A New Receptive Vocabulary Size Test for French

Roselene Batista and Marlise Horst

Abstract: Researchers have developed several tests of receptive vocabulary knowledge suitable for use with learners of English, but options are few for learners of French. This situation motivated the authors to create a new vocabulary size measure for French, the *Test de la taille du vocabulaire* (TTV). The measure is closely modelled on Nation's (1983) Vocabulary Levels Test (VLT) and follows the guidelines written by Schmitt, Schmitt, and Clapham (2001). Initially, a pilot version was trialled with 63 participants; then an improved version was administered to 175 participants at four proficiency levels. Results attest to the TTV's validity: mean scores across the four frequency sections decreased as the tested words became less frequent, and more proficient learner groups outperformed less proficient groups. The TTV in its current form is intended to be of practical use to teachers and learners, but it is also expected to evolve; ideas for future improvements are discussed.

Keywords: frequency, French L2 vocabulary, vocabulary size, assessment

Résumé : Des chercheurs ont développé plusieurs tests de vocabulaire réceptif pour les apprenants d'anglais, mais les options pour les apprenants de français ne sont pas nombreuses. Ce scénario a motivé les auteurs à créer un nouvel outil qui mesure la taille du vocabulaire en français, le Test de la taille du vocabulaire (TTV). Cet outil repose sur le modèle du *Vocabulary Levels Test*, conçu par Nation (1983), et suit les directives proposées par Schmitt, Schmitt et Clapham (2001). Premièrement, une version pilote a été testée auprès de 63 participants, ensuite une version améliorée a été complétée par 175 participants de quatre niveaux de compétence distincts. Les résultats confirment la validité du test: les moyennes obtenues par les participants à chacune des quatre sections décroissent au fur et à mesure que les mots deviennent moins fréquents et les groupes plus avancés ont obtenu des moyennes plus élevées par rapport aux groupes moins avancés. Le TTV, dans sa forme présente, se veut un outil pratique, conçu pour être utilisé par des enseignants et des apprenants. Néanmoins, on espère que le test évolue; quelques pistes de réflexion sur des améliorations seront discutées.

Mots clés : évaluation, fréquence, taille du vocabulaire, test en français

doi:10.3138/cmlr.2820

This article reports on the development and trialling of a new test for learners of French. The test is a measure of vocabulary size or breadth, which is defined as the number of words a learner of a new language can recognize and link to basic meanings (Milton, 2009; Nation, 2013). Being able to approximate the vocabulary sizes of learners of French, English or any other new language is useful for researchers and educators and also for learners. Researchers use tests of vocabulary size to answer questions about the development of second language (L2) lexis and its relationship to other aspects of language knowledge. For example, Stæhr's (2008) investigation of Danish learners of English found that recognition knowledge of 2,000 frequent word families – consisting of a headword and its basic inflected and derived forms – was an important predictor of success on reading, writing, and listening exams given at the end of their secondary schooling. The connection to reading comprehension in Stæhr's study was particularly strong, with vocabulary size accounting for 72% of the variance in scores. In educational contexts, tests of vocabulary size can be helpful in placing students in language courses, and classroom teachers can use such tests to diagnose their learners' needs and select appropriate materials. Learners, too, are eager to know where they stand, and vocabulary size scores can provide clear information upon which learners can act.

An important characteristic of receptive size measures is the frequency-informed selection of test items. Analyses of large corpora have resulted in lists of the most frequent word families or lemmas of a language (a lemma differs from a family in that it includes a headword and its inflections but does not include derived forms). These frequency lists are used by size-test builders to systematically sample vocabulary items from a range of frequency levels. Test-takers are asked to indicate their ability to recognize the meanings of the sampled vocabulary in some way, for example, by choosing correct definitions in a multiple-choice format. Estimations of size are then based on test-takers' performance at each of the various frequency levels sampled by the test.

