LexITA: A quick and reliable assessment tool for Italian L2 receptive vocabulary size

Simona Amenta [1,2] Linda Badan [3] & Marc Brysbaert [2]

[1] Centre for Mind/Brain Sciences, University of Trento

[2] Department of Experimental Psychology, Ghent Univers

[3] Department of Translation, Interpreting and Communication, Ghent University

To be published in Applied Linguistics.

E-mail: linda.badan@ugent.be

Abstract

In language and second language acquisition research, it is important to have a measure for tracking the proficiency level of participants. Lexical competence is fundamental for communicative purposes in a given language, and vocabulary tests are a reliable measure to assess lexical proficiency. That is why vocabulary tests have a central role in language proficiency assessment. Although many people study Italian as second language (L2), an easy-to-use vocabulary test to measure lexical proficiency is still missing. In this work, we aim to fill this gap by presenting LexITA, which is an objective, reliable, and quick assessment of Italian receptive vocabulary. LextITA was validated on students of Italian L2 and 20 showed to be a valid measure to assess vocabulary knowledge of L2 speakers spanning different levels of proficiency.

Introduction

People differ in language proficiency. This is particularly true for non-native speakers acquiring a second language (L2), but it is also true for native speakers, who for instance show substantial differences in the number of words they recognize (Author 2016). In recent years, empirical language researchers have become more aware of these individual differences. As a result, it is no longer accepted to present a study's results without proper assessment of the participants' language proficiency, which goes beyond subjective self-rating. Researchers have focused on vocabulary size because it is a good indicator of language proficiency and can be assessed efficiently (see for instance Ouellette 2006; Author *et alii* 2017; Anderson and Freebody 1983; Read 2000). An important requirement of a vocabulary size test for research is that it measures lexical knowledge objectively, reliably and rapidly, so that it can easily be integrated in a study.

In the present article, we describe the efforts we made to develop an Italian receptive vocabulary test, so that vocabulary size can be estimated objectively and efficiently in empirical studies of Italian, filling a gap in the field of linguistic and psycholinguistic research on Italian as L2.

Assessing language proficiency and vocabulary in research context

The easiest way to measure language proficiency is to ask people how good they are. Often these subjective assessments are accompanied by a language background questionnaire (e.g. Li *et alii* 2014). Although self-ratings provide useful information (LeBlanc and Painchaud 1985), they are limited in a number of ways (Marian *et alii* 2007). For a start, estimating one's own level is rather difficult and usually provides a crude measure. Second, people may use different comparison groups. Native (L1) speakers tend to compare themselves with other native speakers, whereas L2 speakers often compare themselves to other L2 speakers (Chan and Chang 2018). Third, subjective proficiency ratings may have different meanings in different cultures, making it difficult to compare

international findings in a systematic way. Finally, subjective assessments may be influenced by language anxiety (MacIntyre *et alii* 1997) or self-serving biases.

Because of the problems with subjective assessment, researchers need to complement such assessments with more objective information. One possibility is to use a commercial test. There are, however, two problems with such tests. The first is that they are copyright protected and must be purchased. This makes it difficult to use the same test throughout an entire research area, so that the various studies can be compared to each other. A second limitation is that many of the commercial tests are quite demanding in terms of time and resources needed for their administration and provide more information than is needed. As such, they are difficult to integrate in many research contexts, where participants are tested for circa one hour. For these reasons, language researchers have developed new resources, specifically designed for language research, so that (i) the test is freely available to all researchers, and (ii) it does not introduce a burden that is too high for inclusion in a typical one-hour language experiment.

Vocabulary tests have a prominent place in language proficiency assessment, because lexical competence is central to communicative competence (Meara 1996) and vocabulary tests are a reliable predictor of lexical competence (e.g., Ouellette 2006 for L1; Author *et alii* 2017; for L2, Anderson and Freebody 1983; Mezynski 1983; Read 2000). For this reason, in order to assess language proficiency of participants, language researchers mostly rely on a vocabulary test.

There are several ways to test the multidimensional nature of vocabulary knowledge. In fact, researchers diverge in defining what vocabulary or lexical knowledge is. It is often defined as a bulk of interconnected sub-knowledges, such as knowledge of morphological forms and meanings, grammatical knowledge and social constraints in the use of words (Nation 1990, 2001; Ringbom 1987). Lexical knowledge has also been defined as a *continuum* consisting of different levels of knowledge of a word, from a superficial to a deep level (Faerch *et alii* 1984, Palmberg 1987, Henrikseen 1999). A distinction that is often made is between the dimensions of *breadth* and *depth* of lexical knowledge (Anderson and Freebody 1981, Read 1993, Qian 2002, Wesche and Paribakht

1996, Gyllstad 2013). *Depth* of knowledge refers to the quality of lexical knowledge, which reveals how well words are known with respect to their various senses and uses in contexts. *Breadth* refers to the size of vocabulary knowledge and simply refers to the number of words that a person knows (Read 2000).

The majority of vocabulary tests assess *breadth* of knowledge rather than *depth* (for an overview of different tests developed for English as L2, see Schmitt 2000). However, several studies have shown a high correlation between both concepts (see Gyllstad 2007 for discussion).

The two best known tests for English in L2 research are the Vocabulary Levels Test (Nation 1990) and the Vocabulary Size Test (Meara and Jones 1987, 1990). The first test presents words of different frequency levels to the participants and requests them to match the word to definitions or synonyms. In contrast, in the Vocabulary Size Test, a version of which is described in this paper, participants have to decide whether strings of letters represent words they recognize or not. To correct yes-responses to unknown words, the test includes non-words, *i.*e. made-up sequences of letters that resemble words but that do not exist in the language. The score is obtained by subtracting the percentage of yes-responses to the non-words from the percentage of yes-responses to the existing words. The Yes/No format adopted by the Vocabulary Size Test is based on a model originally used in L1 research (Zimmerman *et alii* 1977) and updated with pseudo-words by Anderson and Freebody (1983). This format was promoted for L2 use by Meara and Buxton (1987), who argued that it was the easiest and quickest way to get a useful measure of language proficiency.

It has been argued that this test format may have some limitations that, however, have been showed not to hamper its validity. First, it has been argued that the Yes/No format is a mono-dimensional format that measures superficial knowledge of words. It does not offer information of what (and how much) a learner knows about the words. However, Yes/No test results correlate well with more demanding proficiency tasks (Harrington and Carey 2009; Zhang *et alii* 2019). Furthermore, it has been shown that vocabulary size is not only a relevant factor to obtain fluency in

speech (Coady *et alii* 1993), but it is also a valid predictor for reading comprehension (Anderson and Freebody 1981) in both L1 and L2 (Nation 1990, 1993, 2001).

Secondly, a Yes/No size test is a *receptive* vocabulary test in which a learner needs to recognize an item as a word in a target language. The test does not measure whether the learner is able to use the word or not. However, we agree with Meara (1996) who argued that the vocabulary size measured with a Yes/No test does more than just offering a rough measure of the number of words that a learner can recognize. L2 learners (as well as L1 speakers), during the acquisition of a word from exposure to a language, will inevitably learn more than just to recognize its form (Nation 1990). Moreover, the importance of receptive recognition is underlined by the fact that "it is generally assumed that words are known receptively first and only later become available for productive use, which is why it is most useful to think in terms of a receptive to productive continuum, representing increasing degrees of knowledge of a word." (Eyckmans 2004:15).

