

LEXTALE_FR A FAST, FREE, AND EFFICIENT TEST TO MEASURE LANGUAGE PROFICIENCY IN FRENCH

Marc Brysbaert*
Ghent University

Recent research in our lab has confirmed that performance-based assessment of second language proficiency provides more useful information than subjective ratings. The objective assessment in those studies was based on the LexTALE tests published by Lemhöfer and Broersma (2012). These tests are available for English, Dutch, and German, but not for French. In the present study we report the construction of a French equivalent. Participants get a random sequence of 56 French words of varying difficulty and 28 French-looking non-words. They have to indicate which words they know. The results showed a big difference (effect size of $d = 3.6$) between native speakers and speakers of French as a second language. Because the test is not at ceiling level for native speakers, it can be used for the assessment of French language proficiency both in first and second language.

Psycholinguistic researchers investigating second language (L2) processing typically measure language proficiency by asking participants to rate their own performance on Likert scales (Lemhöfer & Broersma, 2012). Usually, a distinction is made between proficiency in listening, speaking, reading, and writing.

A problem with subjective assessments, however, is that they may not always be valid and comparable across studies. Lemhöfer and Broersma (2012), for instance, correlated them with translation tests and commercially available proficiency tests, and found considerable differences between a Dutch and a Korean sample of bilingual students. Subjective ratings may also be influenced by the context of the study. Because they are not performance based, they are vulnerable to demand characteristics. As a result, the assessment may be different depending on whether the participants take part in the experiment because of degree requirements or because they need money. Finally, as we will see below, the validity of self-rated assessments may also

* Marc Brysbaert, Department of Experimental Psychology, Ghent University.
This research is supported by an Odysseus grant awarded by the Government of Flanders to the author.
We would like to thank Katrien Lievois and Steve Majerus for their kind cooperation in collecting the validation data.
Correspondence concerning this article should be addressed to Marc Brysbaert, Department of Experimental Psychology, Ghent University, Henri Dunantlaan, 2, 9000 Gent. E-mail: marc.brysbaert@ugent.be

depend on the heterogeneity of the sample, as participants have a tendency to compare themselves to a restricted reference group.

In search of an alternative for subjective self-assessments, Lemhöfer and Broersma (2012) took inspiration from the literature on L2 education. In this literature, a number of tasks have been developed to measure language proficiency of students applying for courses. These tasks consist, among other things, of (word) translation tasks, tasks assessing (word) meanings with multiple choice questions, and grammatical judgment tasks (see Deltour, 1993, for an example of such a test in French). Lemhöfer and Broersma (2012) were particularly charmed by a test Meara developed in the 1980-90s (Meara & Buxton, 1987; see also Meara, 1992). This test simply asks participants to go through a list of stimuli and to indicate which words they know. To correct for response bias, the stimulus list includes plausible nonwords, typically in a ratio of 1 to 2. In the English proficiency test Lemhöfer and Broersma (2012) distilled from Meara's work, 20 nonwords were presented among 40 English words, and participants were asked to indicate which words they knew. This takes less than 5 minutes. Lemhöfer and Broersma (2012) found that the results on the basis of the test, which they called LexTALE (Lexical Test for Advanced Learners of English), in most analyses correlated more with the criterion than the subjective assessments. They further developed tests for German and Dutch (available at www.lextale.com), which unfortunately have not been normed and validated yet.

We became convinced of the usefulness of LexTALE, when we started to use it in our research. Diependaele, Lemhöfer, and Brysbaert (2013) reanalysed the larger word frequency effect in L2 speakers than in native speakers (L1) in the progressive demasking task. They observed that the differences in the frequency effect could completely be explained by differences in vocabulary size as estimated with LexTALE. That is, participants with a small vocabulary size had a larger word frequency effect than participants with a large vocabulary size, independently of whether the language was L1 or L2. In another study, Khare, Verma, Kar, Srinivasan, and Brysbaert (2013) tested a finding of Colzato, Bajo, van den Wildenberg, Paolieri, Nieuwenhuis, La Heij, and Hommel (2008), which suggested that the attentional blink is larger in bilinguals than monolinguals. The attentional blink refers to 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. Khare et al. (2013) asked a large group of Hindi-English bilinguals to complete an attentional blink task and measured their L2 proficiency both with self-assessment and LexTALE. Whereas no correlation was found between self-assessments and attentional blink, a significant positive correlation was found between the LexTALE scores and the size of the attentional blink: participants with high proficiency scores for English L2

