

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

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

Jose Aguasvivas^{1,2*}

Manuel Carreiras^{1,2,3}

Marc Brysbaert⁴

Pawel Mandera⁴

Emmanuel Keuleers⁵

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, Tilburg University; Tilburg, 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

26

Abstract

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

43

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

45

69 appearance of that word within multiple sources is counted. Other properties, however, require
70 participants to complete questionnaires asking about different subjective dimensions that cannot
71 be automatically computed from corpora, and that may vary depending on participants'
72 characteristics (e.g., valence, arousal, age of acquisition; Gierut & Dale, 2007). In this sense,
73 Keuleers and Marelli (n.d.) distinguish between unelicited properties –those that can be obtained
74 from linguistic resources using computational methods– and elicited properties that can be
75 obtained directly from participants' elicited behaviour.

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

82 The source on which distributional measures for words are based can influence the
83 expected results of LDT. For instance, the performance of younger adults is better predicted by
84 frequencies obtained from internet sources (Balota, Cortese, Sergent-Marshall, Spieler, & Yap,
85 2004; Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010), while the frequencies of a
86 corpus based on movie subtitles in the US, but not in the UK, better predicts the performance of
87 US students (Brysbaert & New, 2009). There is not a unique corpus that can fully capture the
88 heterogeneity of a language's vocabulary across different individuals. Due to this, Keuleers and
89 Balota (2015) suggest using approaches where participants can assess word properties in
90 conjunction with corpus information. Under this novel *crowdsourcing* approach, online
91 platforms function as a vehicle for the assessment of properties from a vast number of raters.

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

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

113 Crowdsourced lexical decision megastudies have numerous advantages. First, they allow
114 for massive data collection at a reduced cost by distributing the experiment through an online

115 platform and providing alternative incentives to participants (e.g., sending scores via e-mail; see
116 Dufau et al., 2011). Second, the effects of continuous variables (like frequency) can be treated as
117 such without the need to categorise them (Keuleers et al., 2012). Third, the studies provide
118 normative information on performance from a vast number of participants on many words (and
119 nonwords). Fourth, virtual experiments can be run within the database to evaluate novel
120 hypotheses or better control stimuli selection, and computational models of word recognition can
121 be evaluated against the data (Stadthagen-Gonzalez, Imbault, Pérez Sánchez, & Brysbaert,
122 2017). Finally, the data from multiple megastudies can be combined to produce meta-
123 megastudies, drawing inferences about language processing beyond the scope of a specific
124 language (Myers, 2016).

125

126

Word accuracy as an indicator of vocabulary size

127

128

129

130

131

132

133

134

135

136

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).

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

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

153

154 **Extrinsic factors affecting LDT**

155

156 *Age.* With time, individuals can encounter and learn novel words in both their native and
157 other languages. Studies measuring the effect of age on vocabulary knowledge tend to conclude
158 that, independently of the format used (e.g., multiple choice, production, lexical decision),
159 vocabulary increases drastically throughout early adulthood, then flatten in middle-age, only to

160 then decline gradually or hold steady through late adulthood (Bowles & Salthouse, 2008;
161 McCabe, Roediger, McDaniel, Balota, & Hambrick, 2010; Singer, Verhaeghen, Ghisletta,
162 Lindenberger, & Baltes, 2003; Singh-Manoux et al., 2012). Recent LDT megastudies suggest
163 that vocabulary keeps increasing with age, and does not decline as previously thought (at least
164 not in the participants that take part in the test), suggesting that age is one of the most relevant
165 predictors of vocabulary size (see Brysbaert et al., 2016a). Furthermore, the effect of intrinsic
166 properties such as frequency and age of acquisition seems to be mediated by age, with a decrease
167 in the size of the effect as age increases (Davies, Birchenough, Arnell, Grimmond, & Houlson,
168 2017). Also, lexical decision response time appears to remain largely unaffected by age (Schröter
169 & Schroeder, 2017). While slowing response times in other tasks is often attributed to an ageing-
170 related decline in information processing capacities, it can, in fact, reflect increased information
171 processing demands (Ramscar, Hendrix, Love, & Baayen, 2014; Ramscar, Hendrix, Shaoul,
172 Milin, & Baayen, 2014; Ramscar, Sun, Hendrix, & Baayen, 2017).

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

183 words and discard non-words more accurately than those with lower education level (Kosmidis,
184 Tsapkini, & Folia, 2006).

185 **Geographic location.** It is known that language varies across social and regional
186 contexts, which is the subject of study of sociolinguistics and dialectology (Eisenstein,
187 O'Connor, Smith, & Xing, 2010). These variations also suggest that vocabulary, albeit similar in
188 size, might be composed of different words depending on the location of the speaker, as is the
189 case with Latin-American versus Castilian Spanish (Aguasvivas et al., 2018). By using
190 geotagged material, inferences can be drawn on lexical, syntactic, and semantic variations not
191 only across countries but also within regions of the same country (Kulkarni, Perozzi, & Skiena,
192 2016). This is particularly interesting for countries like Spain, in which linguistic policies
193 acknowledge the country's multilingual and multicultural character, allowing some communities
194 to increase the presence of languages other than Spanish in compulsory education (Huguet,
195 2007). Despite this, there is scarce tradition of research on the linguistic aptitudes of individuals
196 within these regions (Huguet, Lapresta, & Madariaga, 2008). For this study, we are interested in
197 knowing whether Spanish vocabulary size is similar within these regions as compared to regions
198 where both the educational and social context is limited to Spanish. Furthermore, we are
199 interested in comparing Spanish across multiple Spanish-speaking countries.

200 **Multilingualism.** Before megastudies were run, small-scale studies comparing bilinguals
201 and monolinguals on linguistic tasks suggested that bilinguals showed decreased lexical retrieval
202 capacity (Portocarrero, Burright, & Donovanick, 2007), less verbal fluency (Bialystok, Craik, &
203 Luk, 2008), and greater interference in lexical decisions (Gollan & Acenas, 2004). They all
204 pointed to disadvantages that arose due to (a) individuals dividing their word usage between the
205 languages they know, and (b) multilinguals being exposed less to a specific language than a

206 monolingual person (Gollan, Montoya, Cera, & Sandoval, 2008). However, contrary to these
207 early findings (and researcher intuitions), Keuleers et al. (2015) found not only that L1
208 vocabulary size was larger in bilinguals, but that L1 vocabulary size increased with the number
209 of languages the participants reported to know. This is a critical finding that deserves close
210 attention, and the use of a parallel megastudy approach in a different language will allow us to
211 test its replicability. Overall, Keuleers et al.'s conclusion regarding multilingualism and
212 vocabulary size is that vocabulary in a language might be aided by the knowledge of other
213 languages, mainly because the knowledge of extra languages gives people more diverse contexts
214 in which to learn words. Given that many of these words are cognates in several languages (have
215 the same form and meaning), knowing words in a second language is likely to increase
216 knowledge of the same words in the native language. For instance, knowing the Spanish word
217 *siesta* increases the English vocabulary as well. This line of argumentation fits well with recent
218 evidence demonstrating the role of cognate words in the process of language learning (e.g.,
219 Casaponsa, Antón, Pérez, & Duñabeitia, 2015).

