Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size

Cristina Izura[1*], Fernando Cuetos[2] and Marc Brysbaert[3]

[1]Department of Psychology, Swansea University,

Singleton Park, Swansea, SA2 8PP, UK

[2]Departamento de Psicología, Universidad de Oviedo,

Plaza Feijoó s/n, 33003, Oviedo, Spain

[3]Department of Experimental Psychology, Ghent University,

Henri Dunantlaan 2, B-9000 Gent, Belgium

*Corresponding Author: Izura Cristina, Department of Psychology, Swansea University at Singleton Park, Swansea SA2 8PP, Wales, UK. Tel.: +441792-513344; fax: +441792-295679. E-mail address c.izura@swansea.ac.uk

Lextale-Esp: Un test para la rápida y eficaz evaluación del tamaño del vocabulario en español.

Cristina Izura[1*], Fernando Cuetos[2] and Marc Brysbaert[3]

[1]Department of Psychology, Swansea University,

Singleton Park, Swansea, SA2 8PP, UK

[2]Departamento de Psicología, Universidad de Oviedo,

Plaza Feijoó s/n, 33003, Oviedo, Spain

[3]Department of Experimental Psychology, Ghent University,

Henri Dunantlaan 2, B-9000 Gent, Belgium

*Corresponding Author: Izura Cristina, Department of Psychology, Swansea University at Singleton Park, Swansea SA2 8PP, Wales, UK. Tel.: +441792-513344; fax: +441792-295679. E-mail address c.izura@swansea.ac.uk

Abstract

The methods to measure vocabulary size vary across disciplines. This heterogeneity hinders direct comparisons between studies and slows down the understanding of research findings. A quick, free and efficient test of English language proficiency, LexTALE, was recently developed to remedy this problem. LexTALE has been validated and shown to be an effective tool for distinguishing between different levels of proficiency in English. The test has also been made available in Dutch, German, and French. The present study discusses the development of a Spanish version of the test: Lextale-Esp. The test discriminated well at the high and the low end of Spanish proficiency and returned a big difference between the vocabulary size of Spanish native and non-native speakers.

Keywords: Spanish proficiency test, vocabulary size, first language, and second language

Resumen

Los métodos para medir el tamaño del vocabulario varían según las disciplinas. Esta heterogeneidad dificulta las comparaciones entre estudios y enlentece la comprensión de los hallazgos. Para remediar este problema, recientemente ha sido desarrollado un test de competencia lingüística en inglés que es rápido, eficaz y gratis, el LexTALE. El LexTALE ha sido validado y ha demostrado ser una herramienta eficaz para distinguir entre distintos niveles de competencia lingüística en inglés. El test también se ha realizado en holandés, alemán y francés. El presente estudio presenta la versión española del test; Lextale-Esp. El test mostró una buena discriminación entre los niveles altos y bajos de competencia en español y reveló grandes diferencias entre el tamaño de vocabulario de nativos y no nativos.

Palabras clave: Test de competencia lingüística en español, tamaño del vocabulario, primer idioma y segundo idioma.

Introduction

Measuring language proficiency is important for educators and researchers. Two critical aspects are vocabulary size and grammatical knowledge. As vocabulary size provides valuable information for teaching (e.g., learning progress, motivation, best level to start a program with, etc.), is rather easy to measure, and is particularly interesting for researchers interested in word recognition, many existing language tests focus on this variable.

Schmitt (2000) gives a review of the vocabulary tests developed over the years including those that were not validated. Arguably the best known test for English vocabulary is the Vocabulary Levels Tests (VLT, Nation, 1990). It is a test based on word frequency. It estimates language proficiency on the basis of the number of words correctly identified at five different frequency levels, defined by ranking the words from most frequent to least frequent and grouping them in bands of 1000 words. The VLT includes words from the second, the third, the fifth, and the tenth band, together with a group of words typical in academia. It comprises 60 words per level, which are presented in sets of six. The learner has to match three out of the six presented words with one of the three definitions provided. The test works well but loses some discriminatory power at the high proficiency end.

Another well-known test for English vocabulary size among second language teachers is The Eurocentres Vocabulary Size Test (EVST; Meara & Jones, 1987, 1990). The EVST is a computerized test based on ten frequency bands of 1000 word each (from

Thorndike & Lorge, 1944). The test uses the lexical decision paradigm and consists of 150 items. Two thirds of the items are real words and one third invented nonwords. Test items are intermixed randomly. Participants have to indicate which words they know. The nonwords are used to correct for response bias (i.e., saying one is familiar with a word that cannot be known). The variation of word frequencies in the EVST is large enough to include words that are unfamiliar even to native speakers (such as *myosote, leat*, and *algorism*). The final score is automatically generated by the programme following a relatively complex assessment, because the test gauges word knowledge in a gradual way. It starts with the easiest (most frequent) words presenting a sample of 10 words and 5 nonwords. If the participant's performance is high enough, the programme goes on to assess word knowledge from the next frequency band and so on until accuracy falls below a pre-specified criterion. When that happens, a rough score is computed based on the accuracy observed in the last two frequency bands (e.g. if the participant's accuracy is 100% up to frequency band 5 and then decreases drastically, the assumption made is that the participant knows between 5,000 and 6,000 words). At that point, the way of testing changes towards a more detailed assessment by presenting words from the frequency band at which accuracy started to decline.