Frequency-informed size testing is based on the assumption that the most frequent words of a language will be learned early, while less frequently encountered words will be learned later. An overview of research by Milton (2009) provides evidence that this assumption is sound: in studies of both English and French, L2 learner populations scored highest on the section of the size test that assesses the most frequent vocabulary, with a pattern of decreasing scores on sections that test less frequent vocabulary. But it is also clear that in the cases of certain words and certain learners, learning may not always follow a

© *CMLR/RCLV*, doi:10.3138/cmlr.2820

strict frequency order. Research by Bardel, Gudmundson, and Lindqvist (2012) shows that Swedish-speaking learners of French can readily recognize some infrequent French words (e.g., *électrique*, *vocation*) due to their resemblance to words that have been borrowed into Swedish. These researchers also identified "thematic" words that are rather infrequent in French generally but are likely to be known to learners because they occur frequently in classroom input. Nonetheless, frequency appears to be a powerful factor in vocabulary learning across groups of learners. Milton's (2009) analysis of learnability factors found the frequency of a word to be a much stronger predictor of its being learned than cognateness, word length, and part of speech.

It is important to note that the number of L2 word families that a learner knows receptively is just one of several measurable dimensions of vocabulary knowledge (see Nation [2013] for an overview); measures of a variety of other kinds of lexical knowledge have been developed and tested. Instruments used by researchers interested in the acquisition of French, for example, include measures of lexical diversity in speech production (Tidball & Treffers-Daller, 2007), "depth" measures that assess learners' ability to recognize collocates and other word associations (Bogaards, 2000; Greidanus, Bogaards, van der Linden, Nienhuis, & de Wolf, 2004), and profiling software that identifies proportions of advanced lexis in speech samples (Bardel et al., 2012). A 2014 study by Forsberg Lundell and Lindqvist uses innovative measures of productive collocation ability and lexico-pragmatic knowledge. Generally, the instruments mentioned above have proved their usefulness in answering questions about relationships between different kinds of lexical knowledge and effective ways of distinguishing between groups of varying proficiency levels. But most of them target fairly advanced university learners of French, and they are accessible mainly to researchers. In our view, there is a need for a new, freely available French vocabulary size test suited to assessing learners of a wide range of proficiencies. It is also important that the test be easy to administer and that it produce readily interpretable scores. Our size test was designed with these practical goals in mind.

In this paper, we detail the development of the *Test de la taille du vocabulaire* (henceforth the TTV) and report the results of administering it to 175 immigrant learners of French in Québec. But first we take a closer look at test formats that have been developed to assess L2 vocabulary size and review earlier size findings, with particular attention to learners of French.

© *CMLR/RCLV*, doi:10.3138/cmlr.2820

Investigating receptive vocabulary size

One widely used size test is the Eurocentres Vocabulary Size Test by Meara and Jones (1990), along with its computerized version, X-Lex, by Meara and Milton (2003); it is available in English and several other languages and is the only test of which we are aware that assesses L2 French vocabulary size. The format requires test-takers simply to check the box next to a word if they know its meaning; sample items from the English version are shown in Box 1; henceforth we refer to this format as the "checklist test." A notable feature of the checklist format is the inclusion of plausible non-words among the target items (e.g., *galpin* in Box 1). These function as a check on overestimations, such that ticking a non-word as known results in a downwards adjustment of the test-taker's score. In another well-known test, the Vocabulary Levels Test (Nation, 1983; Schmitt, Schmitt, & Clapham, 2001), test-takers are asked to identify the correct simply worded definition of a target English word in a matching format (Box 2). The Vocabulary Size Test (VST), an instrument designed for learners of English by Nation and Beglar (2007), presents target words in short, contextualized sentences with four multiple-choice answer options (Box 3).

Box 1. Sample questions from the checklist test

1 galpin	[ ]	2 impulse	[ ]	3 suggest	[ ]
4 advance	[ ]	5 peculiar	[ ]	6 benevolate	[ ]
7 indicate	[ ]	8 needle	[ ]	9 destruction	[ ]

Box 2. Sample cluster from the Vocabulary Levels Test

1. desolate	
2. fragrant	
3. gloomy	_____ good for your health
4. profound	_____ sweet-smelling
5. radical	_____ dark or sad
6. wholesome	

Box 3. Sample question from the Vocabulary Size Test

MINIATURE: It is a miniature.
a. a very small thing of its kind
b. an instrument to look at small objects
c. a very small living creature
d. a small line to join letters in handwriting