220

221 **Intrinsic factors affecting LDT**

222

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

227 **Word frequency.** The word frequency effect is one of the most robust and well-

228 documented effects of the word recognition literature (Brysbaert, Mandera, & Keuleers, 2018). It

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

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

248 For this study, we tackle the question of how word frequency relates to vocabulary
249 knowledge. The frequency measure used in this study was extracted for each word from the
250 EsPal database using the Zipf scale (Duchon et al., 2013), which is roughly equivalent to the
251 base 10 logarithm of the frequency per billion words and ranges from 1 to 7 (for a detailed

252 description of the scale, see van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The higher the
253 value in Zipf scale, the more frequent a word is seen in the corpus.

254 **Orthographic neighbourhood size.** The time required to recognise a printed word also
255 seems to depend on the degree of orthographic similarity it has to other words in the language
256 (Diependaele et al., 2012). In the traditional definition (Coltheart, Davelaar, Jonasson, & Besner,
257 1977), a word's orthographic neighbourhood (N) is the number of words that have the same
258 length as that word, but that differ in exactly one letter (e.g., *cake – lake*). A higher value for the
259 orthographic neighbourhood implies that a word has more similarity to existing words. A more
260 recent definition (Yarkoni, Balota, & Yap, 2008), operationalises orthographic neighbourhood
261 density as the average Levenshtein distance (Levenshtein, 1966) between a word and its 20
262 nearest orthographic neighbours (OLD20). Higher values in this measure indicate a sparser
263 neighbourhood, as the average distance between the target words and its neighbours is larger.

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

272 **Length.** The number of characters in a word can greatly influence the time required to
273 recognise it, as the individual requires more grapheme-phoneme conversions during reading.
274 Most studies have traditionally controlled for this variable instead of including it, which has led

275 to an overshadowing of its possible effect on word recognition time and accuracy (González-
276 Nosti, Barbón, Rodríguez-Ferreiro, & Cuetos, 2014). In this aspect, Acha and Perea (2008)
277 compared beginner (children), intermediate, and adult readers in a Spanish LDT showing that,
278 while the length effect for words was robust in children and disappeared in adults, the effect of
279 the length of non-words followed the opposite pattern. They suggested that in a fully developed
280 lexical system, access to known word representation occurs automatically while accessing
281 unknown words or non-words requires letter-by-letter decoding (Acha & Perea, 2008).

282

283

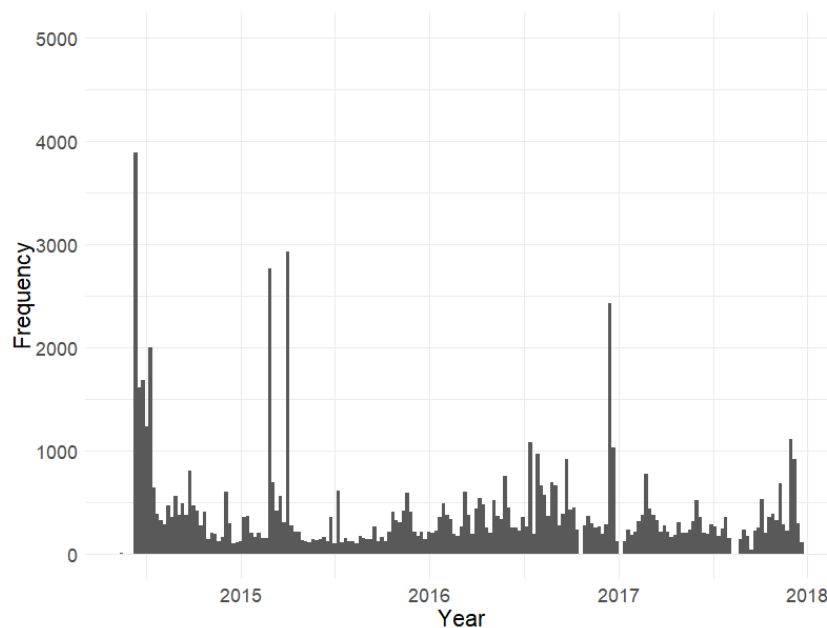
Method

284

Participants

286

287 We collected data from May 12th, 2014 to December 19th, 2017 (see Figure 1). Up to
288 that point, 209,351 participants had finished 282,576 tests by completing one (80.0%), two
289 (14.1%), three (3.3%), or more sessions (2.6%). Most of the data (68.9%) were acquired during
290 the first month of the experiment when a radio advertising campaign was run to attract the
291 public's attention. Participants also had the option of publishing their results via social networks,
292 which attracted new participants in a snow-ball sampling fashion. Additionally, before the
293 experiment, participants were able to voluntarily provide information about their sex, age,
294 country of origin, education level, handedness, number of known foreign languages, best foreign
295 language, and geolocation information. The raw version of this data for native Spanish speakers
296 is presented in the SPALEX database made available in Aguasvivas et al. (2018) and it can be
297 retrieved from <https://figshare.com/projects/SPALEX/29722>.



298

299 **Figure 1.** Frequency of participation per year. Each line represents a week. Participation in the
 300 year 2014 represented 73.77% of the data, while 2015 represented 9.20%, 2016 10.30%, and
 301 2017 6.74% of the data. Gaps in the distribution of responses correspond with maintenance
 302 periods of the online platform.

303

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

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

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

335 Using only the geolocation information of participants within Spain, we identified those
336 that were located in official bilingual communities (Basque Country, Catalonia, and Galicia). A

337 group of participants living in official monolingual communities that matched the number of
338 participants in the bilingual communities (Andalusia, Castile and Leon, Castile-La Mancha,
339 Madrid, and Murcia) was also selected for comparison purposes. Furthermore, we limited the
340 number of foreign languages reported by these participants to match monolingual and bilingual
341 profiles. A total of 1,679 participants (885 bilinguals) were therefore extracted from the database
342 and stored for a separate analysis.

343

344 **Materials**

345

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

358

359 **Procedure**

360

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



372

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

376

377 Before the beginning of the experimental session, each participant had the option to fill in
 378 the demographic questionnaire and provide their geolocation information voluntarily. Answering
 379 these questions was not required to proceed with the experiment, but participants not answering
 380 them were not included in the analyses. After the questionnaire screen, participants were
 381 instructed to place their fingers in the instructed position (buttons or keys) and press a button to
 382 begin the experiment. The stimuli were always presented in a vertically and horizontally centred

383 position on the screen, and a blue progress bar on the top of the screen informed participants of
384 their advancement through the experiment (see also Figure 2). Responses were automatically
385 coded into correct and incorrect responses, and response time (RT) was recorded in milliseconds
386 for each response. It is important to note that in Aguasvivas et al. (2018) we tested whether the
387 70/30 word to nonword ratio introduced bias in the accuracy scores by using the LD1NN
388 algorithm (Keuleers & Brysbaert, 2011). The results indicated that if participants were to base
389 their decisions only on the statistical characteristics of presented words and nonwords, they
390 would be 2.6 times more likely to identify a stimulus as a word than as a nonword. Values from
391 other studies range from 0.34 to 4.1, depending on how nonwords are created. We also tested the
392 reliability of RT scores by using the split-half method, obtaining Spearman-Brown corrected
393 reliability of 0.92 for words and 0.91 for nonwords.