A third limitation of the test format is that the words are displayed in isolation, totally *de-contextualized*. The Yes/No test adopts the discrete-knowledge approach (defined also as *trait view*, see Chapelle 1998), which sees vocabulary knowledge as a construct distinct and separated from other components of language. Despite the limits of the context-independent nature of the Yes/No test, researchers such as Cameron (2002) and Stanovich (1980) point out that vocabulary size highly correlates with reading comprehension. Also Read (2000) notices that the more the assessment of words are contextualized, the less clear is to what extent is the learner's performance is influenced by their word knowledge.

LexTALE

The format of the Vocabulary Size Test got a strong impetus in L2 research when Lemhöfer and Broersma (2012) developed a version for high-proficiency English L2 speakers.

Lemhöfer and Broersma (2012) called their test LexTALE, a Lexical Test for Advanced Learners of English. It is a receptive Yes/No vocabulary size test, targeting the breadth of vocabulary. Lemhöfer and Broersma (2012) opted for this format, because it is the least demanding and still provides a useful measure of language proficiency (e.g., Beeckmans *et alii* 2001, Eyckmans 2004, Huibregtse *et alii* 2002). For instance, Harrington and Carey (2009) reported a highly consistent correlation between Yes/No vocabulary scores and placement decisions based on listening, grammar and writing tests at an Australian university. Similarly, Zhang *et alii* (2019) reported a correlation of .76 between the scores on the Yes/No task and whether or not participants were able to translate the words of the test.

The LexTALE has 60 items, subdivided in 40 words and 20 non-words. Words vary in difficulty, operationalized by the frequency with which they occur in the language. Words were selected in such a way that the low frequency words should be known only to participants with high proficiency levels, whereas high frequency words should be known to virtually all participants. Non-words were added to discourage participants from saying "yes" to words they did not recognize.

Lemhöfer and Broersma (2012) used the test to assess the English vocabulary knowledge of 72 Dutch and 87 Korean learners of English L2 at rather high proficiency levels. They reported a correlation of .75 between the LexTALE scores and performance on a translation task in the Dutch native speakers, but a lower correlation of .51 for the Korean native speakers. The test further gained in importance when it was reported that differences in vocabulary size, as measured with LexTALE, predicted differences in word processing efficiency. Participants with large vocabularies recognized words faster and were less influenced by the frequencies of the words. This was true both in L2 and in L1 (Diependaele *et alii* 2013).

In addition, LexTALE has other practical advantages: (i) It is fast, taking only three to five minutes to complete; (ii) it is freely available for research purposes; (iii) it is easy 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 non-words selected; (iv) it can be administered in a paper-and-pen version as well as on a computer.

As a result, the LexTALE has become a standard test in English L2 research. It easily allows comparing participants groups in various studies. In addition, the scores can be used for research about individual differences in language processing (as done by Diependaele *et alii* 2013).

Lemhöfer and Broersma (2012) also developed Dutch and German versions of LexTALE. They tried to make the three tests equally difficult for native speakers, so that test scores could be compared not only within a language but also between languages (which is important for research on bilinguals).

Because of the advantages of the LexTALE tests, similar tests were developed for French, Spanish, and Chinese. A difference with the Lemhöfer and Broersma (2012) tests, however, is that no attempt was made to equate the difficulty across languages. Instead, it was found more important to try and develop the best possible test for each language separately. As a result, the new tests were no longer named LexTALE (with TALE in capital letters as part of an acronym) but Lextale-Fr (Author 2013), Lextale-Esp (Izura *et alii* 2014) and LEXTALE_CH (Chan and Chang 2018). The format remained the same, but the number of items and the difficulty of the items differed, in order to increase the reliability of the test and to make the test wide enough so that it could be used both for L2 and L1 speakers (e.g., Ferré and Author 2017). In order to do so, Lextale-Fr includes 84 items (56 words and 28 non-words); Lextale-Esp and LEXTALE_CH each have 90 items (60 words and 30 non-words). All tests were presented to groups of L2 and L1 speakers. The initial versions of the tests included more stimuli, which were subsequently pruned on the basis of an item analysis (see below).

Measuring lexical knowledge in Italian

In the present article we propose a new test of lexical knowledge for Italian, drawing inspiration from the Vocabulary Size Test and the LexTALE. However, because our test differs from the

original LexTALE tests, we propose to call it the LexITA test.[1] Our new test is addressed for use in empirical research with learners of Italian as L2, and it encompasses all proficiency levels from low to high, rather than being limited to high-proficiency speakers only.

Despite the relevance of the Italian language in empirical research, a freely available, short, reliable vocabulary test, aimed at adults that approach Italian as a second language is still missing.

Examinations to obtain language certification for Italian L2, like the Certificate of Italian as Foreign Language (CILS) or the Certificate of the Italian Language (CELI), just to mention two of the seven available certifications, do not include a specific receptive vocabulary test that can be easily adapted to a research context. Psycholinguistic and linguistic studies on Italian (both L1 and L2) currently rely on the Peabody Receptive Vocabulary Test (Stella *et alii* 2000) if they need an objective assessment of participants' lexical knowledge. However, the test is copyright protected and has not yet been tested and normed for adults (being a test for children primarily). In the absence of a standard test to measure receptive vocabulary for adult learners of Italian L2, researchers rely on subjective estimates or ad-hoc lexical decision tasks (e.g., Primativo *et alii* 2013). Although the latter method could give a good estimate of vocabulary (see Meara, 1996; Meara and Buxton, 1987), a problem with ad hoc tests is that they vary from study to study, hence limiting comparability of proficiency levels of participants across studies.

With the present work, we intend to create and test a reliable measure of lexical knowledge for Italian, which will provide Italian L2 researchers with a resource to assess participant's vocabulary objectively.

Methods

The procedure for the creation of LextITA followed closely the procedure successfully adopted for the development of the Lextale tests, and, in particular, the French the Spanish, and the Chinese versions.

Materials

Ninety words were extracted from Subtlex-IT (Crepaldi *et alii* 2012). Subtlex-IT is a corpus of word frequencies based on television (tv shows and movies) subtitles. The use of subtitle frequencies to study written words recognition is widely used in psycholinguistic studies in many languages (e.g, Keuleers *et alii* 2010; Cuetos *et alii* 2012; Van Heuven *et alii* 2014; Mandera *et alii* 2015), and this type of frequency are particularly reliable in predicting visual word recognition (Author and New, 2009, Author *et alii* 2011; Author et alii 2018; Heister and Klieg 2012; Soares *et alii* 2015; Herdağdelen and Marelli 2017).

Item words were selected in order to cover a wide frequency range: from very high frequency words, like *viaggiare* 'to travel' or *spiaggia* 'beach', that are likely to be recognized to most speakers, to very low frequency words, like *fregio* 'frieze' and *liuto* 'lute', that are likely to be recognized only to proficient L1 speakers. Overall, we included 28 words with less than 1 occurrence per million words (pm), 23 that ranged between 1 and 5 occurrences pm, 14 raging between 6 and 10 occurrences pm, 16 between 11 and 20 occurrences pm, 7 between 21 and 100 occurrences pm, and 2 higher than 100 occurrences pm. Frequency and length distribution for the word items is shown in Table 1. Of the 90 word items, 52 were nouns, 26 were verbs, and 12 were adjectives.