What are some of the typical vocabulary sizes identified using these instruments? An investigation using the VST, reported by Nation (2013), found that learners of English who were able to perform adequately in undergraduate studies at an English-medium university had vocabulary sizes of 5,000 to 6,000 word families. Learners studying at the doctoral level were found to have a vocabulary size of around 9,000 English families. Analyses of the coverage of frequency lists for a variety of text types by Nation (2006) show that learners would need knowledge of the 8,000 to 9,000 most frequent English word families to understand 98% of the vocabulary that occurs in novels written for native speakers. The 98% criterion is based on research by Schmitt, Jiang, and Grabe (2011) and others (see Nation [2013]), which indicates that knowledge of 98% of the words in a text is a reasonably good guarantee that it will be comprehended adequately. Size research has also investigated native speakers; a study by Goulden, Nation, and Read (1990) indicates that university-educated adults may know around 20,000 English word families.

What does vocabulary size research have to say about learners of French? In a study tellingly entitled "Language Lite," Milton (2006) reports that after hundreds of hours of French study over seven years in British secondary school programs, learners were found to have a recognition vocabulary size of only 1930 lemmas (*SD* = 475), according to mean scores on the checklist test. Similar modest figures are reported by David (2008) in a study that also investigated secondary learners in Britain using the checklist test. In a follow-up to his study of secondary learners, Milton (2008) investigated university students. He reports that after an additional four years of study at a British university, including a year abroad spent in France, students' mean French vocabulary size reached 3,326 lemmas (*SD* = 579).

Although the research discussed above sheds some light on the amount of vocabulary needed to complete school French programs in Britain, many other questions remain. To our knowledge, the

© *CMLR/RCLV*, doi:10.3138/cmlr.2820

vocabulary sizes that learners would need to read a French novel without assistance, follow the dialogue of a movie, study at a French-medium university, or achieve other learning goals they may have are largely undetermined. Corpus counts by Cobb and Horst (2004) suggests that knowledge of words on the 2,000 most frequent French list is likely to be a powerful asset, offering a possibly even higher level of known-word coverage than in English. But to our knowledge this potential has not been investigated experimentally with L2 learners. Nor are we aware of research that specifies the number of words that native speakers of French can recognize, as indicated by their performance on a size test.

One explanation for this research shortfall may be the unavailability (until recently) of good corpus-based frequency lists for French. Another may be the limitations of the single available size measure for French that might be used to address such questions, namely the checklist test. Several researchers have found that the non-words used as a check on overestimations in this test are a source of unreliability, with learners in one context far more likely to risk saying "yes" to non-words than those in another (Milton, 2009). Eyckmans, van de Velde, van Hout, & Boers (2007) report that 60% of non-words were identified as real by the Belgian students they investigated. The fact that "yes" answers to real words are unverifiable is another concern: when a test-taker indicates that a word is known, it must be taken on faith that the definitional meaning he or she has in mind is correct. David (2008) notes the need for another type of vocabulary test to confirm research findings based on the yes-no checklist instrument. These problems are relevant to conducting experimental research, but it is also possible that the checklist test lacks credibility with classroom teachers and learners because the self-report format may not look like a "real" test.

All these concerns informed our decision to create and trial a new French size test suitable for use in classrooms, the *Test de la taille du vocabulaire* (TTV). Our study has two main purposes: The first is to report on the development of the test itself; the second is to assess its effectiveness by administering the test to groups of L2 French learners at different proficiency levels, and by interpreting the results. In this we have used as a model the updated Vocabulary Levels Test (VLT) for English by Schmitt et al. (2001), for reasons discussed in the Methodology section below. To test the performance of the new measure, we investigated the following research questions:

(a) Is the TTV implicational, such that learners of French score higher on the test of the 2,000 most frequent lemmas, lower on the 3,000-level words, lower still on the 5,000, and so on?

© *CMLR/RCLV*, doi:10.3138/cmlr.2820

(b) Is the test able to distinguish between groups of varying levels of proficiency? That is, do learners in higher-proficiency groups have larger vocabulary sizes than learners in lower-proficiency groups?

The focus of the first question is the validation of the frequency aspect of the test. Since word frequency has been shown to be a strong predictor of L2 word learning in previous research, we hypothesize that the testing will reveal a pattern of decrease in scores on the four sections corresponding to the decrease in frequency of the targeted words. The question concerning proficiency level also explores the extent to which the test is functioning as intended; we hypothesize that more proficient learners (as identified by their performance on a placement test) will have higher scores than less proficient ones.