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

401

402

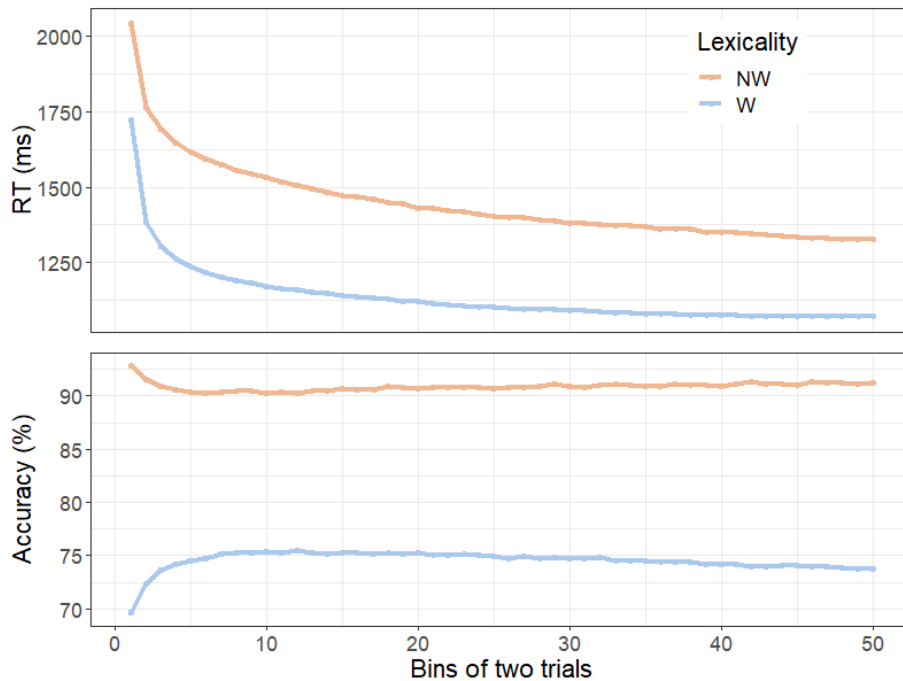
Results

403

404 We calculated a score for each participant by subtracting the percentage of false alarms
405 (incorrectly accepted nonwords) from the percentage of hits (correctly accepted words). This

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

414



415

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

420

421 Variables affecting vocabulary knowledge

422

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

429 Due to the amount of observations and terms in the regression, we opted to run a first
 430 model including all factors and their two- and three-way interactions. We then selected only

431 those terms that accounted for more than 0.5% of the variance. After the first iteration, only the
432 main effects remained. Table 1 shows the results of the final model for the score of the
433 participants, which accounted for 28% of the variance in scores ($R^2 = 0.278$, $F = 4851.914$, $p <$
434 0.001 , 95% CI [0.27, 0.28]. While most of the factors were significant in the initial model, the
435 surviving terms after applying the criteria were age ($F = 34751.097$, $p < 0.001$, $\eta^2 = 0.164$, 95%
436 CI [0.161, 0.168]), geographic location ($F = 17142.431$, $p < 0.001$, $\eta^2 = 0.081$, 95% CI [0.079,
437 0.083]), education level ($F = 828.432$, $p < 0.001$, $\eta^2 = 0.016$, 95% CI [0.015, 0.017]), reported
438 number of foreign languages ($F = 1103.272$, $p < 0.001$, $\eta^2 = 0.005$, 95% CI [0.005, 0.006]), and
439 gender ($F = 929.117$, $p < 0.001$, $\eta^2 = 0.004$, 95% CI [0.004, 0.005]).

440

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

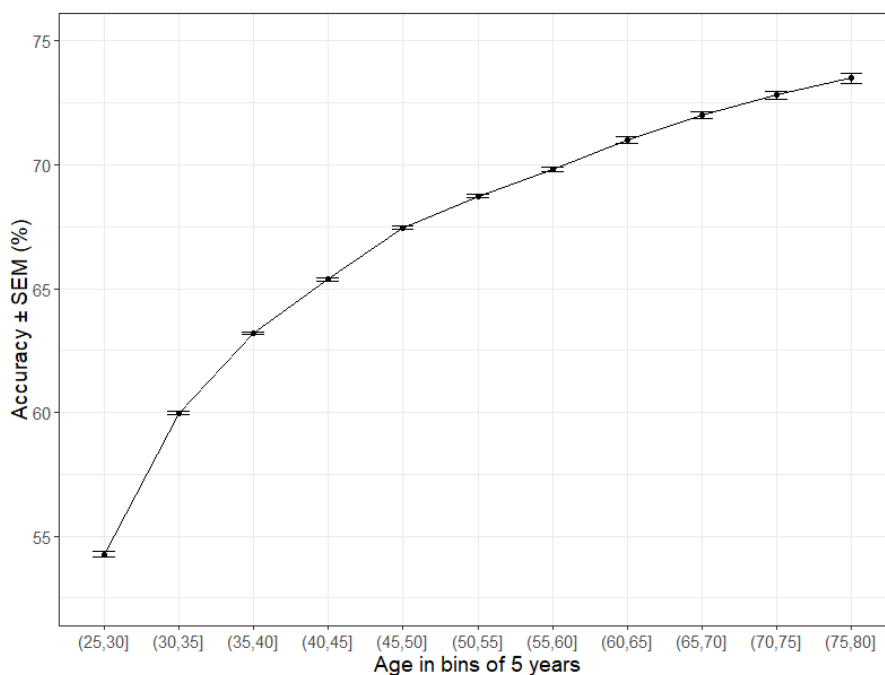
Term	df	SS	F	<i>p</i>	η^2	95% CI [LOW, HIGH]
Log(Age)	1	340.424	34751.097	<0.001	0.164	[0.161, 0.168]
Location	1	167.929	17142.431	<0.001	0.081	[0.079, 0.083]
Education	4	32.462	828.432	<0.001	0.016	[0.015, 0.017]
No. foreign lang.	1	10.808	1103.272	<0.001	0.005	[0.005, 0.006]
Gender	1	9.102	929.117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

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

445

446 The effect of age on score reflects the fact that vocabulary size increases with age. This is
 447 illustrated in Figure 4, showing that the knowledge of Spanish vocabulary is about 55% (about
 448 25,000 words in our test) between the ages of 25 and 30, and it increases up to 75% (around
 449 34,000 words) by 75 to 80 years of age. This idea is consistent with previous studies in English
 450 (Brybaert et al., 2016a). However, contrary to vocabulary declining in late adulthood, as
 451 previous studies suggest (McCabe et al., 2010), our results show that until 80 years of age,
 452 vocabulary keeps increasing, at least for the people who took part in our study.