<Insert Table 1 about here>

Together with the word items, we compiled a list of 90 non-word items, created with Wuggy (Keuleers and Author 2010), an algorithm that generates non-words resembling the language in its

ortho-phonotactic structure. The advantage of using this system is that the non-words are legal items in the language and have the right degree of difficulty. If the non-words are too easy, participants can pick out the words without knowing them (that is, without processing them lexically; Keuleers and Author 2011). If the non-words are too difficult, they may create confusion and are sometimes more likely to be accepted by L1 than by L2 participants. This is particularly the case for non-words that sound like existing words (e.g., Author 2013). With such non-words, the test risks no longer to be a vocabulary test, but a spelling test. To make sure that we had the right type of non-words, they were tested in a regular lexical decision experiment (with time pressure) with Italian L1 speakers (Crepaldi *et alii* 2015). We selected only the non-words with approximately 90% correct rejections in the speeded lexical decision task.

Procedure

Word and non-word items were randomly included in a list using Google Forms. This system allows collecting data online through a sharable weblink. We administered the questionnaire to university students (see next section). To ensure the validity of the data, the link to the questionnaire was shared by professors during classes, and participants could fill in the questionnaire during the class time or later. The colleagues who shared the questionnaire made clear to the students that the test was to be filled in individually, without the help of other students or of online or offline instruments. We also made clear that each test was completely anonymous, and that it was not part of the course evaluation. We believe that by sharing the link during class time, all participants were motivated to fill in the questionnaire in a serious and honest way without external help.

We divided the questionnaire in two parts. The first part included instructions, privacy information and consent. To make sure that this part could be understood by beginner learners of Italian, we provided an English translation after the Italian instructions (see Appendix). This part was followed by some questions regarding the participants' age, gender, education level, L1, number of (eventual) other languages spoken, level of Italian (from A1 to C2), subjective estimate

of ability to read, write, understand (orally), and speak Italian (scale 1 to 5). The second part of the questionnaire contained the 180 initial items of LexITA. The system generated a novel random permutation for each participant, so that we were able to exclude effects due to list composition. Participants were presented with the question "Is this an Italian word?", and had to answer yes or no to all 180 items.

Participants

We presented our list of items to two groups: one composed of proficient, native speakers of Italian (L1 group), and one composed of learners of Italian as a second language (L2 group). The L1 group consisted of 58 university students (20 males, 37 females and 1 n.d.), recruited at the University of Milano-Bicocca (Milan, Italy), or Erasmus students at X University. Most of them were students of Communication. Their mean age was 24 (SD=7.8; range= 18-57; mode= 19). Of these 58 subjects, 55 also spoke one or more languages besides Italian, one did not speak any language other than Italian, and two gave no answer to this question.

The L2 group consisted of 141 participants (26 male, 111 female, 3 other, 1 n.d.) recruited at X University and at the University of Leiden (Netherlands). Their mean age was 25 (SD=11.6; range=17-72; mode=18). The L1 of the participants is listed in Table 2. Participants from this group were students of Communication or Translation. Since they were all enrolled in formal courses of Italian as L2, we could also collect information about their proficiency level, objectively assessed, according to the Common European Framework of Reference for Languages (CEFR)[3]. Following this classification, participants were divided into proficiency levels as follows: 46 were beginners (31 at A1 level, 15 at A2 level), 60 were intermediate (26 at B1 level, 34 at B2 level), and 35 were advanced learners (27 at C1 level, 8 at C2 level).

<Insert Table 2 about here>

Results

The first aim of our analyses was to select items with a good discriminative power that could be included in the final version of LexITA, which we wanted to have 60 word items and 30 non-word items (as in the French, Spanish, and Chinese Lextale tests). To achieve this aim, we analysed the responses to the items using point-biserial correlation and Item Response Theory. Once we identified the items that discriminated between participants, we analysed the responses as a function of proficiency levels and we calculated the reliability of the test by computing Cronbach alpha and intra-class correlation.

Selecting items for LexITA

To assess the quality of the items, we analysed words and non-words separately, first running a point-biserial correlation between the response to each item and the participants' total accuracy. For each subject we computed their mean accuracy to word items and non-word items separately. Then we computed two correlations: one between the accuracy to the single word item and the mean word item accuracy, and one between the single non-word item and the mean non-word item accuracy. For the purpose of these analyses, both groups were included (L1 and L2). Point-biserial correlation coefficient ranges between -1 and + 1. For a good word item, the coefficient is expected to be positive: Participants who indicate they recognized the item tend to have a higher overall score than participants who indicate they do not recognize the item. A negative correlation indicates that someone who indicates they recognize the item, overall has a lower score than someone who indicates they do not recognize the item. Such correlation indicates that there is something strange about the word.

All but one word item showed a positive correlation (from .02 to .65). The exception was the word *economico* "economic", which showed a small negative correlation and was removed from further analyses. We may speculate why the word *economico* was more likely to be selected as a

word by participants with low overall performance. One reason might be that the word is a cognate of the words “economisch” in Dutch and “economic” in English. Dutch and English speakers constituted the majority of our L2 group.

All non-word items showed a positive correlation (from .19 to .74).

Point-biserial correlations give important information about the usefulness of an item, but this is not sufficient. Following Author (2013), what we want in a test are items that span across the whole difficulty range, and that have good discriminative power. In other words, we want to include not only easy and difficult items, but also items in-between to make fine distinctions at intermediate proficiency levels. To this aim we applied an Item Response Theory (IRT) analysis to our items. The IRT analysis takes into consideration both the performance level of the participants and the difficulty level of the items. To perform this analysis, we used the *ltm* R package (Rizopoulos 2006), and ran the two-parameter logistic model on word items and non-word items separately. This model is based on two parameters: Difficulty and Discriminative power.

In this model, difficulty is represented by the location on the x-axis (i.e., how far to the right or to the left the curve is displaced). This difficulty parameter is operationalised as the point on the x-axis where the item response curve crosses the 0.5 probability value on the y-axis. Ideally, we want to select items that cover the full range of difficulty. The item discrimination value is represented by the steepness of the response curve in its middle section. A steeper curve is index of better discrimination power of the item. Ideally, we want to select items with steep curves. For example, looking at Figure 1, we can see that *liuto* “lute” is more difficult di *frondoso* “leafy” and that they are both more difficult than *stretto* “narrow” and *cannella* “cinnamon”. We can also see that *stretto* and *cannella* have the same degree of difficulty, however *stretto* has higher discriminative power than *cannella*.

<Insert Figure 1 about here>

Following Izura *et alii* (2014), we ordered the word items following their difficulty parameter, and then selected them on the basis of their discrimination power, by extracting those with the higher value at intervals of roughly 1/30th of the difficulty range. This procedure guarantees the selection of the most discriminative items – the ones that better allow us to separate proficient respondents from less proficient respondents - for a wide range of difficulty levels. Looking at Figure 1a, it can be seen that for an ideal test we lacked some really difficult words, which would allow us to make a distinction at the very high end of the proficiency level. Few words made a distinction at the high end of the x-axis and these words did not have strong discrimination power.

We repeated the same procedure for the non-word items. All in all, we selected 60 word items and 30 non-word items to be included in the test (frequency and length of the selected items are reported in Table 3).

<Insert Table 3 about here>

Scoring of LexITA and comparison between proficiency levels

The items selected through the previous analyses, and included in the final version of LexITA, cover a wide range of difficulty and have good discriminative power. As a final step in our validation effort, we assessed how LexITA is able to differentiate between different proficiency groups.

Following Author (2013), we computed the test score as:

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

This formula makes sure that guessing behaviour is penalized (Izura et *alii* 2014). A maximum score of 60 can only be obtained by saying yes to all the words and to none of the non-words.

We then assessed how participants with different proficiency levels performed on the test. Figure 3 and Table 4 show the distribution of scores according to the CEFR levels.