showed a larger attentional blink than participants with low proficiency scores. In this study participants were paid for their participation, which may have been an incentive for some to rate their proficiency as higher than warranted (some participants indeed showed rather high levels of false alarms on the nonwords).

LexTALE tests are currently available for English, Dutch and German. For our own research, it would be good to have a similar test in French. Meara (1992) developed a test for students taking French courses in the UK (available at <http://www.lognostics.co.uk/tools/index.htm>; retrieved on June 4, 2012), but no empirical data are available about its performance. In addition, the test differs from the LexTALE format, as it has five word frequency bands and only one band of nonwords, which drastically increases the word/non-word ratio. The words also came from the upper frequency tail, meaning that chances of a ceiling effect for native speakers are high, excluding research like Diependaele et al.'s (2013) in which performance in L2 is compared to performance in L1. A ceiling effect is also likely to be a problem for two tests developed by Cobb (available at <http://www.lexutor.ca/tests/>; retrieved on June 4, 2012). These tests include but very high-frequency words belonging to the top one or two thousand, which are known to everyone with basic knowledge of French. A more interesting test was assembled and tested on a small group in a Master's Thesis of Hommersom (2003). Here again, however, we did not have the feeling that the selection of words and nonwords had been given enough consideration.^[1]

Because of the shortcomings in the existing tests, we decided to build a new test from scratch, based on data recently collected within the Lexique 3.72 enterprise (New, Pallier, Brysbaert, & Ferrand, 2004) and the French Lexicon Project (FLP; Ferrand, New, Brysbaert, Keuleers, Bonin, Meot et al., 2010). In the former (available at <http://www.lexique.org/>; retrieved on June 4, 2012) word frequencies for 135 thousand French words (47,342 lemmas) have been collected from written sources and film subtitles. In addition, the database contains a variable, *deflem*, which says how many participants (on a total of some 20 per word) in an offline word judgment task indicate they know the word. This is an interesting source to select words of various difficulty levels. Other such information can be found in FLP. In this project lexical decision times were collected for over 38 thousand French words and the same number of nonwords. For each word and nonword there is information about the reaction time and the percentage correct. This makes it easier to

1. There is also the opportunity to indicate which French words one knows on the website <http://www.abysum.com/Mesmots/>, in order to get an estimate of one's vocabulary size. However, because the website includes only two nonwords per session, the interpretation of the outcome critically depends on the participants' honesty to only select the words they know.

select words from the entire difficulty range. An initial sample of 60 words and 60 nonwords was selected (see below) and subsequently tested. Although the existing LexTALE tests only include 40 words and 20 nonwords, we opted to have a larger number to start with, so that we could drop items in case they turned out not to be good.

Method

Materials

Sixty words of various frequencies and difficulty levels were selected from Lexique and FLP. Given that we wanted to cover the entire range of French speakers, going from very little L2 knowledge to native speaker, we selected words going from high-frequency words known to everyone (*cheveux*, *pomme*, *église*) to low-frequency words known by only a few native speakers (*capeline*, *tanin*, *treillage*). All in all, 17 words had a film lemma frequency of less than 1 per million words (pm), 11 had a frequency between 1-5 pm, 16 had a frequency between 5-10 pm, 9 had a frequency between 10-20 pm, 6 had a frequency between 30-100 pm, and one (*cheveux*) had a frequency of more than 100 pm (see the file with supplementary materials for more information). Only words having little orthographic overlap with Dutch translations were selected (i.e., no cognates). Compounds and derived words were avoided (see below for two that were missed and how they performed). A further 60 nonwords were selected from FLP. Here again the selection ranged from nonwords easily rejected by native speakers to nonwords often eliciting errors in speeded lexical decision (with a majority of the latter). All stimuli can be found in the Supplementary Materials.