453



454

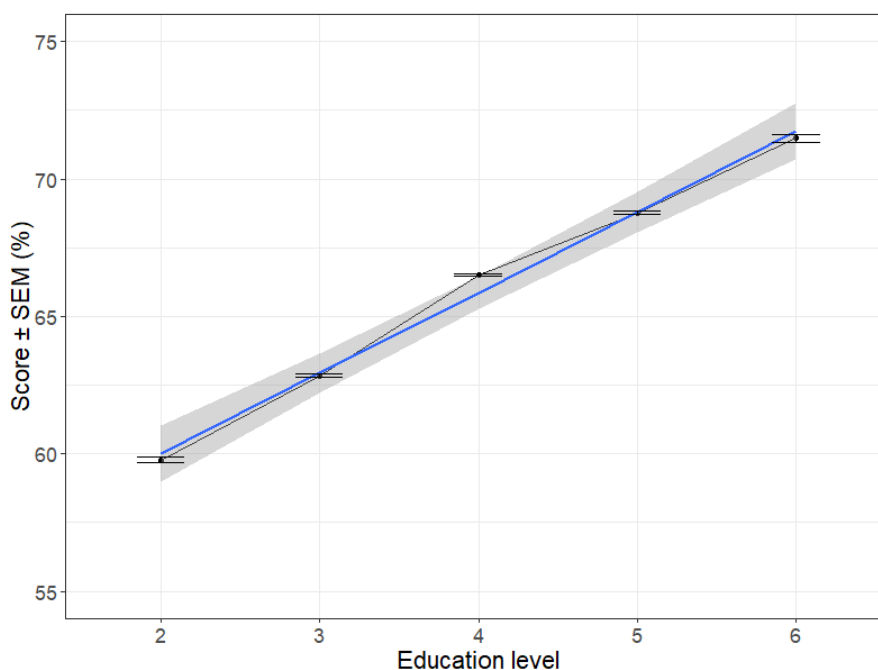
455 **Figure 4.** Score increases as a function of age. Age is plotted in bins of five years. Score is
 456 plotted in percentage. SEM = standard error of the mean.

457

458 Although we expected vocabulary size to be similar across different Spanish-speaking
 459 locations, differing only in words used, results show that on average, native speakers from Spain
 460 ($M = 69.2$, $SD = 10.0$) have a larger vocabulary size than native speakers from Latin-America
 461 ($M = 61.5$, $SD = 11.7$). The difference was of about 8% or around 3,500 words in our database.
 462 A likely factor in this difference is the fact that our word list did not contain typical Latin-
 463 American words. This fact was also evidenced in Aguasvivas et al. (2018; Figure 2), who
 464 observed there is a gap between Latin-American and Spanish speakers in the knowledge of about
 465 30% of the words in this test.

466 Following previous findings, education level plays an important role in vocabulary size.
 467 Figure 5 shows the effect of education level on scores. For a student of secondary school, the
 468 mean score is 59% ($SD = 12.2$), which is more than half of the vocabulary in this test. Moreover,

469 the score seems to increase linearly with the education level. For PhD students, the mean score is
 470 71% (SD = 9.9). This implies a progressive increase of up to 12% or about 5500 words.

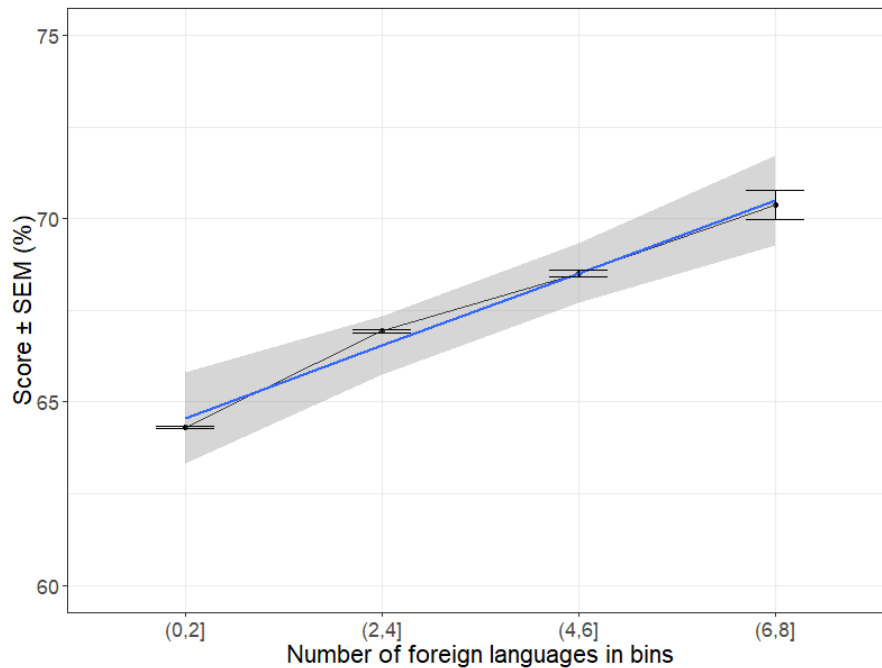


471

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

474

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



483

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

488

489 Finally, there seem to be small differences in vocabulary size according to the gender of
 490 the participants. These differences suggest that male participants score on average, about 2%
 491 higher than female participants. Although the difference was present throughout all ages, an
 492 informal exploration revealed that it was slightly larger for respondents older than 35.

493 Nevertheless, it is important to note that these differences only represent a very small effect size
 494 barely surviving our criterion of 0.5% of variance explained, and considering the potential
 495 misconceptions that could arise from a lengthy discussion of this difference, we decided to
 496 withhold hypothetical interpretations in this regard.

497

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

508 As done in the analysis of the vocabulary size, we applied the criterion of 0.5% variance
509 explained to successively eliminate two- and three-way interactions. Table 2 shows the estimates
510 for the final model, which explained almost 50% of the variance ($R^2 = 0.49$, $F = 8432.185$, $p <$
511 0.001 , 95% CI [0.48, 0.49]). In this model, frequency ($\beta = 1.06$, $p < 0.001$, 95% CI [1.03, 1.09]),
512 length ($\beta = 1.22$, $p < 0.001$, 95% CI [1.19, 1.25]), and orthographic neighbourhood measured by
513 old20 ($\beta = -0.80$, $p < 0.001$, 95% CI [-0.83, -0.78]) significantly predicted average accuracy.
514 Furthermore, frequency showed a significant interaction with both length ($\beta = -1.28$, $p < 0.001$,
515 95% CI [-1.33, -1.23]), and old20 ($\beta = 0.82$, $p < 0.001$, 95% CI [0.77, 0.86]). Overall, the longer
516 and more frequent a word is, the easier it is to recognize it. However, the fewer neighbours it has,
517 the harder it is to recognize.