<Insert Figure 2 about here>

<Insert Table 4 about here>

The figure and table show that the performance of participants was coherent with their proficiency levels. Unfortunately, we do not have comparison data for L2 participants from other tests of the same type, as, for example, none of the Lextale tests or even the original LexTALE had considered the levels of proficiency according to the CEFR. For the L1 group we see a ceiling effect, with next to perfect responses and a small standard deviation. This confirms the impression conveyed by Figure 1a that our test is well placed for L2 speakers but should have included more difficult words for native speakers. At the same time, it should be remembered that our L1 group was a quite homogeneous high-performing group, since it was composed of university students of Communication.

Table 5 shows the correlations between the scores on the test and the self-assessed abilities to read, listen, write and speak (scale 1 to 5). Correlations are high, certainly when both groups are taken together. Within each group, the correlations are a bit lower due to restricted range within each group. This is particularly true for the L1 group with its ceiling level.

<Insert Table 5 about here>

Interestingly, for the L2 group the correlations seem to be higher for reading and writing. We may speculate that this result depends on the fact that LexITA is a written test. Therefore, it may be more telling of these abilities than the ability to speak or understand oral Italian. Alternatively, it may also depend on the difficulty to give a self-evaluation for oral abilities such as speaking or understanding spoken language.

Item consistency was measured with the Cronbach alpha using the *psych* R package (Revelle 2018), which gave a very high value of $\alpha = 0.98$. We also computed the intra-class correlation (ICC) coefficient for binary data with the same R package and obtained an ICC3k coefficient of 0.97 ($p<.001$). When we limit our analysis to the L2 group, which is the target group for this test, we obtain $\alpha = 0.97$ and ICC3k=0.97 ($p<.001$)[4], which are both very high scores, indicating a strong reliability of the test.

Testing the final version of LexITA

In the previous section we analysed the responses of 199 subjects to the initial version of LexITA, which comprised 90 words and 90 nonwords. Using point-biserial correlation and IRT analysis, we selected the 60 word items and 30 non-word items with the best discriminative power. When we limited the analysis to these items we observed that the LexITA scores corresponded well with the participants' proficiency levels and correlated positively with their self-ratings of their ability to read, write, speak and understand Italian.

However, we should be mindful of the fact that, although the items we selected were the ones with good discrimination power, still the participants responded to them together with other items that may have had an influence on the responses. Even though we controlled for possible effects of list composition by randomizing the items at each presentation, we cannot be 100% sure that overall the composition of the total test had no influence.

Therefore, we administered the final version of LexITA, consisting of the selected 60 words and 30 non-words, to a new group of 187 subjects following the same procedure of the previous

study (test included in Google Form and randomized item presentation). The participants group was composed of 42 L1 speakers (mean age = 39.5; sd=12.1; range= 20-59) recruited through a mailing list of European schools of Italian as L2 (mainly teachers, teaching assistants, administrative personnel). The L2 group was composed of 145 subjects (mean age = 27.5; sd=14.5; range= 15-76), also recruited through a mailing list of European schools of Italian as L2, and at the Utrecht University, X University, Leiden University, and Manchester University. Also in this case, we asked participants to report the CEFR level corresponding to the one reached in their language courses (assigned by previous *ad hoc* examinations). However, since we do not have the strict control on the data collection procedure that we had in the first study, we cannot exclude in principle that some individuals reported a different proficiency level than their assigned one. We do not see this as an actual problem for our study, on the contrary, in this last study we purposefully relaxed the control over our participants pool since we wanted to validate LexITA as a test that can be administered freely, also outside tightly controlled environments (like a laboratory or a classroom) and still be a reliable assessment instrument. Administering the test in a less controlled environment represents a tough challenge to the validity of LexITA, but for the same reason, we believe that this tougher condition could give a stronger indication that the test can convey reliable results in different contexts. The first languages of participants belonging to the L2 group are listed in Table 6, while their self-reported proficiency level is presented in Table 7.

<Insert Table 6 about here>

<Insert Table 7 about here>

Before scoring the test, we removed 7 participants from our analysis because they responded "yes" to all items (also the non-words), which we considered a strategic behavior rather than an actual response. The distribution of scores across proficiency levels was very much in line with what we observed in the previous test, as shown in Figure 4 and Table 8. Only at the two lowest

levels, A1 and A2, did we see a different pattern, arguably because the participants used other criteria than their proficiency level to decide whether they belonged to the A1 or A2 category (see the introduction). The few participants who indicated they were at the very first level, actually performed slightly better than the larger group who considered themselves at the second level. This is a pattern reminiscent of the findings related to language anxiety, the fear of using L2 in public situations (MacIntyre *et alii* 1997). A comparison between scores of the first and second phase is presented in Table 8.

<Insert Figure 3 about here>
<Insert Table 8 about here>

We also asked the participants to rate their knowledge of Italian on a scale from 1 to 9. The correlation between the LexITA scores and the self-ratings was r = .70, which is slightly lower than what we observed in the first study, but still a high correlation.

When comparing the results of the first and the second study, it is good to keep the differences in procedure in mind. The first administration of the test was much more controlled (the link to our online questionnaire was shared by teachers in class), the participant group was very homogeneous (university students of language-related degrees), and the CEFR proficiency levels were established in a more objective way (as part of the students' enrollment at university). In the second study we shared the test via a mailing list to participants who only had a common interest of learning Italian as L2, and they had to report their CEFR level themselves. Given these differences, the remarkably convergent results in both studies are reassuring that the LexITA scores are quite robust and apply to a large range of Italian L2 speakers. The results of the L1 groups are very comparable as well and once more suggest that the test is too easy for proficient L1 speakers.

Finally, we assessed the reliability of the test by computing the Cronbach alpha for this set. We used the alpha function included in the *psych* R package (Revelle 2018). This gave a value of

Cronbach α = 0.96, which is slightly lower than what we obtained in the first study, but still indicates very high reliability. We computed the intra-class correlation (ICC) coefficient for binary data with the same R package. We obtained a coefficient of ICCk3= 0.96 (p<.001). When we limit the analysis to the L2 group, we obtain a value of Cronbach α = 0.96 and ICCk3 = 0.96 (p<.001)[5], indicating that the final version of the test is very reliable for the target group.

Discussion

Although many people study Italian as L2, allowing researchers to study Italian L2 processing, an easy-to-use vocabulary test to measure lexical proficiency has been lacking. Vocabulary size is well known as predictor of reading and word recognition (see the references in the introduction), but only a small number of Italian studies measure this variable. As a result, information about the proficiency level of the participants and about individual differences in proficiency is absent in many scientific papers. This is similar to the situation in English before the publication of LexTALE by Lemhöfer and Broersma (2012).

With our work, we aim at filling this gap by providing the community of researchers of Italian L2 with a simple test that makes it possible to reliably and objectively assess Italian L2 lexical knowledge. We have created LexITA by following the procedure used by Author (2013), Izura *et alii* (2014), and Chan and Chang (2018) for the French, Spanish, and Chinese Lextale tests respectively. These three tests followed a standard procedure for the selection of items and expanded the number of words and non-words to cover a wider range of proficiency, and they have rapidly been picked up by researchers in the languages involved (e.g., Brand and Ernestus 2018; Declerck *et alii* 2018; Molinaro *et alii* 2017). Indeed, it is to be expected that language research will become hard to publish in good journals if they do not include objective information about the participants' vocabulary knowledge.

For the construction of a good test, it is important to start with more stimuli than needed, which are presented to groups of various proficiency levels (going from CEFR level A to L1), so that the best items can be retained. If the items are only presented to beginning learners, one risks selecting items that are too easy. If the administration is limited to advanced learners, one might select too many items that are too difficult for beginners. For LexITA, we started with a list of 90 words and 90 non-words, from which we selected 60 words and 30 non-words that covered a large range of item difficulties.