Administration

The Meara tests have been administered in different ways. One point of difference is whether participants must respond to each stimulus or just indicate the words they know. Lemhöfer and Broersma (2012) asked their participants to make a yes/no decision to all items. Given that our test had more trials and because we expected some of our participants to know very few words, we feared such a practice could be demotivating. Therefore, we simply asked participants to indicate which words they knew. Perea and colleagues (e.g., Gómez, Ratcliff, & Perea, 2007; Moret-Tatay & Perea, 2011) have shown that a go/no-go task, in which participants only respond to words, is as good as the traditional lexical decision task and may even be better, because participants can focus entirely on the words.

The instructions were as follows (they are available in three languages: Dutch, English, and French): *Hi, this is a test of French vocabulary. On the next page you will find 120 sequences of letters that look “French”. Only some of them are real words. Please, indicate the words you know (or of which you are convinced they are French 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 “oui”, “tu”, “jamais”, “université” and “oublier”, you indicate this as follows:*

| | | | |
|----------|------|------------|------|
| | | | |
| Stimulus | Mot? | Stimulus | Mot? |
| ainrir | | jamais | ✓ |
| osatrome | | reure | |
| oublier | ✓ | favernais | |
| avero | | oui | ✓ |
| repole | | université | ✓ |
| tu | ✓ | pourpreme | |

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

The stimuli were presented in a fixed, random order. The test was administered to four groups: (1) first-year students of educational sciences at the Dutch-speaking Ghent University; (2) a few acquaintances of the author with very little knowledge of French; (3) some 40 students from the Dutch-speaking Artesis College who were studying for professional translator; and (4) first-year students from the French-speaking University of Liège taking psychology courses. Each participant was additionally asked to give their gender, native language, the number of years they had taken French courses in school, and their self-rated proficiency in French (max = 10). In total, 316 participants returned the questionnaire. Of these, 152 were native French speakers and 164 had French as second language.

Results

First, for each stimulus we assessed its quality by calculating the point-biserial correlation between the pattern of answers to that question and the overall scores of the participants. This revealed that the correlation was positive for all words, going from .10 for *cheveux* (hair) to .87 for *escroc* (crook). So, for each word we found that students knowing few words were less likely to indicate they knew the word than students knowing many words. In contrast, nearly half of the nonwords had negative correlations, meaning they were more likely to be selected by persons with good knowledge of French than by persons with poor knowledge. This was particularly true for nonwords with missing accents (*bergere* instead of *bergère* [shepherdess], *vouter* instead of *voûter* [to bend]), nonwords that were part of fixed expressions (*tatin* from *tarte tatin* [Tatin pie]), and pseudohomophones of less well-known words (pseudohomophones are nonwords that sound like words, such as *ivoir* instead of *ivoire* [ivory], *cigard* instead of *cigare* [cigar]). The 30 nonwords with the lowest item-total correlations were discarded from the set.

In a second step, we ran an item response theory (IRT) analysis on the remaining 90 items (with the R package *ltm*; Rizopoulos, 2006). An IRT analysis assumes that all items of a questionnaire measure the same latent trait (language proficiency). It allows the researcher to see how items are responded to throughout the ability range. This analysis showed that the items not only differed in their difficulty but also in their discrimination power. Discrimination 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). As can be seen in Figure 1, discrimination was higher, for instance, for *cheveux* (hair), *cloche* (bell), and *infâme* (infamous) than for *fascine* (faggot) and *canoter* (boating or to row).

The IRT analysis further indicated that four words did not perform well. They were the rather easy words *pomme* (apple) and *accourir* (to rush) and the extremely difficult words *coterie* (coterie) and *coutil* (twill). The few people who selected the last two words were not the highest performing, so that it looked as if the words were better known in the mid-range than at the high end. *Pomme* was known by nearly everyone, but was missed by a high-performing participant (in all likelihood because of a slip of attention). Finally, *accourir* was missed by a rather high percentage of participants throughout the ability range, presumably because it is a derivative of the high frequency word *courir* (to run). No precision was lost by dropping these four words. This also allowed us to drop two further less well performing nonwords: *col-lie* (possibly interpreted as a dog or mistaken for *colis* [parcel]) and *collition* (probably mistaken for *collision* [collision] or *coalition* [coalition]).