The equation used in the EVST also considers overestimation of the number of words known by adjusting the final score for the number of nonwords that were responded to positively, following signal detection theory (Zimmerman, Broder, Shaughnessy, & Underwood, 1977). The test was commissioned by a group of schools that provided short and intensive courses of English as second language (L2) and that needed a

quick placement test. The high correlations between the EVST scores and measures of reading comprehension, listening comprehension and grammatical accuracy indicated that the test was able to correctly classify students in the appropriate proficiency levels. The results of EVST correlated highly with VLT (Mochida & Harrington, 2006; but see Cameron, 2002; Meara & Jones, 1988, for some cautionary notes).

As observed by Lemhöfer and Broersma (2012), vocabulary tests do not seem to be well-known among psycholinguistic language researchers. Most studies on word recognition do not include information about the language proficiency of their participants. This is particularly the case for studies in the native language (L1), which seem to be based on the assumption that first-year students form a homogeneous population without interesting variation (see Andrews & Hersch, 2010; Chateau & Jared, 2000; Diependaele, Lemhöfer, & Brysbaert, 2013; Yap, Balota, Tse, & Besner, 2008, for counterevidence). Proficiency differences are acknowledged more in research on second-language (L2) processing. However, the standard way to assess proficiency here is to make use of self-assessments or language history questionnaires (e.g., Dunn & Fox Tree, 2009; Li, Sepanski, & Zhao, 2006).

Lemhöfer and Broersma (2012) presented the Lexical Test for Advanced Learners of English (LexTALE) as a new and validated test of vocabulary knowledge in English at rather high proficiency levels. It is based on the EVST and includes 60 items (40 words and 20 nonwords) for which the test takers have to indicate whether or not they know the word. Both EVST and LexTALE aim to measure language proficiency by

estimating vocabulary size, but at different levels. EVST is meant to place beginning students in the right grade; LexTALE was designed as a standard tool to assess language proficiency of participants in psycholinguistic experiments. Both tests use word frequency as the basic criterion for words of various difficulty levels. Words were selected in such a way that some should be known to participants with low proficiency levels, whereas others are known only to participants with high proficiency levels. Because it is expected that most participants will not know all words, the number of nonwords is smaller than the number of words (typically in a ratio of 1 to 2). To compute the final score, both tests take into account the number of words correctly identified and the number of false positives, that is the nonwords that are "recognized" as existing words.

A difference between EVST and LexTALE is that the latter is easier to administer. Participants are simply given the full list of stimuli and their score is calculated on the basis of the number of words and nonwords selected. EVST requires access to the computer program for the adaptive presentation of stimulus materials. The LexTALE scores have been validated by correlating them with word translation scores and the scores of a commercial language test (the Quick Placement Test; Lemhöfer and Broersma, 2012).

Further evidence for the usefulness of LexTALE was provided by Diependaele et al. (2013). They observed that participants with low scores on the test had a much steeper word frequency effect in a visual word recognition experiment than participants with high scores (see Yap et al., 2008, for a similar finding). Furthermore, differences in

vocabulary size entirely accounted for the observation that people have a larger word frequency effect in L2 than in L1 (i.e., once differences in vocabulary size were taken into account, there was no distinction in the word frequency effect between L2 and L1 any more). In a related study, Khare, Verma, Kar, Srinivasan, and Brysbaert (2013) used the LexTALE scores in an attempt to replicate and extend a finding reported by Colzato, Bajo, Wildenberg, Paolieri, Nieuwenhuis, La Heij and Hommel (2008). These authors investigated the attentional blindness phenomenon (i.e., the finding that when participants are asked to identify two targets in a rapid series of visual stimuli, they often fail to report the second target if it occurs between 100-500 ms after the first target). They observed that the attentional blindness effect was stronger in bilinguals than in monolinguals. Khare et al. (2013) examined whether this implied that the effect would also be stronger in highly proficient Hindi-English bilinguals than in less proficient bilinguals. The authors indeed observed the expected correlation, but only when English proficiency was measured with LexTALE, not when it was measured with a self-assessment questionnaire.

Lemhöfer and Broersma (2012) further developed LexTALE tests for Dutch and German (see www.lextale.com), which has the advantage of the potential standardization across languages. Unfortunately, these tests have not yet been normed or validated. Inspired by the findings with the English LexTALE, Brysbaert (2013) compiled an analogue test for French, which he called LEXTALE_FR. This test included 84 items (56 words, 28 nonwords) rather than the original 60, to further increase the reliability of the test and to better cover the entire range of language proficiencies, so that the same test could be used for L1 and L2 speakers. The latter

was checked by presenting the test to L1 and L2 speakers of French. Both groups showed a healthy variance in performance, with no signs of a floor effect for the L2 speakers or a ceiling effect for the L1 speakers.