LexITA is a test for Italian L2 speakers as it turned out to be too easy for the sample of native speakers (96% correct in Italian L1 speakers). In this respect, it is easier than LexTALE (88% correct in English L1 speakers; Dijkgraaf *et alii* 2017; and 89% correct in Dutch L1 speakers; Vander Beken and Author 2018), Lextale-Esp (90% correct in Spanish L1 speakers; Ferré and Author 2017; Izura *et alii* 2014), Lextale-Fr (88% correct in French L1 speakers; Declerck *et alii* 2019), or LEXTALE-CH (71% correct in Mandarin L1 speakers; Chan and Chang 2018). Native Italian-speaking students can be assessed with the test, to compare their proficiency to that of the L2 speakers, but the test is not refined enough to measure individual differences within this group.

The reliability of the test is high (above .9). The validity is guaranteed by all the research in other languages showing that scores on Yes/No vocabulary tests correlate well with other measures of language proficiency (Harrington and Carrey 2009; Lemhöfer and Broersma 2012; Zhang *et alii* 2019). We also saw that L1 speakers performed much better than L2 speakers and that, for the latter, there was a clear relationship with the participant's CEFR levels: Beginners (A level) totalized the lowest scores, while advanced learners (C level) totalized the highest scores, still slightly below that of L1 speakers (Figures 3 and 4). It is also important to note that LexITA is able to discriminate between L2 speakers from the highest proficiency level (C1 and C2) and native speakers of Italian, capturing differences between these two groups.

A limitation, specific to LexITA, is that in our validation efforts we did not manage to test many participants who were native speakers of languages typologically similar to Italian, i.e.

Romance languages like Spanish, Catalan, Portuguese, Romanian, and French (we only had 12 overall). In fact, it would be interesting to collect more norms for these populations to assess to what extent the similarity between L1 and L2 influences the test scores. We have carefully chosen and tested the items covering a wide range of difficulty, and, to the best of our knowledge, tried to avoid the inclusion of obvious cognates, which would represent an advantage for some language groups over others. This makes it possible to use the test with Romance L1 speakers as well. At the same time, we do not think that a test becomes better if all possible cognates with each and every language are avoided, as this may result in a fairly unrepresentative sample of words (see Izura *et alii*, 2014, and Ferré and Author 2017, for similar concerns in Spanish).

The creation of LexITA and its similarity in structure and scale to the Spanish, French and Chinese Lextale tests represents an important step forward in the creation of a set of comparable measures for vocabulary knowledge in different languages. Developing LexITA following the same criteria adopted for the creation of the French, Spanish, and Chinese Lextale, will be beneficial to all the researchers interested in comparing progress in L2 acquisition when more than one language is involved. The homogeneity of methods to measure vocabulary size favours direct comparison across disciplines boosting the understanding of research findings.

Availability

LexITA is a reliable and quick assessment of Italian vocabulary in an easy manner. Completing the test requires circa 5 minutes, and it can be done with pen and paper or on a computer/smartphone/tablet, which makes it an easy and economic option. We administered our items in a random permutation but other tests (e.g., Lextale-Esp) found equally good results with a fixed presentation order. So, we do not have any strong recommendation toward one presentation form or the other. Of course, if the test is presented in its pen and paper version, it will not be possible to apply a random permutation at every presentation unless preparing different versions of the test beforehand.

A copy of LexITA items is included in the Appendix to this manuscript.

References

Anderson, R. C., & Freebody, P. (1981) Vocabulary Knowledge. In J.T. Guthrie (Ed.), *Comprehension and Teaching: Research Reviews* (pp.77-117). Newark, DE: International Reading Association.

Anderson, R.C. & Freebody, P. (1983) Reading comprehension and the assessment and acquisition of word knowledge. In B. Huxton (Ed.), *Advances in Reading/Language Research*. 2 (pp. 231-256). Greenwich, CT: JAI Press.

Author (2013). X

Author et alii (2016). X

Author et alii. (2011). X

Author, & New, B. (2009). X

Author, et alii (2017). X

Author, et alii (2018). X

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M. & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. *Language Testing*, 18, 235-274.

Brand, S., & Ernestus, M. (2018). Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: Evidence from native listeners and learners of French. *The Quarterly Journal of Experimental Psychology, 71(5)*, 1240-1259.

Cameron, L. (2002). Measuring vocabulary size in English an additional language. *Language Teaching Research* 6(2), 145-173.

Chan, I.L., & Chang, C.B. (2018). LEXTALE_CH: A quick, character-based proficiency test for Mandarin Chinese". In the *Proceedings of the 42nd Annual Boston University Conference on Language Development*.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research*, (pp. 32-70). Cambridge: Cambridge University Press.

Coady, J, Magoto, J., Hubbard, P., Graney, J. & Mokhtari, K. (1993). High frequency vocabulary and reading proficiency in ESL readers. In T. Huckin, M. Haynes and J. Coady (Eds.), *Second Language Reading and Vocabulary* (pp. 3-23) Norwood, NJ.: Ablex.

Crepaldi, D. et alii (2012). Subtlex-IT: http://crr.ugent.be/subtlex-it/

Crepaldi, D. et alii (2015). Quality, not quantity: Register is more important than size in corpusbased frequency estimation. In the *Proceedings of the 19th Conference of the European Society for Cognitive Psychology*, September 17-20 2015, Paphos, Cyprus.

Cuetos, F. et alii (2012). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, *33*(2).

Declerck, M., Snell, J., & Grainger, J. (2018). On the role of language membership information during word recognition in bilinguals: Evidence from flanker-language congruency effects. *Psychonomic Bulletin & Review, 25(2)*, 704-709.

Declerck, M., Koch, I., Duñabeitia, J. A., Grainger, J., & Stephan, D. N. (2019). What absent switch costs and mixing costs during bilingual language comprehension can tell us about language control. *Journal of Experimental Psychology: Human Perception and Performance*, 45(6), 771.

Diependaele, K. et alii (2013). The word frequency effect in first and second language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology, 66*, 843-863.

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917-930.

Eyckmans, J. (2004). *Measuring Receptive Vocabulary Size*. Utrecht: LOT.

Faerch, C., Haastrup, K., & Phillipson, R. (1984). *Learner language and language learning*. Clevedon: Multilingual Matters.

Ferré, P., & Author (2017). X

Gyllstad, H. (2007). *Testing English collocations: Developing tests for use with advanced Swedish learners*. PhD Thesis. Lund: Lund University.

Gyllstad, H. (2013). Looking at L2 Vocabulary Knowledge Dimensions from an Assessment Perspective – Challenges and Potential Solutions. In C. Bardel, B. Laufer, & C. Lindqvist (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, Eurosla Monographs Series, 2, (pp. 11-28). EUROSLA.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. System, 37(4), 614-626.

Heister, J., & Kliegl, R. (2012). Comparing word frequencies from different German text corpora. Lexical resources in psycholinguistic research, 3, 27-44.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, 21(2), 303–317.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. Cognitive science, 41(4), 976-995.

Huibregtse, I., Admiraal, W. and Meara, P. (2002). Scores on a yes–no vocabulary test: correction for guessing and response style. *Language Testing* 19, 227–45.

Izura, C. et alii (2014). LexTALE-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, *35*(1), 49-66.

Keuleers, E. et alii (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, *42*(3), 643-650.