Methodology

We begin with an account of the development of a pilot version of the TTV. Subsequent sections describe the methodology of the main validation study.

Piloting the test

Design

As mentioned, in choosing the format for a size test for French to complement the existing yes-no checklist test (and to avoid some of its limitations), we were interested in verifiable responses whereby test takers "prove" that they know a word by identifying a correct definition. Tests of English vocabulary size that could serve as models are the VLT shown in Box 2 and the VST shown in Box 3. Both have verifiable answer formats and strong track records in experimental research (Read, 2000; Schmitt et al., 2001; Nation, 2013). An important reason for eventually choosing the VLT is its efficient presentation using question clusters (see Box 2 for an example of a cluster). In each cluster, test-takers consider six answer options and make matches to three definitions. Since the same set of six answer options is "recycled" three times within the cluster, a great deal less reading is required on the part of test takers (and less writing on the part of test designers) than is the case in the standard multiple-choice format used by the VST, which presents four answer options for each target word.

A possible point in favour of the VST is its sampling of English words from all of 14 corpus-based frequency bands, which gives it the ability to test a wide range of learner knowledge. But French word lists drawn from a large modern corpus at 14 different levels of

frequency were not available to us. However, we were able to take advantage of recent work by Lonsdale and Le Bras (2009), whose list of the 5,000 most frequent lemmas is based on a 23-million word corpus of current written and spoken international French. Unlike earlier French corpora that are based largely on written language (e.g. Baudot, 1992; Verlinde & Selva, 2001), this corpus has a substantial spoken component (50%). The Lonsdale and Le Bras list was used to construct a pilot version of the TTV following the VLT, with sections that sample the 2,000, 3,000, and 5,000 frequency levels. The VLT also has sections that test the 10,000 level as well as Coxhead's (2000) Academic Word List, a list of families that occur frequently in university textbooks. Like the VLT, the TTV tests words at the 10,000 level, but since the Lonsdale and Le Bras lists go only as far as the 5,000 frequency level, we turned to an older list by Baudot (1992) to create this part of the test. The decision to include the same frequency levels as the VLT was made with a view to enabling eventual comparison studies of L2 English and French vocabulary development. However, we did not include a section on the TTV that parallels the Academic Word List section on the VLT. Such a list has not been determined for French, and it may well not exist. Research by Cobb and Horst (2004) indicates that while a distinct academic lexis (largely Greco-Latin) is characteristic of English, this is likely not the case in French or other Romance languages.

Creating the pilot TTV involved first sampling the test words and distractors at random from the 2,000, 3,000, 5,000 and 10,000 frequency lists (henceforth referred to as 2K, 3K, 5K, and 10K) and creating test clusters for each level. Each cluster consists of six words from the same word class and three simply worded definitions. In the pilot version of the TTV, there were six noun clusters, three verb clusters, and three adjective clusters per section; this distribution reflects (roughly) the representation of word classes on the Lonsdale and Le Bras (2009) list. Twelve clusters were created for each section of the pilot test with a view to retaining the ten that functioned best in the final version. The definitions of the target words were kept short (to reduce reading to a minimum) and syntactically simple. To help ensure comprehensibility, definitions consisted entirely of words taken from a more frequent level than the test words themselves: words tested in the 2K frequency section have definitions using words from the 1K list in the Lonsdale and Le Bras (2009) list; test words on the other sections are defined using words taken from the 1K and 2K lists. A sample item from the 5K frequency section is shown in Box 4. Five native speakers of French took the test and achieved perfect or near perfect scores.

Box 4. A noun cluster from 5K frequency section of the TTV

1. brouillard		
2. coïncidence		
3. farce	_____	une histoire qui fait rire
4. instituteur	_____	ce qui empêche de voir loin
5. pneu	_____	un professionnel de l'éducation
6. soumission		

The pilot test was administered to 63 adult immigrant learners from a variety of first language backgrounds in a government-sponsored program at a school in Montreal. All were enrolled in French courses specially conceived to integrate newcomers into Québec society. There were two proficiency levels: intermediate and advanced. Most students completed the test in 30 to 35 minutes. Once the tests were scored, facility and discrimination indices were calculated (following Fulcher [2010]) to explore the measurement characteristics of the test clusters and the word-definition matching items within them.