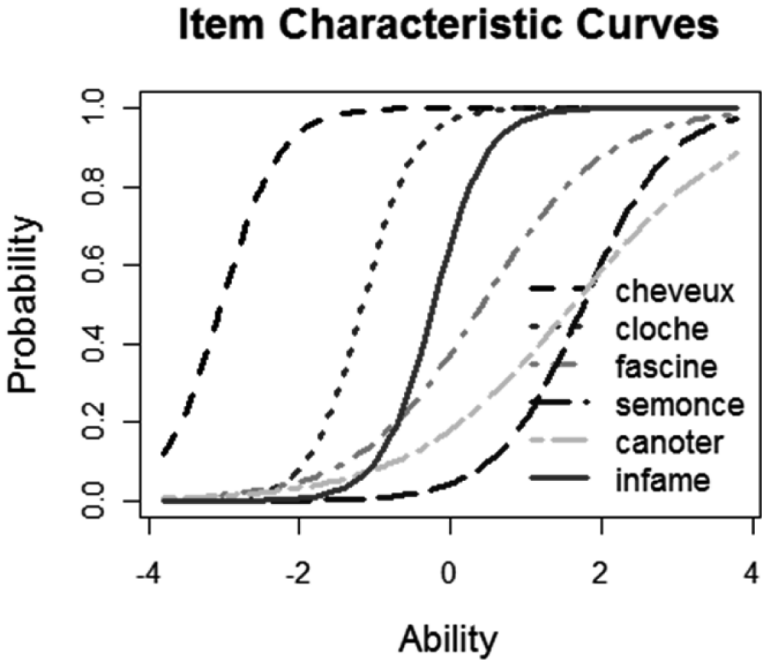


Figure 1

Item characteristic curves for some words included in the questionnaire. From this IRT analysis, one can conclude that cheveux was easier than cloche and infâme, etc. At the same time, some words have a steeper transition from not known to known than others. In particular canoter has a rather flat transition, arguably because it is a specialist word, not known to all of the high performing participants. Fascine also has a less steep transition, possibly because it is a form of the verb fasciner (fascinate) as well as a low-frequency noun (referring to a bundle). The steepness of the curve is called item discrimination

In a third step, we looked at different ways to summarise the participants' scores based on the remaining 56 words and 28 nonwords. Lemhöfer and Broersma (2012) tried out three measures of which the easiest performed best. For the present test it is calculated as follows:

$$\frac{(N_{\text{word trials correct}} + 2 * N_{\text{nonword trials correct}})}{112}$$

This measure counts the mean percentage of correct trials, taking into account that the nonwords trials are only half in number (hence, the multiplication by

2 and the division by 112 instead of 84, the number of items). The Lemhöfer and Broersma measure in theory goes from .0 to 1.0, but in practice only from .5 to 1.0, because to get values below .5 the participants have to select more nonwords than words as “items known”. The measure is the linear equivalent of an equation used at Ghent University in multiple choice exams to correct for guessing. It is (applied to the present situation):

$$\text{Ghent score} = N_{\text{words selected}} - 2 * N_{\text{nonword selected}}$$

This equation goes from less than 0 (when more nonwords are selected than words) to 56 (when all word and nonword trials are correct).

Finally, because the questionnaire data are the outcome of signal detection (SDT) decisions, it is possible to calculate an SDT measure of sensitivity, such as d' . The easiest way to do this is by making use of a built-in Excel function (Stanislaw & Todorov, 1999):

$$D' = \text{NORMSINV}\left(\frac{N_{\text{words selected}}}{56}\right) - \text{NORMSINV}\left(\frac{N_{\text{nonword selected}}}{28}\right)$$

To avoid calculation errors, 56 was replaced by 55.5 if $N_{\text{words selected}}$ reaches the maximum, and 0 by 0.5 if no nonwords have been selected (i.e., $N_{\text{nonword selected}} = 0$).