Keuleers, E., & Author (2010). X

Keuleers, E., & Author (2011). X

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *Tesol Quarterly*, *19*(4), 673-687.

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

Li, P., Zhang, F., Tsai, E., & Puls, B. (2014). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition*, *17*(3), 673-680.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language learning*, *47*(2), 265-287.

Mandera, P. et alii (2015). Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, *47*(2), 471-483.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940-967.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), Performance and competence in second language acquisition (pp. 35–53). Cambridge: Cambridge University Press.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*, 142–151.

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

Meara, P., & Jones, G. (1990). Eurocentres vocabulary size test 10KA. *Zurich: Eurocentres*.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of educational research*, *53*(2), 253-279.

Molinaro, N., Giannelli, F., Caffarra, S., & Martin, C. (2017). Hierarchical levels of representation in language prediction: The influence of first language acquisition in highly proficient bilinguals. *Cognition, 164*, 61-73.

Nation, I. S. P. (1990). Teaching and Learning Vocabulary. Teaching Methods. *United States: Cengage Learning, Inc. Retrieved September*, *9*, 2017.

Nation, I.S.P. (1993). Vocabulary size, growth, and use. In Schreuder, R. & Weltens, B. (Eds.), *The Bilingual Lexicon*, (pp.115-134) Amsterdam: Benjamins.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554-566.

Palmberg, R. (1987). Patterns of vocabulary development in foreign-language learners. *Studies in Second Language Acquisition*, 9, 201-220.

Primativo, S., Rinaldi, P., O'Brien, S., Paizi, D., Arduino, L. S., & Burani, C. (2013). Bilingual vocabulary size and lexical reading in Italian. *Acta psychologica*, *144*(3), 554-562.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10, 355-371.

Read, J. (2000). *Assessing vocabulary* (pp. 1-85). Cambridge: Cambridge University Press.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.4.

Ringbom, H. (1987). *The Role of the First Language in Foreign Language Learning*. Clevedon: Multilingual Matters.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. Journal of statistical software, 17(5), 1-25.

Schmitt, N. (2000). *Vocabulary in language teaching*. Ernst Klett Sprachen.

Signorell, A., et mult. al. (2019). DescTools: Tools for descriptive statistics. R package version 0.99.27.

Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J. J., ... & Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. The Quarterly Journal of Experimental Psychology, 68(4), 680-696.

Stanovich, K. (1980). Towards an interactive-compensatory model of individual differences in the development of reading. *Reading Research Quarterly* 16, 32-71.

Stella, G., Pizzoli, C., & Tressoldi, P. E. (2000). Peabody test di vocabolario recettivo. *Torino: Omega Edizioni*.

Van Heuven, W. J. et alii (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

Vander Beken, H. & Author (2018). X

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review* 53, 13–40.

Zhang, X., Liu, J., & Ai, H. (2019). Pseudowords and guessing in the Yes/No format vocabulary test. *Language Testing* 21(1), 101-118.

Zimmerman, J., Broder, P.K., Shaughnessy, J.J. & Underwood, B.J. (1977). A recognition test of vocabulary using signal-detection measures and some correlates of word and non word recognition. *Intelligence* 1, 5-13.

Table 1. *Characteristics of the word items included in the initial test with 90 words.*

Distribution	Length (letters)	Frequency per million
min	4.0	0.02
1st quartile	7.0	0.66
median	8.0	4.50
3rd quartile	9.0	12.13
max	14.0	118.98
mean	7.8	11.13
SD	1.8	21.71

Table 2. First phase of the evaluation: *L2 group participants' first language (*Two bilingual subjects were further removed from validation analyses)*

Language	Number of participants
Dutch/Flemish	102
English	15
Spanish	5
French	4
Russian	4
German	2
Polish	1
Arabic	1
Chinese	1
Slovakian	1
Albanian	1
Romanian	1
Bulgarian	1
Bilingual (Flemish/Italian)*	2

Table 3. *Characteristics of word and non-word items included in the final version of LexITA.*

Distribution	Words		Nonwords	
	Length (letters)	Frequency per million	Length (letters)	Frequency per million
min	5	0.02	4	NA
1st quartile	7	0.75	5	NA
median	8	4.03	7	NA
3rd quartile	9	11.93	8	NA
max	14	118.98	12	NA
mean	7.9	11.85	7.1	NA
SD	1.9	23.07	2.1	NA

Table 4. *Mean performance and scores range in all proficiency levels considered in the first phase of the evaluation*

Proficiency Level	A1	A2	B1	B2	C1	C2	L1
Mean	12.97	20.27	35.96	40.32	43.78	50.50	57.93
SD	6.59	6.61	7.76	6.61	7.51	7.13	2.15
Range	2 - 32	7 - 31	10 - 51	18 - 51	22 - 59	34 - 59	50 - 60

Table 5. *Correlation between test scores and self-assessed ability to read, listen, write and speak Italian in the first phase of the evaluation*

	L1	**L2**	**ALL**
Reading	0.44	0.73	0.83
Listening	0.46	0.67	0.82
Writing	0.38	0.76	0.86
Speaking	0.42	0.70	0.83

Table 6. Second phase of the evaluation: *L2 participants' first language*

Language	Number of subjects
English	40
Dutch	30
Chinese	29
German	13
Finnish	9
French	7
Spanish	3
Japanese	2
Portuguese	2
Russian	2
Turkish	2
Estonian	1
Korean	1
Latvian	1
Polish	1
Slovenian	1
Tagalog	1

Table 7. *Participants' distribution over 6 considered proficiency levels*

L2 levels	Number of subjects
A1	11
A2	45
B1	28
B2	24
C1	21
C2	15

Table 8. *Comparison of LexITA scores between the first and the second phase of the evaluation*

	Proficiency Level	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**L1**
Phase 1	Mean (range)	13.0 (2 - 32)	20.3 (7 - 31)	36.0 (10 - 51)	40.3 (18 - 51)	43.8 (22 - 59)	50.5 (34 - 59)	57.9 (50 - 60)
Phase 2	Mean (range)	20.4 (4 - 58)	16.7 (-7 - 60)	26.0 (5 - 59)	36.9 (13 - 58)	47.5 (29 - 60)	53.1 (38 - 59)	57.9 (32 - 60)

Figure 1. Item response curve for 4 word items used as an example to show the procedure we followed to select word and non-word items to include in the test.