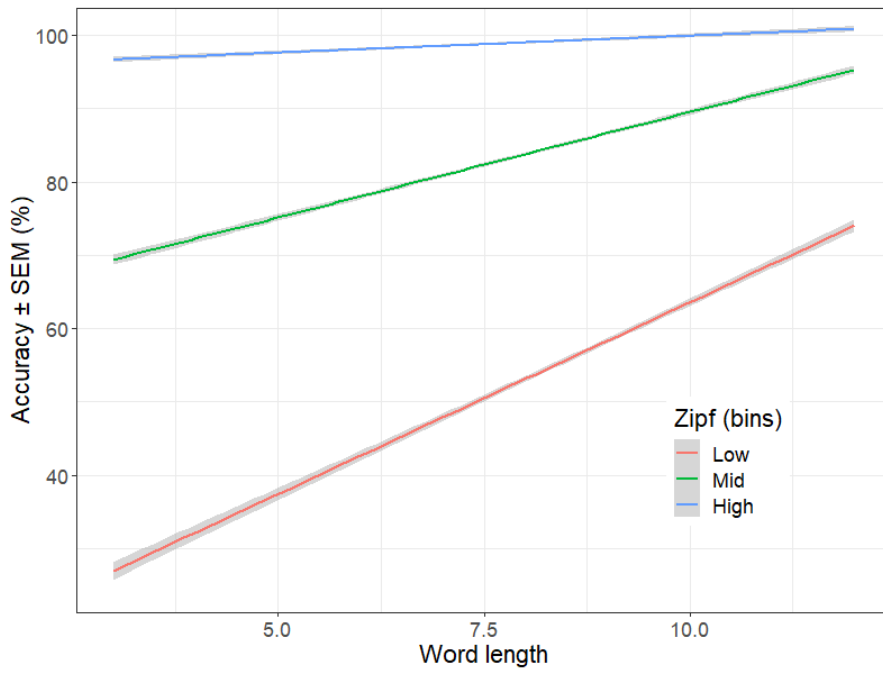
518

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

Predictor	<i>b</i>	<i>b</i>		<i>beta</i>	<i>sr</i> ²	<i>sr</i> ²		<i>r</i>	Fit
		95% CI	[LOW, HIGH]			95% CI	[LOW, HIGH]		
(Intercept)	-0.23**		[-0.24, -0.21]						
Zipf	0.26**		[0.26, 0.27]	1.06	0.07		[0.07, 0.07]	0.59**	
Length	0.16**		[0.16, 0.17]	1.22	0.07		[0.07, 0.08]	0.15**	
Old20	-0.26**		[-0.27, -0.25]	-0.80	0.04		[0.03, 0.04]	-0.01*	
Zipf * Length	-0.04**		[-0.04, -0.04]	-1.28	0.03		[0.03, 0.03]		
Zipf * Old20	0.07**		[0.06, 0.07]	0.82	0.02		[0.01, 0.02]		
									$R^2 = 0.485^{**}$
									95% CI [.48, .49]

520 *Note.* A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents
521 unstandardized regression weights. *beta* indicates the standardized regression weights. *sr*² represents the semi-partial
522 correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a
523 confidence interval, respectively. Zipf indicates zipf transformed frequency. Old20 indicates orthographic
524 neighborhood. * indicates $p < .05$. ** indicates $p < .01$.
525

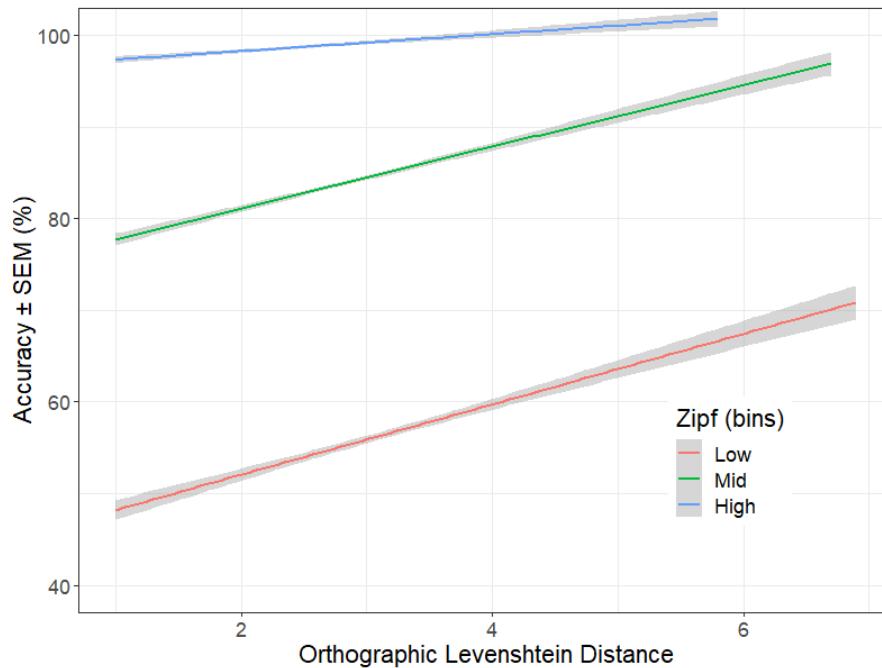
526 Figure 7 shows the interaction between word length and frequency. For high-frequency
527 words, length seems to become almost irrelevant in correctly recognising the word. On the other
528 hand, word length seems to aid word recognition for lower frequency words. This interaction has
529 been previously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



536

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

540



541

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

545

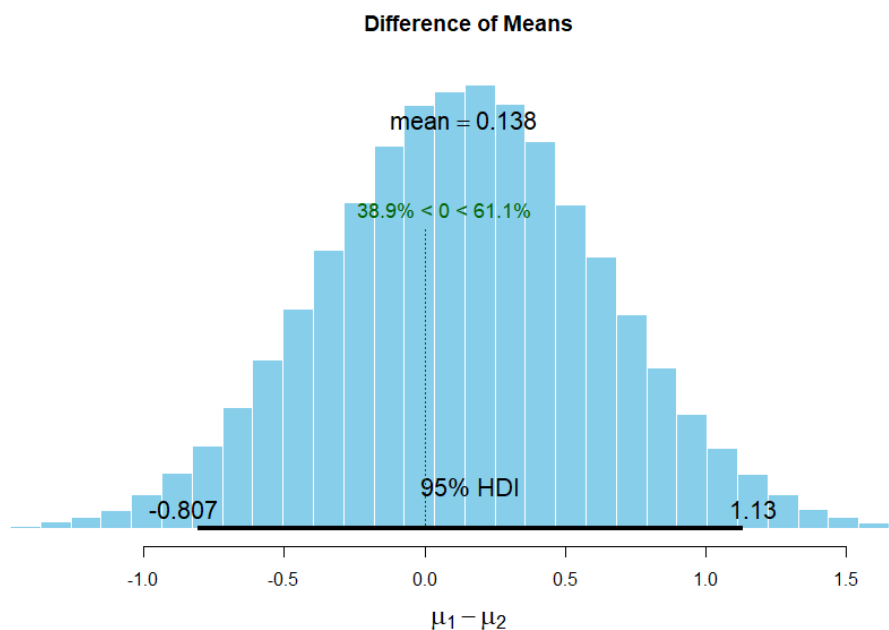
546 **Vocabulary size in bilingual and monolingual communities within Spain**

547

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

556 monolingual group consisted of 794 participants, and the bilingual group included 885
557 participants.