Figure 2 (p. 31), shows a scatterplot of the Ghent scores and the d' -values for the 316 participants who completed the questionnaire. Because the correlation between both measures is .984, for most practical purposes they will lead to the same conclusion.

The Ghent scores ranged from -13 to 56. There was a big difference in performance between the native French speakers ($M = 42.8$, $N = 152$, $SD = 6.71$) and the respondents who had French as second language ($M = 8.3$, $N = 164$, $SD = 11.61$; effect size $d = 3.64$). Most of the latter participants had Dutch as native language, although some had German, Luxemburgish, Spanish, Turkish, or English as mother tongue.

Figure 3 (p. 32) shows the correlation between the Ghent score and the years of French courses in school. Although the correlation is clearly present ($r = .641$, $N = 304$ ^[2], $p < .01$), there is a considerable degree of scatter in the data. Particularly noteworthy is the big group with 8 years of French courses in school and poor performance on the test. This is the group having taken compulsory French courses in primary (2 years) and secondary education (6

2. Not all participants filled in this part.

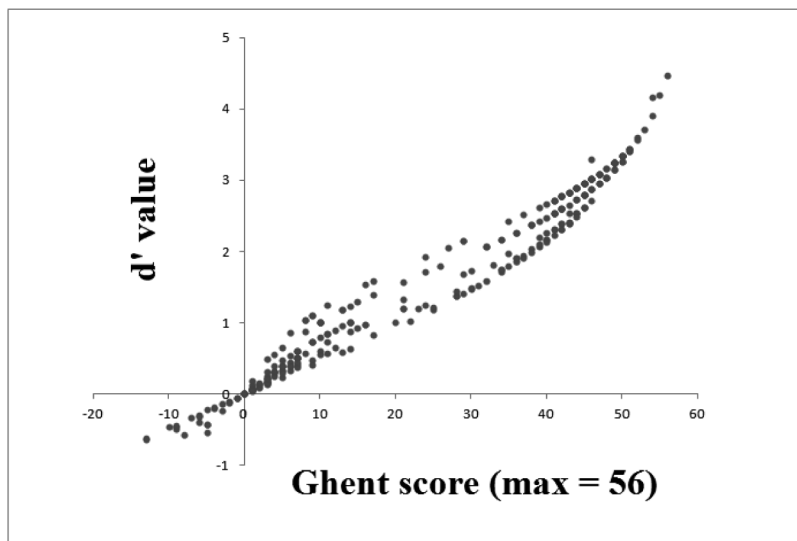


Figure 2

Relationship between the Ghent score and d'

Due to different weights given to false alarms (nonwords selected), there is some difference in d' values for equal Ghent scores in the midrange. d' also gives slightly more weight to the best performing participants. Still, the overall correlation between both measures is .984, meaning that for practical purposes it will not make a difference which score is used

years), apparently without much result as far as the tested words are concerned.

Finally, Figure 4 (p. 32) shows the relationship between the self-assessments and performance on the test. Again, the correlation is there, but far from perfect ($r = .579$, $N = 310$, $p < .01$). Particularly in the midrange (5-8) very different LEXTALE scores are obtained from people giving themselves the same subjective assessment.

Discussion

We have presented the French equivalent of the LexTALE vocabulary tests, which in turn are an offspring of the tests proposed by Meara (Meara, 1992; Meara & Buxton, 1987). For continuity reasons, we propose to call this test LEXTALE_FR (LEXical Test for Advanced LEarners of FRench). As Figure 2 shows, it covers the entire proficiency range from people without (much)