The facility index (FI) is the proportion of test-takers who answer an item correctly. The FIs for the matching items ranged from as low as 0.13 obtained for the 10K item *moisi* to as high as 1.00 obtained for the 2K item *hiver*. Items known to all (such as *hiver*) were obvious candidates for discarding. Mean FIs for the 2K, 3K, 5K, and 10K sections of the pilot test (based on the 10 best-functioning clusters in each section) are shown in the third column of Table 1. The figures indicate that the test is working as intended; a large proportion (over 80%) of the test-takers were able to answer the items on the 2K section correctly, and the mean FI scores decrease as the test items become more infrequent. However, we were surprised to see that the mean FI for the 10K items amounted to .55, which indicates that more than half of the test-takers were able to match correct definitions to these supposedly difficult words. In Schmitt et al.'s (2001) validation of the VLT, the mean FIs for the two 10K sections they tested were much lower (.30 and .29). Closer inspection of the words targeted in the 10K section of the TTV, which were drawn from Baudot's (1992) list, revealed that four of them were not as infrequent as might have been expected. For instance, while Baudot lists *pêcheur* as a 10K word, Lonsdale and Le Bras (2009) list it as a 3K word. In view of this discrepancy, the 10K section of the pilot test was discarded and an entirely new set of 13 clusters was created, still using the Baudot list but with careful

Table 1: Facility values and discrimination indices for pilot version ($N = 63$)

Section	Number of items	Item facility		Discrimination index	
		M	*SD*	*M*	*SD*
2K	30	0.82	0.09	0.32	0.08
3K	30	0.72	0.15	0.47	0.11
5K	30	0.65	0.14	0.46	0.10
10K	30	0.55	0.14	0.55	0.08

checking against the newer Lonsdale and Le Bras lists to avoid including overly frequent cases like *pêcheur*. It was not possible to pilot test this new 10K section with a learner group, but two native speakers of French assisted in identifying 10 clusters they deemed to be both well written and challenging for inclusion in the final version of the test. The FI for this new 10K section was recalculated later on the basis of performance in the main study and was found to be .32 (SD = .15). This more plausible figure is in line with the FIs of .30 and .29 found by Schmitt et al. (2001) for the 10K sections of the VLT.

The discrimination index (DI) gives a picture of how well an item discriminates between the top scorers and the bottom ones; in our study, we compared the performance of the top third of the pilot group to the bottom third. The DI for a particular test item is obtained by subtracting the FI of the bottom scorers from the FI of the top scorers (Fulcher, 2010). An extremely easy or extremely difficult test question will have a low DI, since test-takers in both the high and low groups can be expected to perform similarly. According to Fulcher (2010), a test item is discriminating well enough if the DI is .30 or above. The DIs for matching items on the pilot version of the TTV ranged from as low as 0.00, obtained for the 2K item *faim* (a clear candidate for discarding), to as high as 0.90, obtained for the 10K item *fragmentaire*. The mean DI figures for each frequency section (based on the 10 best performing clusters) are shown in the fifth column of Table 1. The DI for the revised 10K section, based on performance in the main study (reported below), was .44 (*SD* = .216). These figures, which are all above the .30 criterion, indicate that the items are discriminating well. However, the means hide the characteristics of clusters, each of which contains three matching items. Discrimination indices for clusters range from .13 to .77.

The elimination of clusters with the weakest measurement characteristics resulted in the final version of the TTV, which is made up of four frequency sections (2K, 3K, 5K, and 10K) with 10 clusters in each

section. Each cluster tests knowledge of three words, for a total of 30 tested words per section and a test total of 120 items.

Interviews

Before moving to the larger study, we probed test-takers' knowledge of tested words in individual interviews. The procedure was intended to explore the extent to which learners actually knew words they had correctly matched to definitions on the TTV. Following a protocol used by Schmitt et al. (2001) in their validation study, a list of 48 tested words, 12 from each of the four frequency sections of the pilot test, was prepared for use in these interviews. In the interviews, which were conducted in French, the first author pointed to the first word on the list and asked: "Can you tell me what this word means?" If the participant was not able to answer the question orally (i.e., he or she could not come up with an acceptable synonym or definition), the participant was given a card with the test word and five answer options (one of which is the correct definition as presented on the TTV). Box 5 shows the card prepared for the word *remporter*. Because the multiple-choice card presents a single word and five definition options, the format differs considerably from the TTV, where answering an item involves reading a single definition and considering six words as possible answer options.