There are several advantages to the LexTALE tests for language researchers. First, it is a fast and effective way of measuring vocabulary size. It takes three to five minutes to complete, is free, and can easily be administered online or in pen and paper format. Second, the use of LexTALE tests as the standard measure of vocabulary size allows direct comparisons between studies. At present, this is virtually impossible given the heterogeneity of measures used (or not used) in various labs. Third, it will make it easier for researchers to investigate individual differences both in language processing (Andrews & Hersch, 2010; Chateau & Jared, 2000; Diependaele et al., 2013; Yap et al., 2008) and in language-related skills such as cognitive control (Bialystok, Craik, & Luk, 2012; Khare et al., 2013).

In the present study we join the effort of standardising the way in which language proficiency is measured across languages by presenting the Spanish version of LexTALE. We take into account the suggestions of Brysbaert (2013) on how to improve the quality of the test by starting off with a slightly larger number of stimuli, which are tested by presenting them to a group of L1 speakers and a group of L2 speakers. Only the stimuli that score well are retained. Brysbaert (2013) noticed that in particular constructing suitable nonwords is a challenge. If they are too easy, one can do the test without knowing what the words mean (Grainger, Dufau, Montant, Ziegler, & Fagot, 2012; Keuleers & Brysbaert, 2011). On the other hand, if the

nonwords are too difficult, they create confusion and are more likely to be selected as words by L1 speakers than by L2 speakers. This is particularly the case for pseudohomophones of low-frequency words (such as *rithm* in English or adesivo in Spanish). These are misspellings of words that retain the phonology and that can only be rejected by participants with very good spelling skills. Because the L1 speakers know the meaning of the word *rhythm* or *adhesivo* referred to by the phonology but do not know the precise spelling, they are more likely to select this nonword as an existing word than L2 speakers who do not know the word. In order to be able to make a good selection of stimuli, we started off with 90 words and 90 nonwords, to end up with 60 good words and 30 good nonwords.

Method

Materials. Ninety words were selected from a Spanish database of word frequencies based on film subtitles, Subtlex-Esp (Cuetos, González-Nosti, Barbón, & Brysbaert, 2011). The frequency of the words ranged from very high, that is words likely to be known by new learners of Spanish (e.g., *ganar (to win), matar (to kill), playa (beach)*) to very low, which are words only familiar to proficient native speakers (e.g., *cenefa (edging), laud (lute), alpiste (birdseed)*). Overall, 26 words had a frequency of less than one occurrence per million words (pm), 23 had a frequency from one to five occurrences pm, 14 words had a frequency ranging between 6 to 10 occurrences pm, 17 words had frequencies from 11 to 20 pm, 8 words had frequencies between 21 and 100 pm, and two words (*ganar (to win)* and *matar (to kill))* had frequencies above 100 occurrences pm. The majority of words were nouns (n = 52), followed by verbs (n = 26) and adjectives (n = 12).

Next, a list of 90 nonwords was compiled. A number of nonwords came from previous lexical decision experiments we ran in Spanish (González-Nosti, Barbón, Rodríguez-Ferreiro, & Cuetos, F (under revision)). We selected nonwords that in general elicited some 10% errors. To fully match the nonwords to the words, we had to create some new stimuli. This was done on the basis of suggestions provided by the Wuggy algorithm (Keuleers & Brysbaert, 2010). Care was taken to include nonwords with similar endings to Spanish words from different syntactic categories; for instance, nonwords ending as Spanish verbs (*er, ar,* ir) or as Spanish adjectives (*oso, ado*). To ensure that the letter combinations of the nonwords could not be distinguished from the letter combinations of the words without lexical knowledge, we ran an LD1NN test on our stimulus list (Keuleers & Brysbaert, 2011). The LD1NN algorithm calculates whether the letter combinations of the nonwords resemble those of the other nonwords more than those of the words. Such was not the case for the stimuli we selected.

A random permutation was made of the list of words and nonwords. This permutation was presented to all participants in the same order.

Procedure. Following the procedure of Brysbaert (2013) we presented the stimulus list to a group of highly proficient Spanish L1 speakers and a group of Spanish L2 speakers. The L1 speakers were predominantly master students of psychology at the University of Oviedo in Spain, though a few other participants took part after hearing

about the study through word of mouth. This group contained 91 L1 speakers with a mean age of 24 years (range 20-50). The second group consisted of 123 Spanish L2 speakers mainly taking courses at the University of Swansea and the Artesis University College Antwerp[1] (mean age was 25 years; range 16-59). The first language of these participants varied as follows: 68 spoke English as L1, 19 Dutch, 8 French, 4 German, 7 Italian, 3 Romanian, 2 Portuguese, 1 Polish, 1 Slovakian, 1 Lithuanian, 1 Finnish, 1 Albanian, 1 Catalan, and 1 Chinese.