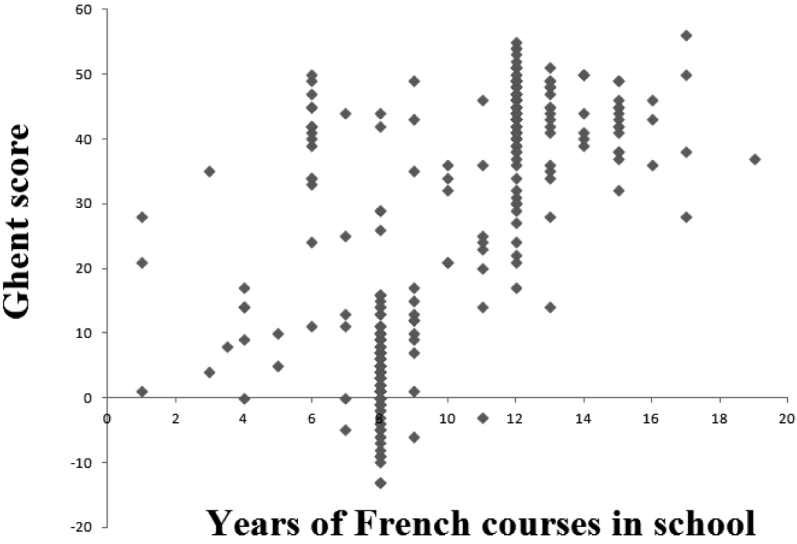


Figure 3

Relationship between years of French courses in school and performance on the test

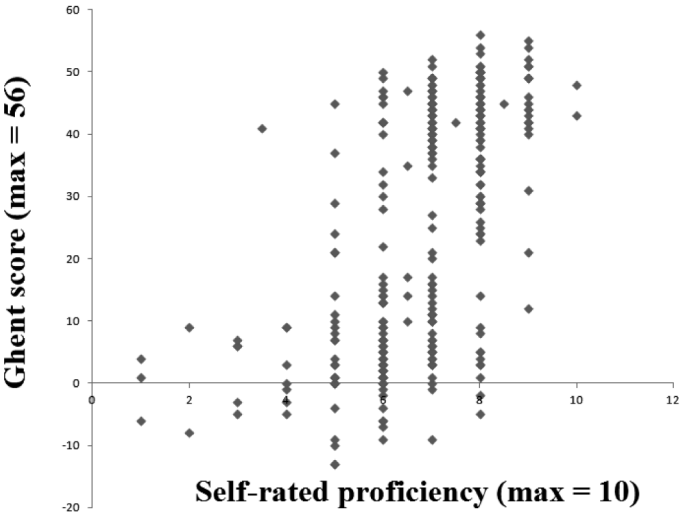


Figure 4

Correlation between self-assessed proficiency and test performance

knowledge of French to advanced native speakers. We wanted such a broad test, because this allows us to further examine Diependaele et al.'s (2013) finding that vocabulary size may have the same effect in L1 and L2. Because the test must cover the entire range, we feel that it would be a bad idea to limit the number of words to the 40 we originally had in mind (in line with Lemhöfer and Broersma, 2012), as all 56 words contributed to the test's performance. The extra items are also useful because they increase the reliability of the test. Cronbach's alpha for LEXTALE_FR according to the present study is .96, against .81 for the English LexTALE. Of course, it would be good to have this value replicated in a new study without the 36 discarded items (see also below for two more, small changes). Although we do not expect big changes, the absence of misleading nonwords may alter the participants' responses to some extent (e.g., making them more inclined to select words they are not 100% confident of).

Our item analysis further reminds us of the importance of this analysis. In particular (and not really expected), a large part of the nonwords turned out to be suboptimal. This may have been due to the way in which we chose them. Because we selected nonwords that induced errors in the FLP lexical decision task, we may have selected nonwords that were too difficult even for native speakers in an offline task. As a result, participants who knew the words from which the nonwords were derived tended to mark the stimuli as known. In particular, nonwords with missing accents were not a good idea, possibly because accents often have to be omitted on electronic devices working with ASCII codes and because accents are left out when words are written in capitals. Also, pseudohomophones of low-frequency words were a problem, in line with Van Orden's (1987) suggestion that these words are often recognised on the basis of their phonology because the orthography is not fully mastered.