Box 5. Sample card used to confirm learners' knowledge of the words tested on the TTV

22 remporter	a. ne pas voir
	b. faire arrêter
	c. gagner un jeu
	d. render plus pauvre
	e. connaitre la valeur

Twelve volunteer participants (a subset of the pilot test group) took part in the 30-minute validation interviews, which took place the day after the pilot test was administered. Five were intermediate-level learners; seven were advanced. The interviews elicited 576 answers (12 participants x 48 test words = 576). In Table 2 the pattern of responses is shown as a matrix. Assessing the extent to which responses on the TTV were a "true" reflection of interviewees' knowledge of the test targets involved counting numbers of matches and mismatches across measurement techniques. In 68% of the cases (390 of 576

Table 2: Comparison of interview results with TTV results

			TTV responses				
			Correct			Incorrect	
Interview	Knew	a	390	(68%)	b	60 (10%)	450
	Did not know	c	46	(8%)	d	80 (14%)	126
			436			140	576

responses), test-takers were able to answer an item correctly on the TTV and also produce (or identify) a correct definition of the target word in the interview (scenario a in Table 2). These "matches" are indicators of validity; arriving at a correct answer on the TTV appears to be based on knowing the word's meaning rather than on mere guesswork. By the same logic, cases of incorrect responses in both formats also serve as indicators of validity; the tested words are really not known. Cases where words were not known on both the TTV and in the interviews (scenario d) amounted to 14% of the total (80 responses). Taken together with the positive results (68 + 14 = 82%), there appears to be considerable congruence in participants' performance on both the test and the interviews. But there were also mismatches, and these threaten the validity of the TTV. One type of mismatch involves the student answering the item correctly on the TTV but not in the interview (scenario c), possibly as a lucky guess. There were 46 instances of this, amounting to 8% of the total. Another kind of mismatch occurs when the learner explains a word correctly in the interview, but did not respond correctly on the TTV (scenario b). There were 60 cases of this, amounting to 10% of the total.

The numbers of cases in all four categories were used to calculate the correlation between performance on the written test and in the interview. The Phi coefficient amounted to .48 (*Phi*, $p < .0001$), which is moderate but not high. Circumstances of the pilot-testing administration may explain why there were more mismatches than expected. One factor that may have increased amounts of guesswork was the students' impression that they needed to answer all of the questions on the test. To reduce the role of guesswork later in the main study, we encouraged test-takers to leave any unknown questions blank; written instructions to this effect were added to the final version of the test and emphasized orally by the test administrator. The larger than expected number of cases where a TTV response was not correct but the student was able to provide a correct meaning in the interview may be explained by the timing of the interviews. These occurred on the day after the TTV was administered, which meant that interview

volunteers had the opportunity to discuss the target words and their meanings or perhaps look them up before the interviews. In summary, it appears that the TTV was able to tap a substantial proportion of the word knowledge that participants actually possessed (82%), but given the problems described, the evidence is not as strong as it might have been.

With this part of the validation process completed and the final version of the test in place, we proceeded to explore its effectiveness with a larger, more diverse population of learners of French.

Testing the final version

Participants

The participants in the main study were 175 adult immigrant learners of French in intensive *francisation* courses at the same Montreal school where the pilot testing had taken place. The 115 females and 60 males had been assigned to one of four proficiency groups (beginning, low-intermediate, upper-intermediate or advanced) based on performance on a four-part placement test that assesses comprehension and production in both written and oral modes. At the time of testing, beginners had spent 330 hours in class, while low-intermediates, upper-intermediates, and advanced learners had had 660, 990, and 1,320 hours of instruction, respectively. Schmitt et al. (2001) emphasize the importance of access to a linguistically and culturally diverse population when investigating this type of vocabulary test. With 39 countries and 21 languages represented in the sample, the participant group clearly met that criterion: L1 backgrounds included Spanish (38), Farsi (26), Romanian (26), Mandarin (23), Russian (20), Arabic (15), Tagalog (9), Portuguese (3), Ukrainian (2), Vietnamese (2), Amharic (1), Bangla (1), Berber (1), Bulgarian (1), English (1), Hungarian (1), Korean (1), Kyrgyz (1), Nepali (1), Tamil (1), and Teochew (1).