Words and nonwords were presented online using Survey Monkey software (http://www.surveymonkey.com). For each stimulus, participants were asked whether this was a Spanish word they knew. The instructions were as shown below. They were available in English for those participants for whom Spanish was the second language and in Spanish for the native speakers.

> *"Hi, this is a test of Spanish vocabulary. You will get 180 sequences of letters that look "Spanish". Only some of them are real words. Please, indicate the words you know (or of which you are convinced they are Spanish words, even though you would not be able to give their precise meaning). Be careful, however: Errors are penalised. So, there is no point in trying to increase your score by adding tallies to "words" you've never seen before!*
>
> *All you have to do is to tick the box next to the words you know. If, for instance, in the example below you recognise "sí", "sacapuntas", "bien", and "casa", you indicate this as follows:*

---

[1] The authors thank María Fernandez-Parra, Rocío Pérez-Tattam, Alicia San Mateo, Anne Verhaert and Katrien Lievois for their kind cooperation.

Estímulo	Palabra?
depiste	
sí	√
coné	
calpar	
joten	
sacapuntas	√

Estímulo	Palabra?
priba	
pelasula	
bien	√
casa	√
lejo	
pretantas	

The results of this test are only useful if you do not use a dictionary and if you work on your own! "

In addition each participant provided information about their gender, number of years they had taken Spanish courses in school, and their self-rated proficiency in Spanish (from 1 "nearly non-existent" to 10 "perfect").

Results

The quality of the test items was assessed first by reviewing the responses to the items using point-biserial correlation and Item Response Theory. A second series of analyses looked at the participants´ responses as a group, providing us with a Cronbach alpha measure of reliability and a measure of criterion validity by comparing the performance of L1 and L2 speakers. These results are described successively.

Item Assessment

The quality of each word and nonword was examined first by computing the point-biserial correlation between the responses to the item and the participants' total

scores. This type of correlation varies between -1.0 and +1.0. A positive point-biserial correlation is expected, as it indicates that a good test performer also performs better on the item than a bad test performer. In contrast, a negative correlation signals an anomaly, because good participants are doing less well on the item than weak participants. All items tested, except one, had a positive correlation (going from r = 0.04 for the non-word *bial* to r = .80 for the word *musgo* (*moss*)). The exception was the non-word *botezar*, which yielded a negative correlation, meaning that it was more likely to be selected as a "word" by participants with a high proficiency score than participants with a low score. In order to achieve high test reliability, it is recommended to remove such negative items before further analyses are run.

A good test contains items equally spread across the entire difficulty range and with good discrimination power. An ideal technique for this, when all items are assumed to measure the same competence (language proficiency), is based on item response theory (IRT). An IRT analysis allows researchers to see how items are responded to throughout the ability range. This gives an idea of the difficulty and the discrimination power of an item (the discrimination power refers to the steepness of the item response curve going from not-known at the low end of the ability range to known at the high end of the ability range). It takes into account both the performance levels of the individuals and the difficulty of the item and, therefore, is more powerful than the point-biserial correlation, because it provides a measure of item difficulty in addition to item quality. We used the R package *ltm* (Rizopoulos, 2006). Figure 1 shows the outcome for a few stimuli. On the basis of the IRT analysis, 60 words and 30 nonwords of various difficulty levels with good discrimination power were selected

(see under Availability). This was done by ordering the items according to difficulty level and taking the items with the best discrimination power at approximately each $1/30^{th}$ of the range covered by the items. Descriptive information related to the final selection of words and nonwords can be found in Table 1.