558 The scores for both groups were subjected to a Bayesian t-test using the BEST package in
559 R (Kruschke, 2013). We opted for a Bayesian framework because it provided a robust test of the
560 differences between the groups, while also being able to test for the null hypothesis of no
561 differences. We used the defaults of the BEST package, which assumes a t distribution as the
562 descriptive model of the data and uses a non-informative prior that is updated with each
563 observation to compute the posterior distributions for the means and standard deviations of both
564 groups, as well as a parameter for normality (5 parameters in total) that are sampled using a
565 Markov Chain Monte Carlo (MCMC) process (Kruschke, 2013). Figure 9 shows the results of
566 the analysis, indicating that vocabulary size in monolingual communities ($M = 69.6$, $SD = 10.2$)
567 did not differ significantly from that in bilingual communities ($M = 69.5$, $SD = 10.1$). The Bayes
568 factor for this analysis indicated strong support for the null hypothesis of no differences between
569 the groups ($BF_{10} = 0.056$). Additionally, the frequentist counterpart showed a similar result ($t =$
570 0.220 , $p = 0.826$).



571

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

573

574

Discussion

575

576 The present study aimed to examine Spanish vocabulary knowledge in a heterogeneous

577 sample of native speakers collected through a massive online LDT. We discuss the results by

578 focusing on the individual factors of the readers that directly affect visual word processing, after

579 briefly summarising the impact of the words' properties in lexical access and vocabulary

580 knowledge.

581 As expected, the frequency with which individuals are exposed to specific words

582 influences how accurately they recognise them. Furthermore, we found an interaction between

583 frequency and length and frequency and orthographic neighbourhood size on word accuracy.

584 Overall, while high-frequency words are correctly recognised irrespective of their length, for

585 low-frequency words, the longer they are, the more accurate participants are at recognising them.

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

597

598 **How do individual differences determine vocabulary size?**

599

600 **Age.** Age effects on vocabulary measures have traditionally reported a decrease in
601 performance for middle- and older-aged individuals (McCabe et al., 2010). Our approach
602 allowed us to test vocabulary across a wide range of ages and words, and the results, in
603 conjunction with Keuleers et al. (2015), suggest that vocabulary knowledge keeps increasing
604 with age in a seemingly logarithmic fashion. This logarithmic trend has also been corroborated in
605 previous simulation studies (Ramscar, Hendrix, Shaoul, et al., 2014). The simple explanation is
606 that, with time, individuals have more probability of encountering and learning novel words.
607 While it is true that some of the previous studies have reported a decline with age in vocabulary
608 knowledge, it is worth noting that they often have used productive vocabulary measures (e.g.,

609 Boston Naming Test; see MacKay, Connor, & Storandt, 2005; Simos, Kasselimis, & Mouzaki,
610 2011).

611 Why do we see these discrepancies? A first explanation might be that the mechanisms
612 required for word recognition do not seem to be affected by age as those required for word
613 production. This would be an interesting topic for further exploration. Nevertheless, an
614 alternative is that most psychometric tests assume that vocabulary is age-invariant, and thus try
615 to extrapolate vocabulary size from a limited set of words in the language, leading to an overall
616 underrepresentation of the effect of age on vocabulary size (Ramscar, Hendrix, Shaoul, et al.,
617 2014). Thus, by using the megastudy approach, we avoid most of the limitations by using a large
618 set of words and assessing vocabulary size across a heterogeneous population.

619

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

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

633
634 **Education.** The robustness of the effect of education level on lexical or semantic access
635 is perhaps one of the reasons why most studies try to control for this variable (Simos et al.,
636 2011). Our results confirm that vocabulary size increases with education. This is to be expected
637 given that higher education level also allows the opportunity to acquire lower frequency words
638 (Tainturier et al., 1992). These results exemplify two important points. The first is the contextual
639 opportunity that higher education offers individuals (Jones, Dye, & Johns, 2017). The likelihood
640 of encountering new words depends highly on the context in which they appear. For instance,
641 corpora analyses show that only the most frequent words appear across all texts, but more than
642 99% of the vocabulary is conditional on contextual factors (Jones et al., 2017). In this case, while
643 the vocabulary size of an individual with a degree in physics and another one with a degree in
644 psychology might contain a lot of overlapping words, a big part of the words they know will be
645 highly dependent on the degree of their choosing, even though the overall vocabulary size
646 appears to be similar (see also Ramscar, Hendrix, Love, et al., 2014). However, both of these
647 individuals will have an increased vocabulary size when compared to individuals with a high-
648 school education level. A larger variety of contexts in which one lives results in a larger number
649 of words known.

650 The second point relates to conscientiousness. Individuals with higher education level
651 might be more aware and careful of their responses, trying to reduce guessing in these types of
652 tasks, which in turn can lead to fewer false alarms, and overall increased performance
653 (Biderman, Nguyen, & Sebren, 2008), especially in an untimed LDT. A brief examination of the

654 data indicates a small but negative correlation between education level and the rate of false
655 alarms in our test, but also a positive correlation with a raw score for words, supporting both of
656 the previously posed arguments.

657

658 **Multilingualism.** The common conception of the effect of multilingualism on vocabulary
659 size is that multilingual individuals are less exposed to words in any of the languages they know
660 (Gollan et al., 2008). If so, the natural prediction is that multilinguals will show decreased
661 vocabulary size as compared to a native speaker of the language (Gollan & Acenas, 2004; Gollan
662 et al., 2008). Previous research with monolingual and bilingual adults and children shows that
663 there is a consistent difference in both productive and receptive vocabulary that does not vary
664 with the language pair of the bilinguals (Bialystok & Luk, 2012; Bialystok, Luk, Peets, & Yang,
665 2010; De Houwer, Bornstein, & Putnick, 2012). Despite this, our results indicate that the
666 knowledge of multiple languages increases Spanish vocabulary size rather than decreasing it.
667 Keuleers et al. (2015) offer a possible explanation for this, suggesting that, because some
668 languages share a big percentage of their vocabulary, the lack of exposure to L1 vocabulary
669 might be compensated indirectly by learning novel vocabulary in a different language. In the
670 case of Spanish and due to its close relation to other romance languages like French, Portuguese,
671 and Italian, indirect vocabulary acquisition might explain increased vocabulary knowledge. Here
672 again, a likely mechanism is that knowledge of various languages increases the variety of
673 contexts in which people learn specific vocabularies.

674 When contrasting different regions within Spain based on their multilingual status, our
675 results indicate moderate evidence towards the null hypothesis, suggesting that there are no
676 reliable differences in vocabulary size between these regions, regardless of the number of

677 languages used at the official level. Bilingual educational policies have been in place for more
678 than 20 years in autonomous communities like Catalonia and the Basque Country, and yet a
679 common criticism has been that students in these communities would not perform on par with
680 students from monolingual communities when their level of Spanish is assessed (Huguet, 2007).
681 While we acknowledge that our assessment of vocabulary size does not encompass other forms
682 of linguistic competence, such as production or comprehension, we did not observe differences
683 between monolingual and bilingual communities in vocabulary size.

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

694

695 **Conclusion**

696

697 The current study offers valuable data regarding individual word processing in Spanish
698 on the largest data collection conducted so far in this language. We tested a large number of
699 participants of varying origins and with different sociodemographic backgrounds, and a

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

721

722 **Acknowledgements:** This research is supported by the Basque Government through the BERC
723 2018-2021 program and by the Spanish State Research Agency through BCBL Severo Ochoa
724 excellence accreditation SEV-2015-0490. This study was also partially supported by grants
725 PGC2018-097145-B-I00, RED2018-102615-T and RTI2018-093547-B-I00 from the Spanish
726 State Research Agency. Work by JA was supported by “la Caixa” Foundation and the European
727 Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie
728 grant agreement no. 713673, and fellowship code LCF/BQ/IN17/116200154004.