Procedures and data analysis

The revised TTV was administered to the participant groups on two consecutive days. Most students completed the test well within the 50-minute time slot.

The first research question pertains to frequency effects. Our hypothesis predicts that learners will know more frequent words than infrequent ones. That is, if the TTV is a valid measure, scores will be high on the 2K section, but lower on the 3K section, and lower still on the 5K and 10K sections. Determining whether this pattern occurred involved scoring the papers and calculating mean scores for the whole participant group for each of the four frequency sections of the test. These section means were then tested for between-group differences,

using a one-way ANOVA. Cronbach's alpha values were determined for each of the four test sections; these indicate internal reliability. We also calculated each participant's overall vocabulary size. This involved converting participants' test performance on each of the four sections (2K, 3K, 5K, and 10K) into percentages. These percentages were then applied to the number of words sampled in each section. In the case where a section of the test samples a 1,000-lemma list, the calculation is straightforward. For example, if a participant answers 90% of the questions correctly, he or she is assumed to know 90% of the lemmas on that list, i.e., 900 lemmas. However, only the 3K section of the test samples a list of 1,000 lemmas. Two sections sample a list containing 2,000 lemmas (2K and 5K), and the last section (10K) samples a list of 5,000 lemmas. Thus the calculation involved applying percentages to 2,000, 2,000, and 5,000 lemmas, respectively, for those parts of the test. Once a participant's figures were obtained for all four parts of the test, they were totalled to arrive at an estimation of his or her overall vocabulary size.

The second question pertains to the TTV's ability to reflect learners' proficiency level. If the test functions as expected, students in the higher proficiency groups will have higher scores (and larger vocabulary sizes) than students in lower groups. Answering this question involved scoring the papers and calculating mean scores on the test as a whole and for each of the four proficiency groups (intact classes). These scores were tested for between-group differences, again using a one-way ANOVA.

Results

The first research question addressed performance of the four frequency sections of the TTV. Means in the entire participant group (N = 175) for each frequency section of the test (maximum score per section = 30) are shown in Table 3. The figures show the expected pattern, with the highest mean score of 20.72 (SD = 6.59) on the section that tested the most frequent (2K) words, and lower scores on sections that tested less frequent words. According to the results of a one-way ANOVA, there were significant differences in the data, $F(3, 699) = 422.82$, $p < .0001$, and post hoc pairwise comparisons showed that all of the differences between means were significant ($p < .01$). As figures in the rightmost column show, the learners as a group know more than two-thirds of the words on the 2K section (69%), but a little less than a third of those on the 10K section (32%). The declining scores across the word sections clearly indicate that the TTV provides a scalable profile of vocabulary frequency levels.

Table 3: Mean scores by frequency section ($N = 175$)

Section	Maximum score	*M*	*SD*	%
2K	30	20.72	6.59	69
3K	30	18.25	7.53	61
5K	30	16.25	7.82	54
10K	30	9.58	6.00	32

Table 4: Mean scores by proficiency level ($N = 175$)

Proficiency level	Maximum score	*M*	*SD*	%	Extrapolation to 10K
Beginning	120	38.87	20.83	32	2699
Low-inter	120	56.29	22.39	47	4068
High-inter	120	73.88	29.50	62	5274
Advanced	120	92.44	13.50	77	6891

Cronbach alpha values for the 2K, 3K, 5K, and 10K sections were .900, .922, .923, and .879, respectively. These figures show that the internal reliability of each of the frequency sections was satisfactory (i.e., near .90 or above) in all four sections. These figures are comparable to those reported by Schmitt et al. (2001) for the VLT.