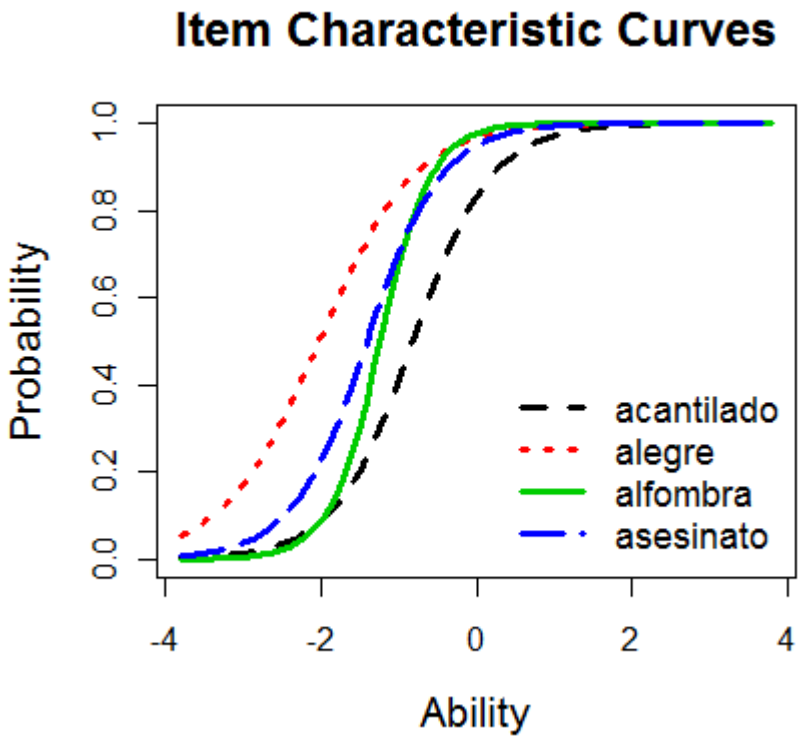

Figure 1. Outcome of an IRT analysis provides interesting information to select stimuli. In this figure, the abscissa represents the language proficiency level (going from low to high), and the ordinate shows the estimated probability of participants knowing the item. Easy items are already known by people with low proficiency levels; hard items require higher proficiency levels. So, the word *alegre* (cheerful) is easier than *acantilado* (cliff). The steepness of the curve indicates how high the discrimination power of the item is. The discrimination power is larger for *alfombra* (carpet) than *asesinato* (murder), possibly because some people at the low end recognize the cognate *assassin* in *asesinato*.

Table 1: Lexical information of the final set of 60 words and 30 nonwords selected to be part of the Lextale-Esp.

	Words	Nonwords
Mean number of letters	6.41	6.63
Mean number of syllables	2.67	2.83
Mean number of phonemes	6.16	6.49
Mean number of orthographic neighbours	6	
Levenshtein´s distance	1.70	
Word frequency (Logarithm +1)	1.85	

Note: Mean number of orthographic neighbours and Levenshtein´s distance from EsPal (Duchon, Peréa, Sebastián-Gallés, Martí, & Carreiras, in press). Word frequency from Subtlex-Esp (Cuetos, González-Nosti, Barbón, & Brysbaert, 2011).

Comparison between groups

Scoring the Lextale-Esp

In line with recommendations made by Lemhöfer and Broersma (2012) and Brysbaert (2013), the test score was defined as:

$$\text{Score} = N_{\text{yes to words}} - 2 * N_{\text{yes to nonwords}}$$

So, a person with 38 Spanish words correct and 5 nonwords erroneously selected as known words, would get a score of 38 – 2*5 = 28. This score accurately penalizes for guessing behaviour, as a test taker who responds randomly (i.e. saying yes to half of the words and half of the nonwords) is expected to have a score around 0. A zero score would also be the outcome of someone responding “yes” to all the items. As it happens, test takers can even obtain a negative score if they are more likely to select nonwords from the list as “known” Spanish words than existing words (a score some

of our L2 participants obtained). Only someone who has all the words correct and did not selected any nonword, gets the maximum score of 60.[2]

The L1 group had a mean score of 53.9 (SD = 6.6; range = 34 to 60). The L2 group had an average score of 11.9 (SD = 17.9; range -16 to 58). This difference is in line with the difference observed by Brysbaert (2013) on the French test.

Figure 2 shows the correlation between the scores on the Lextale test and the self-assessment ratings. Although the correlation is substantial ($r = .82$, $N = 214$, $p < .001$), there are some large divergences for individual participants. Particularly noteworthy are the L2 speakers who give themselves ratings above 6 but still score rather low compared to L1 speakers with the same proficiency ratings. This suggests that L2 speakers use a different criterion for self-assessment than L1 speakers. Similarly, among the L1 speakers participants gave themselves ratings from 6 to 10 although their performance on average was quite similar. The correlation between the Lextale scores and self-assessment was $r = .73$ ($N = 123$) for the L2 group and .10 ($N = 91$) for the L1 group. The low value of the latter group was due to the fact that the L1 speakers were a homogeneous group, all having quite high scores.

[2] For those who like to convert this score to 100, an easy equation is $\%_{\text{yes to words}} - \%_{\text{yes to nonwords}}$ (38 out of 60 words correct is 63.33%; 5 out of 30 nonwords wrong is 16.67%; so the total score is 63.33-16.67 = 46.66%, which equals 28/60)

