhood density effects in novel readers. *Spanish Journal of Psychology*. https://doi.org/10.1017/S113874160000408X

Eisenstein, J., O’Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical Variation. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, MIT, Massachusetts, USA, 9-11 October 2010*.

Ethnologue. (2016). Summary by language size. Retrieved from https://www.ethnologue.com/statistics/size

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French lexicon project: Lexical decision data for 38,840 French words and 38,840 pseudo words. *Behavior Research Methods*. https://doi.org/10.3758/BRM.42.2.488

Gierut, J. A., & Dale, R. A. (2007). Comparability of Lexical Corpora: Word frequency in phonological generalization. *Clinical Linguistics and Phonetics*. https://doi.org/10.1080/02699200701299891

Gollan, T. H., & Acenas, L. A. R. (2004). What Is a TOT? Cognate and Translation Effects on Tip-of-the-Tongue States in Spanish-English and Tagalog-English Bilinguals. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/0278-7393.30.1.246

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2007.07.001

González-Nosti, M., Barbón, A., Rodríguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765 words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-013-0383-5

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

Huguet, Á. (2007). Minority languages and curriculum: The case of Spain. *Language, Culture and Curriculum*. https://doi.org/10.2167/lcc327.0

Huguet, Á., Lapresta, C., & Madariaga, J. M. (2008). A study on language attitudes towards regional and foreign languages by school children in aragon, Spain. *International Journal of Multilingualism*. https://doi.org/10.1080/14790710802152412

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an Organizing Principle of the Lexicon. *Progress in Brain Research*. https://doi.org/10.1016/bs.plm.2017.03.008

Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., & Whalen, D. H. (2012). What lexical decision and naming tell us about reading. *Reading and Writing*. https://doi.org/10.1007/s11145-011-9316-9

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1080/17470218.2015.1051065

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*. https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon*. https://doi.org/10.1075/ml.6.1.02keu

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-011-0118-4

Keuleers, E., & Marelli, M. (n.d.). *Resources for mental lexicon research: A delicate ecosystem*.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1080/17470218.2015.1022560

Kosmidis, M., Tsapkini, K., & Folia, V. (2006). Lexical processing in illiteracy: Effect of literacy or education? *Cortex*, *42*(7), 1021–1027. https://doi.org/10.1016/S0010-9452(08)70208-9

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/a0029146

Kulkarni, V., Perozzi, B., & Skiena, S. (2016). Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media. *Proceedings of the Tenth International AAAI Conference on Web and Social Media*.

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology. Human Perception and Performance*. https://doi.org/10.1037/a0030859

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. https://doi.org/10.1023/A:1022689900470

MacKay, A., Connor, L. T., & Storandt, M. (2005). Dementia does not explain correlation between age and scores on Boston Naming Test. *Archives of Clinical Neuropsychology*. https://doi.org/10.1016/j.acn.2004.03.006

McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology*. https://doi.org/10.1037/a0017619

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*. https://doi.org/10.1037/0033-295X.111.3.721

Myers, J. (2016). Meta-megastudies. *The Mental Lexicon*. https://doi.org/10.1075/ml.11.3.01mye

Pollatsek, A., Perea, M., & Binder, K. S. (1999). The effects of “neighborhood size” in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/0096-1523.25.4.1142

Portocarrero, J. S., Burright, R. G., & Donovick, P. J. (2007). Vocabulary and verbal fluency of bilingual and monolingual college students. *Archives of Clinical Neuropsychology*. https://doi.org/10.1016/j.acn.2007.01.015

Ramscar, M., Hendrix, P., Love, B., & Baayen, H. (2014). Learning is not decline. *The Mental Lexicon*. https://doi.org/10.1075/ml.8.3.08ram

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12078

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The Mismeasurement of Mind: Life-

Span Changes in Paired-Associate-Learning Scores Reflect the "Cost" of Learning, Not Cognitive Decline. *Psychological Science*. https://doi.org/10.1177/0956797617706393

Schröter, P., & Schroeder, S. (2017). The Developmental Lexicon Project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods*. https://doi.org/10.3758/s13428-016-0851-9

Simos, P. G., Kasselimis, D., & Mouzaki, A. (2011). Age, gender, and education effects on vocabulary measures in Greek. *Aphasiology*. https://doi.org/10.1080/02687038.2010.512118

Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., & Baltes, P. B. (2003). The fate of cognition in very old age: six-year longitudinal findings in the Berlin Aging Study (BASE). *Psychology and Aging*.

Singh-Manoux, A., Kivimaki, M., Glymour, M. M., Elbaz, A., Berr, C., Ebmeier, K. P., … Dugravot, A. (2012). Timing of onset of cognitive decline: Results from Whitehall II prospective cohort study. *BMJ (Online)*. https://doi.org/10.1136/bmj.d7622

Solomon, R. L., & Howes, D. H. (1951). Word frequency, personal values, and visual duration thresholds. *Psychological Review*. https://doi.org/10.1037/h0058228

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-015-0700-2

Tainturier, M. J., Tremblay, M., & Lecours, A. (1992). Educational level and the word frequency effect: A lexical decision investigation. *Brain and Language*. https://doi.org/10.1016/0093-

934X(92)90112-R

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1080/17470218.2013.850521

Verhaeghen, P. (2003). Aging and vocabulary scores: A meta-analysis. *Psychology and Aging*. https://doi.org/10.1037/0882-7974.18.2.332

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/PBR.15.5.971