729

730 **Open Practices Statement**

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

733

734

735

References

736 Acha, J., & Perea, M. (2008). The effects of length and transposed-letter similarity in lexical

737 decision: Evidence with beginning, intermediate, and adult readers. *British Journal of*738 *Psychology*. <https://doi.org/10.1348/000712607X224478>

739 Aguasvivas, J. A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A.

740 (2018). SPALEX: A Spanish Lexical decision database from a Massive Online Data

741 Collection. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2018.02156>

742 Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving

743 neighbourhood conflicts. *Psychonomic Bulletin and Review* 2, 4(4), 439–461.

744 Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004).

745 Visual word recognition of single-syllable words. *Journal of Experimental Psychology:*746 *General*. <https://doi.org/10.1037/0096-3445.133.2.283>

747 Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual Word Recognition: The Journey From

748 Features to Meaning (A Travel Update). In *Handbook of Psycholinguistics*.749 <https://doi.org/10.1016/B978-012369374-7/50010-9>

750 Barton, J. J. S., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length

751 effect in reading: A review. *Cognitive Neuropsychology*.752 <https://doi.org/10.1080/02643294.2014.895314>

753 Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive Control and Lexical Access in Younger and

754 Older Bilinguals. *Journal of Experimental Psychology: Learning Memory and Cognition*.755 <https://doi.org/10.1037/0278-7393.34.4.859>

- 756 Bialystok, E., & Luk, G. (2012). Receptive vocabulary differences in monolingual and bilingual
757 adults. *Bilingualism*. <https://doi.org/10.1017/S136672891100040X>
- 758 Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in
759 monolingual and bilingual children. *Bilingualism: Language and Cognition*.
760 <https://doi.org/10.1017/s1366728909990423>
- 761 Biderman, M. D., Nguyen, N. T., & Sebren, J. (2008). Time-on-task mediates the
762 conscientiousness-performance relationship. *Personality and Individual Differences*.
763 <https://doi.org/10.1016/j.paid.2007.10.022>
- 764 Bowles, R. P., & Salthouse, T. A. (2008). Vocabulary Test Format and Differential Relations to
765 Age. *Psychology and Aging*. <https://doi.org/10.1037/0882-7974.23.2.366>
- 766 Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word
767 Processing: An Updated Review. *Current Directions in Psychological Science*.
768 <https://doi.org/10.1177/0963721417727521>
- 769 Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms
770 for 62,000 English lemmas. *Behavior Research Methods*. [https://doi.org/10.3758/s13428-](https://doi.org/10.3758/s13428-018-1077-9)
771 [018-1077-9](https://doi.org/10.3758/s13428-018-1077-9)
- 772 Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of
773 current word frequency norms and the introduction of a new and improved word frequency
774 measure for American English. *Behavior Research Methods*.
775 <https://doi.org/10.3758/BRM.41.4.977>
- 776 Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016a). How many words do we

- 777 know? Practical estimates of vocabulary size dependent on word definition, the degree of
778 language input and the participant's age. *Frontiers in Psychology*.
779 <https://doi.org/10.3389/fpsyg.2016.01116>
- 780 Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016b). The impact of word prevalence
781 on lexical decision times: Evidence from the Dutch lexicon project 2. *Journal of*
782 *Experimental Psychology: Human Perception and Performance*.
783 <https://doi.org/10.1037/xhp0000159>
- 784 Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual
785 word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning*
786 *Memory and Cognition*. <https://doi.org/10.1037/0278-7393.23.4.857>
- 787 Casaponsa, A., Antón, E., Pérez, A., & Duñabeitia, J. A. (2015). Foreign language
788 comprehension achievement: Insights from the cognate facilitation effect. *Frontiers in*
789 *Psychology*. <https://doi.org/10.3389/fpsyg.2015.00588>
- 790 Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In
791 *Attention and Performance VI*. <https://doi.org/10.1006/brln.2001.2475>
- 792 Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word
793 frequencies based on film subtitles | SUBTLEX-ESP: Frecuencias de las palabras españolas
794 basadas en los subtítulos de las películas. *Psicologica*.
- 795 Davies, R. A. I., Birchenough, J. M. H., Arnell, R., Grimmond, D., & Houlson, S. (2017).
796 Reading through the life span: Individual differences in psycholinguistic effects. *Journal of*
797 *Experimental Psychology: Learning Memory and Cognition*.
798 <https://doi.org/10.1037/xlm0000366>

- 799 Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and
800 phonological neighborhood statistics and other psycholinguistic indices in Spanish.
801 *Behavior Research Methods*. <https://doi.org/10.3758/BF03192738>
- 802 De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2012). A bilingual-monolingual comparison
803 of young children's vocabulary size: Evidence from comprehension and production.
804 *Applied Psycholinguistics*. <https://doi.org/10.1017/S0142716412000744>
- 805 Diependaele, K., Brysbaert, M., & Neri, P. (2012). How noisy is lexical decision? *Frontiers in*
806 *Psychology*. <https://doi.org/10.3389/fpsyg.2012.00348>
- 807 Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-
808 based word frequencies as the best estimate of reading behavior: The case of Greek.
809 *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2010.00218>
- 810 Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop
811 shopping for Spanish word properties. *Behavior Research Methods*.
812 <https://doi.org/10.3758/s13428-013-0326-1>
- 813 Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., ...
814 Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can
815 revolutionize research in cognitive science. *PLoS ONE*.
816 <https://doi.org/10.1371/journal.pone.0024974>
- 817 Duñabeitia, J. A., & Vidal-Abarca, E. (2008). Children like dense neighborhoods: orthographic
818 neighborhood density effects in novel readers. *Spanish Journal of Psychology*.
819 <https://doi.org/10.1017/S113874160000408X>

- 820 Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for
821 Geographic Lexical Variation. *In Proceedings of the 2010 Conference on Empirical*
822 *Methods in Natural Language Processing, MIT, Massachusetts, USA, 9-11 October 2010.*
- 823 Ethnologue. (2016). Summary by language size. Retrieved from
824 <https://www.ethnologue.com/statistics/size>
- 825 Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... Pallier, C. (2010). The
826 French lexicon project: Lexical decision data for 38,840 French words and 38,840 pseudo
827 words. *Behavior Research Methods*. <https://doi.org/10.3758/BRM.42.2.488>
- 828 Gierut, J. A., & Dale, R. A. (2007). Comparability of Lexical Corpora: Word frequency in
829 phonological generalization. *Clinical Linguistics and Phonetics*.
830 <https://doi.org/10.1080/02699200701299891>
- 831 Gollan, T. H., & Acenas, L. A. R. (2004). What Is a TOT? Cognate and Translation Effects on
832 Tip-of-the-Tongue States in Spanish-English and Tagalog-English Bilinguals. *Journal of*
833 *Experimental Psychology: Learning Memory and Cognition*. [https://doi.org/10.1037/0278-](https://doi.org/10.1037/0278-7393.30.1.246)
834 [7393.30.1.246](https://doi.org/10.1037/0278-7393.30.1.246)
- 835 Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always
836 means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis.
837 *Journal of Memory and Language*. <https://doi.org/10.1016/j.jml.2007.07.001>
- 838 González-Nosti, M., Barbón, A., Rodríguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the
839 psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765
840 words. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-013-0383-5>