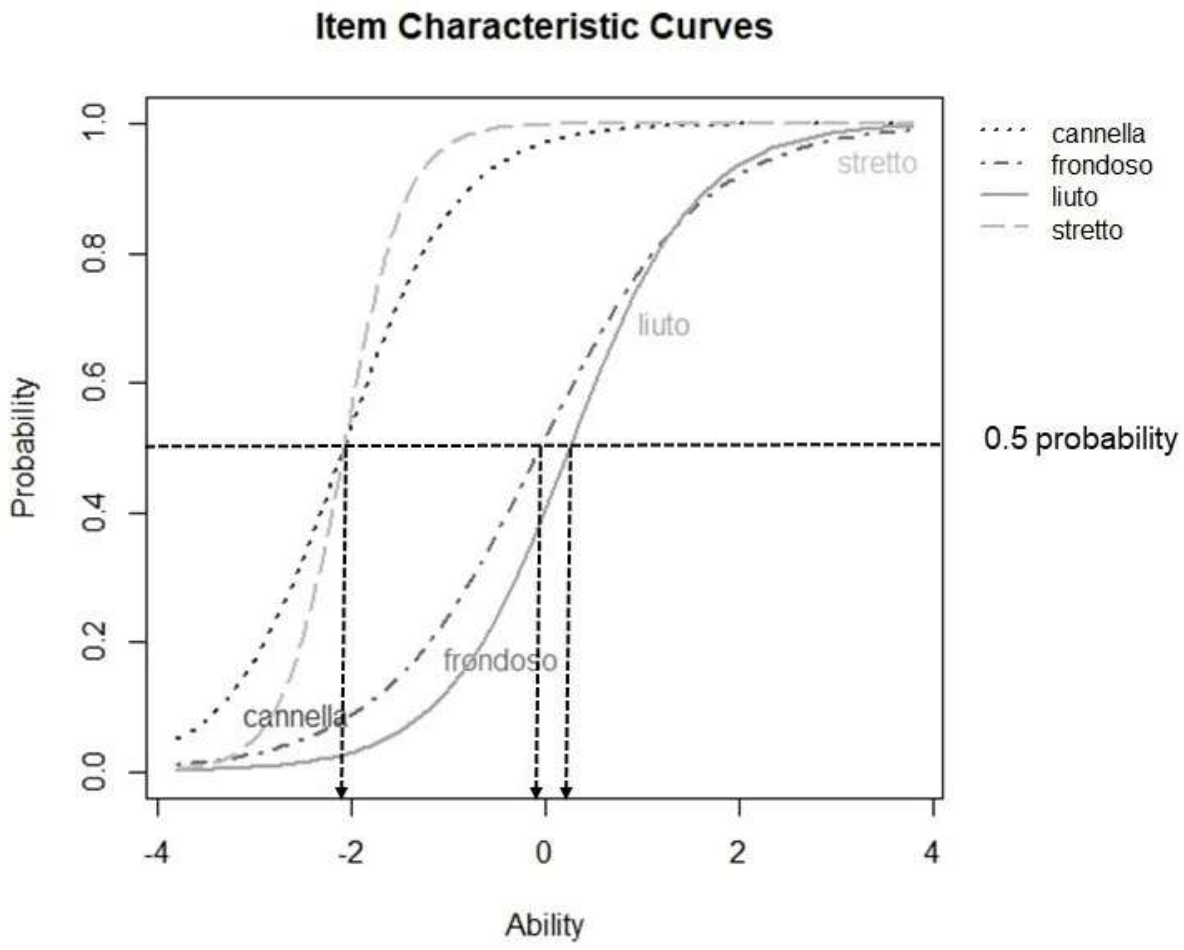


Figure 2. Distribution of scores over proficiency levels (Phase 1).

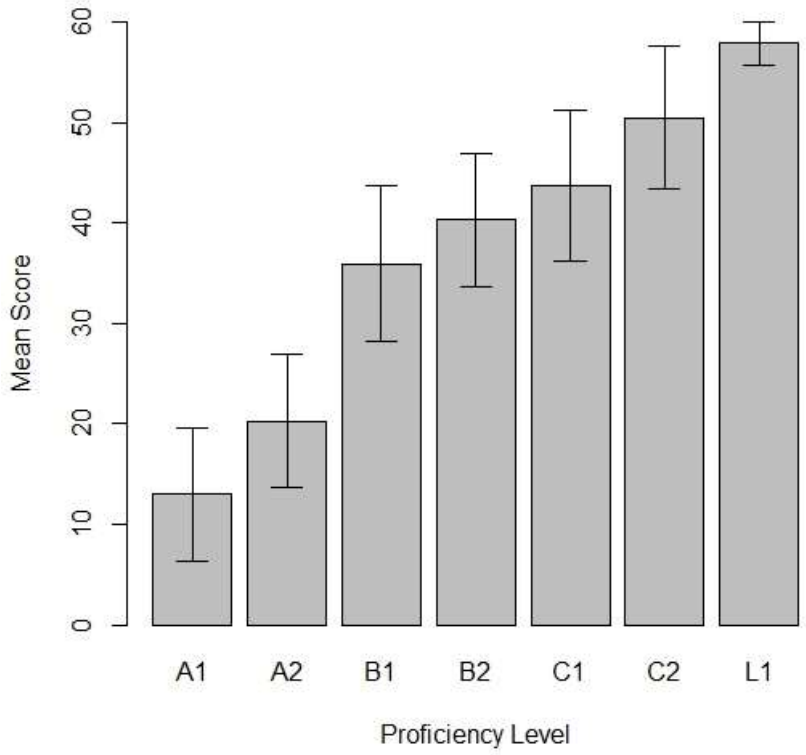
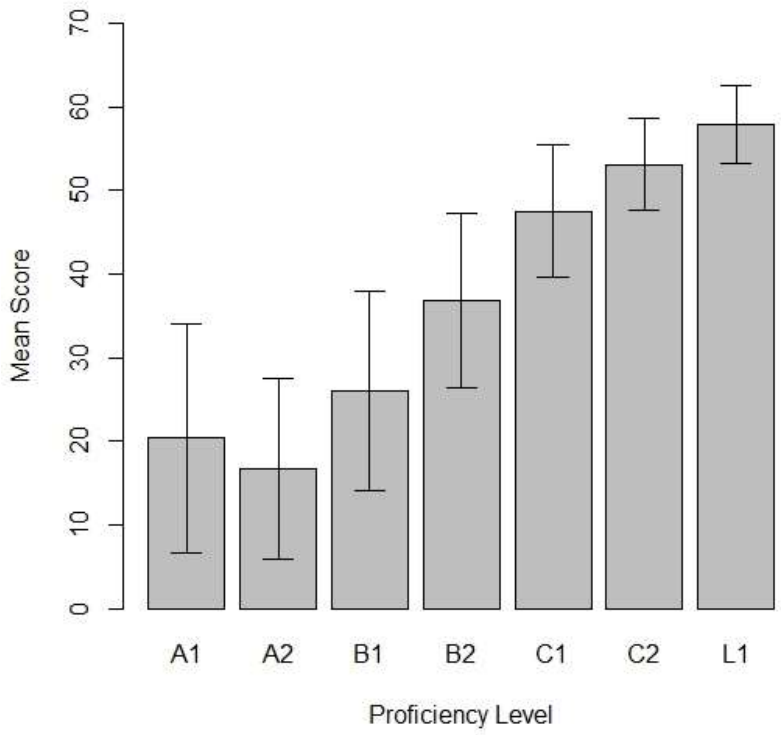


Figure 3. Distribution of scores over proficiency levels (Phase 2).



LexITA

Test del vocabolario italiano

- Il test comprende 90 sequenze di lettere che possono essere parole italiane esistenti oppure no;
- *Seleziona le parole che conosci* (o di cui sei convinto che siano parole italiane, anche se non saresti in grado di dare loro un significato preciso) mettendo una *croce accanto alla casella corrispondente*
- Attenzione: gli errori sono penalizzati! Non è vantaggioso cercare di aumentare il tuo punteggio selezionando sequenze che non hai mai visto prima!

È una parola italiana?

orgoglioso		rasso		matita		anipi	
starnutire		maledire		panagi		uscamo	
inflose		nevicare		pandispagna		cucchiaio	
predorto		avviso		lucchetto		ricercare	
giurare		balbettare		brieca		scondere	
muschio		elerro		scarpa		flasse	
bollitore		cesto		sonta		tirchio	
tarmezione		zucchero		camicia		gunto	
formica		fregio		magro		rapimento	
siscera		pessimo		ascia		annaffiare	
uccidere		ecicitizione		pencato		parda	
pizzicare		frusta		vincere		preoccupazione	
cicogna		spazzolare		prognao		becchime	
coniglio		guadagnare		congelare		fretta	
tappeto		spusa		fraintendere		polveroso	
pappagallo		racozzi		solci		scogliera	
spiaggia		ladro		troluretore		pebuito	
fannullone		incudine		squalo		dimprine	
prugna		stretto		prepimente			
furfante		fidorzato		vese			
liuto		accendere		ancegato		Tot. _____	
baciare		tacchino		gattonare			
martello		capelli		pozzo			
spegente		polmoni		allurazione			

Note per il somministratore:

Come calcolare il punteggio

Il punteggio di LexITA viene calcolato applicando la formula:

$$\text{LexITA punteggio} = N_{\text{sì a parole}} - 2 * N_{\text{sì a non-parole}}$$

Per identificare parole e non parole si può fare riferimento all'elenco sottostante.