A further, in-depth look at the nonwords indicated that two more could be criticised, even though they performed well in the present study.^[3] These were *oeiller* and *replaner*. *Oeiller* is only a true nonword if one makes a distinction between *oe* and *œ*, because *œiller* is a very-low frequency word.^[4] *Replaner* could be interpreted as a prefixed word (*re+planer*), meaning "*planer à nouveau*" (to plane again). To avoid discussion about their status at the high end of the proficiency scale, it seems better to replace these nonwords by *oeuiller* (assuming that everyone who knows the meaning of *œiller*, also knows that its spelling must not include *u*) and *raplaner*.

3. As was kindly pointed out by Boris New.

4. From the medical dictionary: définition A: Sorte de petit manchon ovale et souple, qui sert à accueillir l'œil confortablement sur un viseur ou autre appareil optique; définition B: Voir, relouer, regarder de ses yeux, faire du voyeurisme.

It may be objected that we could have prevented many of the problems with the nonwords by using less word-like examples from the start. This is true, but we think that the nature of the nonwords is one of the reasons why our test is doing well across the entire range. A notion that is becoming increasingly important in psycholinguistic research is that of ‘lexical quality’ (Perfetti, 2007), defined as the degree to which written words are stored in memory as stable, integrated patterns of detailed phonology, orthography, and meaning (to be contrasted with the hazy feeling that a sequence of letters may refer to a word one has encountered a few times in a language and that may have some meaning). Lexical quality increases with print (language) exposure and is thought to be an important notion to understand differences between good and poor readers (Andrews & Hersch, 2010), and between L1 and L2 readers (Diependaele et al., 2013). To tap into this quality, it is important not to make a crude distinction between the words and the nonwords. Indeed, it has recently been shown that the use of easy nonwords makes it possible to do well on a word test even without knowledge of the language involved (Grainger, Dufau, Montant, Ziegler, & Fagot, 2012; Keuleers & Brysbaert, 2011). Particularly good nonwords in the present study were over-regularisations. These are irregular words that have been regularised, such as *metter* (instead of *mettre* [to put]) and *plaiser* (instead of *plaire* [to please]). They are also pseudohomophones of the verb forms *mettez* and *plaisez*. Such nonwords elicited many errors in the beginning language users but not in the advanced speakers.

As for the scoring, we think that the Ghent score is the easiest to calculate and interpret. A guessing participant will obtain a score close to 0. This is also true for someone who does not select a single item (and who according to the Lemhöfer and Broersma score would get .50, because all the nonword trials are correct). Only participants who know the words and are able to avoid the nonwords receive a high score. Authors who feel uneasy about the maximum of 56 can, of course, convert this number to 100 by multiplying the scores with 100/56.

The study of Khare et al. (2013) suggests that the LEXTALE scores are a better measure of language proficiency than subjective self-assessment. Figure 4 indeed shows that particularly for self-ratings of 5-8 there is a big variability in the words known. Only people who give themselves very high or very low values seem to have a good idea of their abilities. One factor that may have contributed to the variability is the fact that many mid-range participants tended to compare themselves to a narrow reference group (this became clear in the talks we had with some of them afterwards). Beginning L2 learners had the tendency to compare themselves to other beginning learners and, therefore, felt they were entitled to award themselves 6/10 even though the number of words they knew was small. Similarly, native speakers

had a tendency to compare themselves to other native speakers and so could give themselves 6/10 even though they knew many more French words than a typical L2 speaker. Indeed, as soon as L2 speakers master the 8,000 most common word families (consisting of a lemma plus its inflected forms and its most productive derived forms), they are considered as rather proficient in that language (Laufer & Ravenhorst-Kalovski, 2010), even though the language as a whole contains more than 30,000 word families.

LEXTALE_FR is not only interesting for second language researchers. Given the importance of lexical quality in L1 language processing (Andrews & Hersch, 2010), information about the participants' LEXTALE_FR scores is likely to be informative for all word recognition research in French. In addition, the test may be of interest to practitioners examining French-speaking patients. For instance, Baddeley, Emslie, and Nimmo-Smith (1993) proposed the Spot-the-Word test as a valid measure of premorbid intelligence in older adults. In this test, the participants were shown pairs of items consisting of a word and a non-word, and they had to identify the word. Correlational studies showed that the test correlated well with estimates of (premorbid) intelligence. Given that the LEXTALE test also involves the selection of words from nonwords, it may have the same properties as the Spot-the-Word test, although this should be tested first, given that the distinction between words and nonwords is rather subtle in the present test (see above).