- 841 Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and
842 naming. *Journal of Memory and Language*. [https://doi.org/10.1016/0749-596X\(90\)90074-A](https://doi.org/10.1016/0749-596X(90)90074-A)
- 843 Huguet, Á. (2007). Minority languages and curriculum: The case of Spain. *Language, Culture*
844 *and Curriculum*. <https://doi.org/10.2167/lcc327.0>
- 845 Huguet, Á., Lapresta, C., & Madariaga, J. M. (2008). A study on language attitudes towards
846 regional and foreign languages by school children in aragon, Spain. *International Journal of*
847 *Multilingualism*. <https://doi.org/10.1080/14790710802152412>
- 848 Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an Organizing Principle of the
849 Lexicon. *Progress in Brain Research*. <https://doi.org/10.1016/bs.plm.2017.03.008>
- 850 Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., & Whalen, D. H. (2012). What lexical
851 decision and naming tell us about reading. *Reading and Writing*.
852 <https://doi.org/10.1007/s11145-011-9316-9>
- 853 Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in
854 psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental*
855 *Psychology*. <https://doi.org/10.1080/17470218.2015.1051065>
- 856 Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior*
857 *Research Methods*. <https://doi.org/10.3758/BRM.42.3.627>
- 858 Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments
859 with the LD1NN algorithm. *The Mental Lexicon*. <https://doi.org/10.1075/ml.6.1.02keu>
- 860 Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual
861 word recognition studies: A lexical decision study on 14,000 dutch mono-and disyllabic

- 862 words and nonwords. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2010.00174>
- 863 Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical
864 decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research*
865 *Methods*. <https://doi.org/10.3758/s13428-011-0118-4>
- 866 Keuleers, E., & Marelli, M. (n.d.). *Resources for mental lexicon research: A delicate ecosystem*.
- 867 Keuleers, E., Stevens, M., Mander, P., & Brysbaert, M. (2015). Word knowledge in the crowd:
868 Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly*
869 *Journal of Experimental Psychology*. <https://doi.org/10.1080/17470218.2015.1022560>
- 870 Kosmidis, M., Tsapkini, K., & Folia, V. (2006). Lexical processing in illiteracy: Effect of
871 literacy or education? *Cortex*, 42(7), 1021–1027. <https://doi.org/10.1016/S0010->
872 9452(08)70208-9
- 873 Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental*
874 *Psychology. General*. <https://doi.org/10.1037/a0029146>
- 875 Kulkarni, V., Perozzi, B., & Skiena, S. (2016). Freshman or Fresher? Quantifying the
876 Geographic Variation of Language in Online Social Media. *Proceedings of the Tenth*
877 *International AAAI Conference on Web and Social Media*.
- 878 Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word
879 recognition for skilled and unskilled readers. *Journal of Experimental Psychology. Human*
880 *Perception and Performance*. <https://doi.org/10.1037/a0030859>
- 881 Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals.
882 *Soviet Physics Doklady*. <https://doi.org/10.1023/A:1022689900470>

- 883 MacKay, A., Connor, L. T., & Storandt, M. (2005). Dementia does not explain correlation
884 between age and scores on Boston Naming Test. *Archives of Clinical Neuropsychology*.
885 <https://doi.org/10.1016/j.acn.2004.03.006>
- 886 McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The
887 relationship between working memory capacity and executive functioning: Evidence for a
888 common executive attention construct. *Neuropsychology*. <https://doi.org/10.1037/a0017619>
- 889 Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis.
890 *Psychological Review*. <https://doi.org/10.1037/0033-295X.111.3.721>
- 891 Myers, J. (2016). Meta-megastudies. *The Mental Lexicon*. <https://doi.org/10.1075/ml.11.3.01mye>
- 892 Pollatsek, A., Perea, M., & Binder, K. S. (1999). The effects of “neighborhood size” in reading
893 and lexical decision. *Journal of Experimental Psychology: Human Perception and*
894 *Performance*. <https://doi.org/10.1037/0096-1523.25.4.1142>
- 895 Portocarrero, J. S., Burright, R. G., & Donovan, P. J. (2007). Vocabulary and verbal fluency of
896 bilingual and monolingual college students. *Archives of Clinical Neuropsychology*.
897 <https://doi.org/10.1016/j.acn.2007.01.015>
- 898 Ramscar, M., Hendrix, P., Love, B., & Baayen, H. (2014). Learning is not decline. *The Mental*
899 *Lexicon*. <https://doi.org/10.1075/ml.8.3.08ram>
- 900 Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive
901 decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*.
902 <https://doi.org/10.1111/tops.12078>
- 903 Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The Mismeasurement of Mind: Life-

- 904 Span Changes in Paired-Associate-Learning Scores Reflect the “Cost” of Learning, Not
905 Cognitive Decline. *Psychological Science*. <https://doi.org/10.1177/0956797617706393>
- 906 Schröter, P., & Schroeder, S. (2017). The Developmental Lexicon Project: A behavioral database
907 to investigate visual word recognition across the lifespan. *Behavior Research Methods*.
908 <https://doi.org/10.3758/s13428-016-0851-9>
- 909 Simos, P. G., Kasselimis, D., & Mouzaki, A. (2011). Age, gender, and education effects on
910 vocabulary measures in Greek. *Aphasiology*.
911 <https://doi.org/10.1080/02687038.2010.512118>
- 912 Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., & Baltes, P. B. (2003). The fate of
913 cognition in very old age: six-year longitudinal findings in the Berlin Aging Study (BASE).
914 *Psychology and Aging*.
- 915 Singh-Manoux, A., Kivimaki, M., Glymour, M. M., Elbaz, A., Berr, C., Ebmeier, K. P., ...
916 Dugravot, A. (2012). Timing of onset of cognitive decline: Results from Whitehall II
917 prospective cohort study. *BMJ (Online)*. <https://doi.org/10.1136/bmj.d7622>
- 918 Solomon, R. L., & Howes, D. H. (1951). Word frequency, personal values, and visual duration
919 thresholds. *Psychological Review*. <https://doi.org/10.1037/h0058228>
- 920 Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of
921 valence and arousal for 14,031 Spanish words. *Behavior Research Methods*.
922 <https://doi.org/10.3758/s13428-015-0700-2>
- 923 Tainturier, M. J., Tremblay, M., & Lecours, A. (1992). Educational level and the word frequency
924 effect: A lexical decision investigation. *Brain and Language*. <https://doi.org/10.1016/0093->

- 925 934X(92)90112-R
- 926 van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A
927 new and improved word frequency database for British English. *Quarterly Journal of*
928 *Experimental Psychology*. <https://doi.org/10.1080/17470218.2013.850521>
- 929 Verhaeghen, P. (2003). Aging and vocabulary scores: A meta-analysis. *Psychology and Aging*.
930 <https://doi.org/10.1037/0882-7974.18.2.332>
- 931 Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of
932 orthographic similarity. *Psychonomic Bulletin and Review*.
933 <https://doi.org/10.3758/PBR.15.5.971>
- 934