L'elenco riporta anche le traduzioni in inglese di tutti gli item "parola".

Accendere (to turn on), allurazione (NW), ancegato (NW), anipi (NW), annaffiare (to water), ascia (axe), avviso (notice), baciare (to kiss), balbettare (to stutter), becchime (birdseed), bollitore (kettle), brieca (NW), camicia (shirt), capelli (hair), cesto (basket), cicogna (stork), congelare (to freeze), coniglio (rabbit), cucchiaio (spoon), dimprine (NW), ecicitizione (NW), elerro (NW), fannullone (idler), fidorzato (NW), flasse (NW), formica (ant), fraintendere (to misunderstand), fregio (frieze), fretta (haste), frusta (whip), furfante (scoundrel), gattonare (to crawl), giurare (to swear), guadagnare (to earn), gunto (NW), incudine (anvil), inflose (NW), ladro (thief), liuto (lute), lucchetto (padlock), magro (thin), maledire (to curse), martello (hammer), matita (pencil), muschio (moss), nevicare (to snow), orgoglioso (proud), panagi (NW), pandispagna (sponge cake), pappagallo (parrot), parda (NW), pebuito (NW), pencato (NW), pessimo (very bad), pizzicare (to pinch), polmoni (lungs), polveroso (dusty), pozzo (well), predorto (NW), preoccupazione (concern), prepimente (NW), prognao (NW), prugna (plum), racozzi (NW), rapimento (kidnapping), rasso (NW), ricercare (to research), scarpa (shoe), scogliera (cliff), scondere (NW), siscera (NW), solci (NW), sonta (NW), spazzolare (to brush), spegente (NW), spiaggia (beach), spusa (NW), squalo (shark), starnutire (to sneeze), stretto (narrow), tacchino (turkey), tappeto (carpet), tarmezione (NW), tirchio (stingy), troluretore (NW), uccidere (to kill), uscamo (NW), vese (NW), vincere (to win), zucchero (sugar).

LexITA

Italian Vocabulary Test

- The test includes 90 sequences of letters which may be existing Italian words or not;
- *Select the words you know* (even if you would not be able recall their exact meaning) *by ticking the corresponding box*;
- Be careful: errors are penalized! Trying to increase your score by selecting sequences you've never seen before can worsen your final score!

Is it an Italian word?

orgoglioso		rasso		matita		anipi	
starnutire		maledire		panagi		uscamo	
inflose		nevicare		pandispagna		cucchiaio	
predorto		avviso		lucchetto		ricercare	
giurare		balbettare		brieca		scondere	
muschio		elerro		scarpa		flasse	
bollitore		cesto		sonta		tirchio	
tarmezione		zucchero		camicia		gunto	
formica		fregio		magro		rapimento	
siscera		pessimo		ascia		annaffiare	
uccidere		ecicitizione		pencato		parda	
pizzicare		frusta		vincere		preoccupazione	
cicogna		spazzolare		prognao		becchime	
coniglio		guadagnare		congelare		fretta	
tappeto		spusa		fraintendere		polveroso	
pappagallo		racozzi		solci		scogliera	
spiaggia		ladro		troluretore		pebuito	
fannullone		incudine		squalo		dimprine	
prugna		stretto		prepimente			
furfante		fidorzato		vese			
liuto		accendere		ancegato		Tot. _____	
baciare		tacchino		gattonare			
martello		capelli		pozzo			
spegente		polmoni		allurazione			

How to score LexITA

To score LexITA simply follow this formula

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

Use the following list to identify word items. English translation for each word item is also reported in the list.

Accendere (to turn on), allurazione (NW), ancegato (NW), anipi (NW), annaffiare (to water), ascia (axe), avviso (notice), baciare (to kiss), balbettare (to stutter), becchime (birdseed), bollitore (kettle), brieca (NW), camicia (shirt), capelli (hair), cesto (basket), cicogna (stork), congelare (to freeze), coniglio (rabbit), cucchiaio (spoon), dimprine (NW), ecicitizione (NW), elerro (NW), fannullone (idler), fidorzato (NW), flasse (NW), formica (ant), fraintendere (to misunderstand), fregio (frieze), fretta (haste), frusta (whip), furfante (scoundrel), gattonare (to crawl), giurare (to swear), guadagnare (to earn), gunto (NW), incudine (anvil), inflose (NW), ladro (thief), liuto (lute), lucchetto (padlock), magro (thin), maledire (to curse), martello (hammer), matita (pencil), muschio (moss), nevicare (to snow), orgoglioso (proud), panagi (NW), pandispagna (sponge cake), pappagallo (parrot), parda (NW), pebuito (NW), pencato (NW), pessimo (very bad), pizzicare (to pinch), polmoni (lungs), polveroso (dusty), pozzo (well), predorto (NW), preoccupazione (concern), prepimente (NW), prognao (NW), prugna (plum), racozzi (NW), rapimento (kidnapping), rasso (NW), ricercare (to research), scarpa (shoe), scogliera (cliff), scondere (NW), siscera (NW), solci (NW), sonta (NW), spazzolare (to brush), spegente (NW), spiaggia (beach), spusa (NW), squalo (shark), starnutire (to sneeze), stretto (narrow), tacchino (turkey), tappeto (carpet), tarmezione (NW), tirchio (stingy), troluretore (NW), uccidere (to kill), uscamo (NW), vese (NW), vincere (to win), zucchero (sugar).

[1] Because Lemhöfer (personal communication, February 12, 2019) expressed unhappiness with the name Lextale used for tests other than the original LexTALE test, we no longer retain that name.
[2] The distribution of item frequency in the item set mirrors the distributions of words in a linguistic corpus (see Author *et alii*, 2018). Similar frequency ranges have been used for the Spanish and French tests.
[3] The Common European Framework for Languages (CEFR) provides a standard for the description of L2 learners 'proficiency. It defines six levels of proficiency (Beginners: A1, A2; Intermediate: B1, B2; Advanced: C1, C2) encompassing reception, production and communicative competences (see www.coe.int/lang-CEFR for more information). The CEFR level was assessed by each university through an ad hoc examination encompassing reading, speaking and listening abilities, as well as vocabulary and grammatical knowledge. On the basis of this exam, students are assigned to a CEFR level. The reported CEFR level in this experiment corresponds to the level to which each student was assigned.
[4] Reliability of the test is confirmed also when computing the Kuder-Richardson index for binary data with DescTools (Signorell *et alii*, 2019). We obtained KR20= 0.94 (confidence level=.05) for the full set of data, and KR20=0.88 (confidence level=.05) when we restrict our set to the L2 group.
[5] We obtained KR20= 0.95 (confidence level=.05) for the full set of data, and KR20=0.92 (confidence level=.05) when we restrict our set to include only the L2 group. These coefficients are consistent with Cronbach Alpha and ICC and show that the test is very reliable for both the full group and for the L2 group.