References

- Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, 139, 299-318.
- Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology*, 32, 55-65.
- 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.
- Deltour, J.J. (1993). *Échelle de vocabulaire Mill Hill de J.C. Raven: adaptation française et normes comparées du Mill Hill et du Standard Progressive Matrices (PM38). Manuel et annexes*. Braine-le-Château: Éditions L'Application des Techniques Modernes SPRL.
- Diependaele, K., Lemhöfer, K., & Brylsbaert, M. (2013). The word frequency effect in first and second language word recognition: A lexical quality account. *Quarterly Journal of Experimental Psychology*. In press.
- Ferrand, L., New, B., Brylsbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for

- 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42, 488-496.
- Gómez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, 136, 389-413.
- Grainger, J., Dufau, S., Montant, M., Ziegler, J.C., & Fagot, J. (2012). Orthographic processing in baboons (Papio papio). *Science*, 336, 245-248.
- Hommersom, M. (2003). *Het testen van receptieve woordenschatkennis*. MA Thesis. University of Utrecht, http://www.ethesis.net/woordenschat/woordenschat_inhoud.htm; retrieved on June 4, 2012.
- 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. *Behavior Research Methods*, 44, 325-343.
- Meara, P.M. (1992). *New approaches to testing vocabulary knowledge*. Unpublished manuscript. Swansea: Centre for Applied Language Studies.
- Meara, P.M., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142-154.
- Moret-Tatay, C., & Perea, M. (2011). Is the go/no-go lexical decision task preferable to the yes/no task with developing readers? *Journal of Experimental Child Psychology*, 110, 125-132.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments & Computers*, 36, 516-524.
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11, 357-383.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17, 1-25.
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31, 137-149.
- Van Orden, G.C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & cognition*, 15, 181-198.

Appendix Availability

LEXTALE_FR is available as pdf with instructions in three different languages (French, Dutch, English). It will also become available on a website. Further information can be found at <http://crr.ugent.be/programs-data>, where there is also a supplementary Excel file with the full results of the present study and more information about the stimulus materials.

Selected stimuli (in the order of the list) with their English translations

cheveux (hair); soumon (NW); cloche (bell); fascine (faggot, fascinate); huif (NW); semonce (lecture); canoter (boating); infâme (infamous); fourmi (ant); cadenas (padlock); racaille (riffraff); pourcine (NW); œillet (carnation); raplaner (NW); plaiser (NW); cerveler (NW); endifier (NW); jamain (NW); ennemi (enemy); pouce (thumb); metter (NW); fosse (pit); inciter (encourage); salière (salt shaker); fouet (whip); censure (NW); clouer (nail); mappemonde (globe); gloque (NW); lézard (lizard); sacher (NW); nouer (establish); occire (slay); écouce (NW); osseaux (NW); rejoute (NW); escroc (crook); hache (ax); parchance (NW); pinceau (brush); poisson (fish); robinet (tap); amadou (coax); peigne (comb); retruire (NW); crayon (pencil); sentuelle (NW); alourdir (to make heavy); marteau (hammer); esquif (skiff); treillage (trellis); dauphin (dolphin); orgueil (pride); amorce (prime); cintre (hanger); chameau (camel); bouton (button); capeline (wide-brimmed hat); lanière (strap); honte (NW); abêtir (stultify); fenêtre (window); écureuil (squirrel); caddie (caddy); détume (NW); oeuiller (NW); balai (broom); prioche (NW); vicelard (pervert); joues (NW); agire (NW); éventail (range); boutard (NW); panier (basket); citrouille (pumpkin); bouilloire (kettle); parir (NW); remporter (win); procureux (NW); tanin (tannin); église (church); indicible (unspeakable); réporce (NW); mignon (cute).

Received June 9, 2012

Revision received September 3, 2012

Accepted September 20, 2012