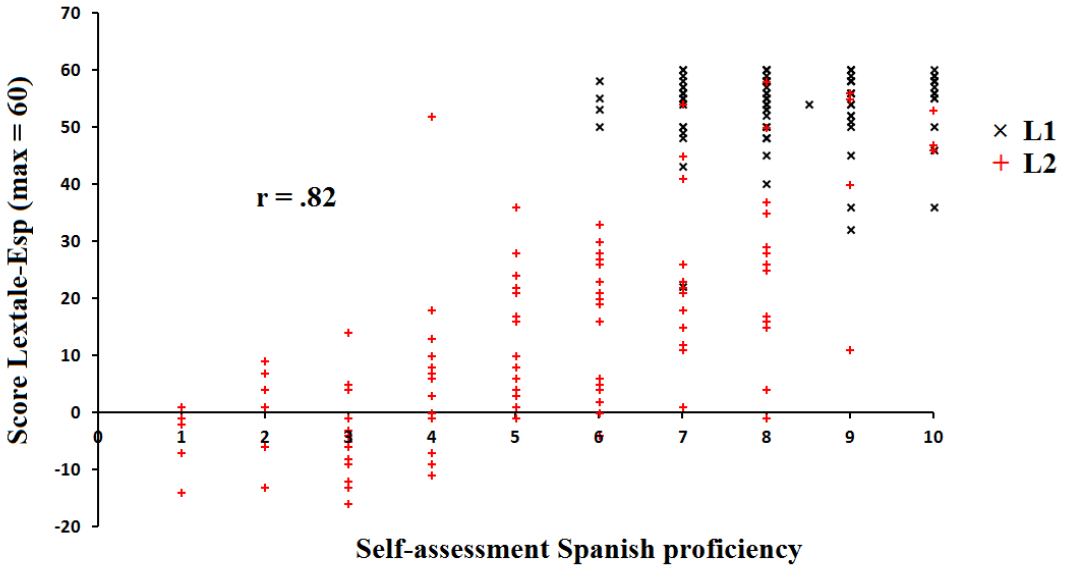

Figure 2: Correlation between self-rating of proficiency in Spanish and the Lextale-Esp score obtained. L2 speakers who rated their proficiency lower than 5 indeed did not know many words; L1 speakers giving themselves ratings of 9 and 10 in general scored well (even though some had less than 40/60). In-between there was more variability.

The reliability of the test was measured with Cronbach's alpha. This gave a value of α = .96. This is a high value, although it must be taken into account that two extreme groups were compared. Still, when the data were limited to the L1 group, reliability remained at a high level of α = .88, and it stayed at α = .96 if the analysis was limited to the L2 group.

To make sure that our findings were not contaminated by the items that were deleted after the IRT analysis, we administered the selected list of 90 stimuli again to a group of L1 speakers and a group of L2 speakers. The L1 group consisted again of

psychology students from the University of Oviedo (N = 102; mean age = 22 years; range 17-58). The L2 group was a group of students having had 2.5 months of Spanish courses at Ghent University at the time of testing[3] (N = 100). So, they were really beginning Spanish L2 speakers, although several of them had taken some Spanish lessons in secondary education or in evening classes. The vast majority rated their proficiency between 1 and 3. Most had Dutch as L1; a few had Russian (2), Bosnian (1), and English (1).

The L1 group completed the online version. The L2 group was given the pen and paper version in a lecture.

Performance of the L1 group was very comparable to that of the initial study (M = 53.8, SD = 6.5, range: 33 to 60), indicating that performance on the list of 90 selected items did not differ depending on when these words were presented alone or in the presence of the 90 items that did not make it). The performance of the L2 group was slightly lower than that of the initial study (M = 7.2, SD = 8.9, range = -17 to 56), as could be expected given that the proficiency level was quite low. Despite the fact that they attended an introductory class of Spanish, seven of the L2 participants gave themselves a rating of 6 and three even a rating of 7. The participants with a rating of 6 had Lextale scores of 7, 22, 24, 9, 5, 20, and -3; those with a rating of 7 had Lextale scores of 7, 21, and -2. An analysis of their performance indicated that they selected more words than students who gave themselves a low rating, but at the same time

[3] The authors thank Ilse Logie for her kind cooperation.

were much more prone to false alarms to the nonwords. In other words, for a few L2 participants thinking they were good Spanish speakers, everything that looked Spanish was a “known” Spanish word.

Discussion

Lemhöfer and Broersma (2012) made a convincing case that researchers should measure the proficiency level of their participants with an objective test. Their message was primarily aimed at L2 researchers, but a similar argument can be made for L1 researchers (Andrews & Hersch, 2010; Chateau & Jared, 2000; Diependaele et al., 2013; Yap et al., 2008). To ameliorate the situation, Lemhöfer and Broersma (2012) presented an English vocabulary test, LexTALE, that allows researchers to get a reliable and valid estimate of vocabulary size in less than four minutes. Subsequent research (Diependaele et al., 2013; Khare et al., 2013) attested to the usefulness of the test.

Although a proficiency test in English is good, it would be better if equivalent tests existed for other languages as well. Lemhöfer and Broersma (2012) developed similar tests for Dutch and German, but did not test or validate them yet. Brysbaert (2013) compiled a test for French, which has good psychometric properties due to careful selection and testing of the stimulus materials.

In the present study, we present the efforts we made to compile a good Spanish test of vocabulary knowledge. Indeed, the combination of English and Spanish is one of the most frequent language pairs examined in research on bilingualism. In addition, much Spanish word recognition research takes place, which would profit from a good and easy-to-use vocabulary test. The test construction followed a paradigm very similar to that employed in the creation of the English LexTALE (and previously the EVST) and consequently has been named Lextale-Esp (Lexical Test for Advance Learners of Spanish).

Our data show that we were able to compile such a test. Two important aspects in the construction were:

1. The selection of good words (from a wide range of frequencies, going from known to nearly everyone to known only by speakers with a very high proficiency level) and the creation of good nonwords (not too easy, not too underhand).

2. Further improvement is possible by presenting the stimuli to groups of different proficiency levels. This allowed us to see which items discriminate well and which create confusion. It also allows researchers to reduce the number of stimuli, because redundancy can be pruned, or to create extra stimuli if gaps need to be filled. Given that a vocabulary test measures a single construct (number of words known), an IRT analysis is well indicated.

Although the high reliability of the test suggests that some further reduction in number of items is possible, we think this would not be a good idea. First, the test as it is now, is quite short (below 5 min administration time). So, the gain in time would be small. Second, we deliberately sought to develop a test that was not prone to floor or ceiling effects, so that it can be used for all language research. This is only possible if the test contains items of various difficulty levels, going from very easy to very difficult. Finally, when the test is used to measure individual differences in more homogeneous groups, it is important to be able to make fine-grained distinctions.

Our test shows the big difference in vocabulary size between native speakers and L2 speakers (as in Brysbaert, 2013). There is virtually no overlap in the scores of L1 and L2 speakers. To some extent this is because we did not have very proficient Spanish L2 speakers. Another factor, however, is that the vocabulary size of native speakers across all possible topics and language registers is rarely attained by L2 speakers. Indeed, someone is considered very proficient in L2 when 8,000 word families are known, whereas the total number of word families in a language is estimated to be more than 30,000 (Laufer & Ravenhorst-Kalovski, 2010).

LexTALe-Exp scores can be used to compare participants within and between studies. For the latter, it would be ideal to have more norming data. On the basis of our findings we can be quite confident that Spanish L1 psychology students will on average have scores around 54/60 (90%). Similarly, beginning L2 learners with unrelated native languages such as Dutch and English are expected to have averages of less than 12/60 (20%). It will be interesting to see how other groups are doing in

this respect. Two variables are likely to be important: Spanish proficiency and the similarity between L1 and L2. As for the latter, we made sure that none of the words used were cognates with the English language and that none of the nonwords were words in Spanish or English. However, a number of words are likely to be cognates with other languages, particularly those languages close to Spanish such as Catalan, French, Portuguese, or Italian. Although it may be tempting to try to avoid the words that may be problematic in the various languages, one must be careful not to construct a test that is too artificial. If two languages have the same root, they are likely to share many everyday words and people who know one language find it easier to learn the other. Taking out these everyday words risks to harm the validity of the test.

Because we do not know how participants with various L1s will perform on the test, it is advised to collect some extra norms if the scores of the Lextale-Esp test are to be used. We are confident that the test, as presented here, is suitable for English-Spanish and Spanish-English bilinguals (in addition to Dutch/Spanish bilinguals), because there is as little overlap between English and Spanish as between Dutch and Spanish. Some caution may be warranted, however, when one wants to interpret the absolute scores of bilinguals with other language combinations.

A different but related issue is the potential influence of the other language on the nonword decisions. This issue was discussed by Meara (1990) in an overview report of the EVST. He argued that a nonword in English like LOYALMENT could cause particular problems to speakers of Spanish as L1. He proposed two reasons for this difficulty: First the fact that LOYAL is a cognate of LEAL in Spanish, and second the

fact that Spanish adverbs are formed by adding the suffix MENTE. Therefore the existing Spanish word LEALMENTE (meaning *loyally*) may induce Spanish speakers to accept LOYALMENT as an English word. Meara (1990) observed, however, that although EVST had different problems for participants with different L1 backgrounds, the overall scores did not seem to differ much. Further research with beginning learners of Spanish in various countries and regions will have to indicate whether the same is true for Lextale-Esp.

An objective proficiency test is better than subjective ratings, because it is less susceptible to response biases (at least when constructed properly). Response biases are particularly a problem when participants are motivated to take part in the study (e.g., because they are paid) or when they want to impress the experimenter. In addition, subjective self-assessments suffer from another problem, as illustrated in Figure 2. Whereas very low ratings are an indication of low proficiency and very high ratings an indication of high proficiency, in-between there is a band of ratings that give rise to quite different levels of performance. Partly, this has to do with response biases in individual participants (too modest or too daring). However, in our experience it also has to do with the fact that raters rarely take into consideration the complete range of proficiency. Beginning L2 learners sometimes give themselves a 6 or 7, because they have the impression they are doing well relative to the other members of their (L2) group. For the same reason, native speakers sometimes give themselves a rating of 6 – 7, because they perceive themselves as performing less well than other proficient L1 speakers. This makes that non-extreme ratings are a mix of

different perceptions about what language proficiency entails. Such is not the case for objects scores such as those of LexTALe.

Availability

The test is very easy to implement in whatever software one wants to use to present stimuli and collect responses (on a desktop, on the internet, on smartphones or tablets, on paper, etc.). The sequence of stimuli we used is the following (words are translated in English; nonwords are indicated as NW):

terzo (NW), pellizcar (pinch), pulmones (lungs), batillón (NW), zapato (shoe), tergiversar (distort), pésimo (abysmal), cadeña (NW), hacha (axe), antar (NW), cenefa (edging), asesinato (murder), helar (freeze), yunque (anvil), regar (water), abracer (NW), floroso (NW), arsa (NW), brecedad (NW), ávido (avid), capillo (NW), lacayo (lackey), lampera (NW), látigo (whip), bisagra (hinge), secuestro (kidnapping), acutación (NW), merodear (prowl), decar (NW), alardio (NW), pandilla (gang), fatacidad (NW), pauca (NW), aviso (notice), rompido (NW), loro (parrot), granuja (rascal), estornudar (sneeze), torpe (clumsy), alfombra (carpet), rebuscar (rummage), cadallo (NW), canela (cinnamon), cuchara (spoon), jilguero (goldfinch), martillo (hammer), cartinar (NW), ladrón (thief), ganar (win), flamida (NW), candado (padlock), camisa (shirt), vegada (NW), fomentar (promote), nevar (snow), musgo (moss), tacaño (stingy), plaudir (NW), besar (kiss), matar (kill), seda (silk), flaco (skinny), esposante (NW), orgulloso (proud), bizcocho (cake), hacido (NW), cabello (hair), alegre (cheerful), engatusar (cajole), temblo (NW), polvoriento (dusty), pemición (NW), hervidor (kettle), cintro (NW), yacer (lie), atar (tie), tiburón (shark), frondoso (leafy), tropaje (NW), hormiga (ant), pozo (well), empirador (NW), guante

(glove), escuto (NW), laud (lute), barato (cheap), grodo (NW), acantilado (cliff), prisa (hurry), clavel (carnation).

In addition, we provide paper versions of LEXTALE_ESP in the supplementary materials, both with instructions in Spanish and in English. In this way the test is easy be use and understand.

References

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbour priming. *Journal Of Experimental Psychology: General, 139,* 299-318.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). *The CELEX Lexical Database Release 2* [CD-ROM]. University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Bialystok, e., Craik, F. I., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences, 16* (4), 240-250.

Brysbaert, M. (2013). LEXTALE_FR: A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica, 53,* 23-37.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research, 6,* 145–173.

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory and Cognition, 28,* 143-153.

Colzato, L.S., Bajo, M.T., van den Wildenberg, W., Paolieri, D., Nieuwenhuis, S., La Heij, W., & Hommel, B. (2008). How does bilingualism improve executive control? A comparison of active and reactive inhibition mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 302–312.

Cuetos, F., González-Nosti, M., Barbón, A., & Brysbaert, M. (2011). Spanish word frequency based on film subtitles. *Psicológica 32,* 133-143.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The Word Frequency Effect in First and Second Language Word Recognition: A Lexical Quality Account. *Quarterly Journal of Experimental Psychology*. In press.

Duchon , A., Perea, M., Sebastián-Gallés, N., Martí, A., Carreiras, M. (in press). EsPal: One-stop Shopping for Spanish Word Properties. Behavior Research Methods.

Dunn, A., & Fox Tree, J. E. (2009). A quick, gradient Bilingual Dominance Scale. *Bilingualism: Language and Cognition, 12* (3), 273-289.

González-Nosti, M., Barbón, A., Rodríguez-Ferreiro, J. & Cuetos, F. (under revision). Effects of the psycholinguistic variables in the lexical decision task in Spanish: A study with 2765 words. *Brain Research Methods.*

Grainger, J., Dufau, S., Montant, M., Ziegler, J.C., & Fagot, J. (2012). Orthographic processing in baboons, *Science, 336,* 245-248.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42,* 627-633.

Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon, 6,* 34-52.

Khare, V., Verma, A., Kar, B., Srinivasan N., & Brysbaert, M. (2013). Bilingualism and the increased attentional blink effect: Evidence that the difference between bilinguals and monolinguals generalizes to different levels of second language proficiency. *Psychological Research*. In press.

Laufer, B., & Ravenhorst- Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22,* 15-30

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behaviour Research Methods, 44* (2), 325-343.

Li, P., Sepanski, S., & Zhao, X. (2006). Language history questionnaires: A web-based interface for bilingual research. *Behavioral Research Methods, 38*, 202-210.

Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. In J. Tommola (Ed.), *Foreign language comprehension and production* (pp. 103-113). Turko: AfinLA.

Meara, P., & Jones, G. (1987). Test of vocabulary size in English as a foreign language. *Polyglot, 8* (1), 1-40.

Meara, P., & Jones, G. (1990). *Eurocentres Vocabulary Size Test* (version E1. 1/K10,MSDOS). Zurich: Eurocentres Learning Service.

Mochida, A., & Harrington, M. W. (2006). The Yes-No test as a measure of receptive vocabulary knowledge. *Language Testing, 23* (1), 73-98.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* Rowley, MA: Newbury House.

Rizopoulos, D. (2006). Ltm: An R package of latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17* (5), 1-25.

Schmitt, N. (2000). *Vocabulary in language teaching.* (J.C. Richards, Ed.). New York: Cambridge University Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

Varma, S. (n.d.). Preliminary item statistics using point-biserial correlation and *p* – values. *Google.*com. Retrieved April 2013 from http://www.google.com

Yap, M. J., Balota, D. A., Tse, C-S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for

opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, & Cognition 34,* 495-513.