The second research question addressed the TTV's ability to discriminate between learners of differing levels of proficiency. If the test is functioning as intended, students in the more advanced groups should have higher scores on the test as a whole (maximum total score = 120) than students in lower groups. Results for the four proficiency groups are shown in Table 4. The means in the third column reveal the expected pattern: the group mean for the beginners is the lowest, at 38.87 (SD = 20.83); as proficiency level increases, so do the means, with the highest mean of 92.44 (SD = 13.50) obtained in the advanced group. According to the results of a one-way ANOVA, there were significant differences in the data, $F(3, 174) = 40.97$, $p < .0001$. Post hoc pairwise comparisons indicated that all of the between-group differences were significant ($p < .01$).

The rightmost column in Table 4 shows the mean vocabulary sizes in the various groups. These are plausible figures, although they are higher than those reported by Milton (2008) in his study of British school learners using the checklist test. We return to this point in the Discussion section.

Discussion

The goal of this study was to develop and validate the TTV, a new measure of receptive vocabulary size for French L2 learners. The test is modelled on the VLT for English by Nation (1983) and revised by Schmitt et al. (2001). It samples words from frequency lists derived from a corpus of international spoken and written French by Lonsdale and Le Bras (2009) and assesses knowledge of 120 words in total, 30 from each of four frequency levels (2K, 3K, 5K, and 10K). Results of validation interviews showed that the interview verifications matched test performance in over 80% of cases. When the TTV in its piloted and improved form was administered to 175 learners of French at four levels of proficiency at a school in Montreal, it functioned as expected: The sections that tested less frequent words proved more difficult than sections with more frequent words; means for the four frequency sections differed significantly. Since research has shown that learners generally acquire more frequent words before they acquire less frequent ones (Milton, 2009; Nation, 1990), the frequency findings speak to the validity of the test. Performance of individuals did not always follow this neatly descending pattern, however. For example, several learners in the advanced group scored higher on the 3K and 5K sections than on the 2K section. Milton notes similar results in his 2009 overview of vocabulary size testing. One possible explanation for this finding comes from Milton's 2007 study, which compared learners with "normal" profiles to those with a 2K deficit and found evidence of an aptitude effect. More research of this type is needed to understand how individual learners respond to frequency in the input to which they are exposed. Also, as Bardel et al. (2012) have found, exposure to thematic classroom vocabulary and the availability of L1 cognates can facilitate the learning of infrequent words; these factors may have been in play here.

The testing also identified proficiency differences in the expected direction: the higher the proficiency level of the group, the greater the mean scores on the test. These differences were statistically significant. The finding that greater vocabulary sizes were associated with more advanced proficiency (as determined by the school's placement measure) lends credibility to the test and points to its potential usefulness in helping to place students in language courses.

How do size findings identified using the TTV compare to those of other studies? An important source of previous size estimations is Milton's 2006 cross-sectional study, which reports mean vocabulary sizes based on yes-no checklist test scores for L2 French learners in Britain. Fortunately for the purposes of comparison here, he also

reports estimated amounts of time spent in class. In terms of hours of instruction, the beginning Québec participants in the TTV study, who have completed 330 hours of study, can be seen as roughly comparable to British secondary learners in year 5, who have completed an estimated total 351 classroom hours (78 + 58.5 + 58.5 + 78 + 78), according to Milton's figures. As shown in the first row of Table 4, the mean vocabulary size for the Québec learners amounts to an estimated 2,699 words. This stands in marked contrast to the mean size of just 852 words reported for the British learners after a similar amount of time in class. Another comparison might be made between the low-intermediate Québec learners with 660 hours spent in class and the British learners, who are reported to have spent a total of 643.5 hours in class by the end of seven years of secondary school. Again, the difference is large. The mean size for the Québec learners shown in the second row of Table 4 is estimated at 4,068, while the British figure is 1,930 (with considerable variability in both groups). Arguably, these very great differences call the TTV's measurement capabilities into serious question.

But are the Québec figures wildly implausible? There are several reasons to think they are not. First, the TTV is designed to measure size through the 10K frequency level, while the maximum level assessed on the checklist test is 5K. This gives the Québec learners a considerable advantage in terms of opportunities to demonstrate word knowledge. Another explanation pertains to the frequency lists used to build the two size tests. The checklist test draws on the Baudot (1992) list, which is based on a corpus of written materials, but the TTV draws for the most part on work by Lonsdale and Le Bras (2009), whose corpus contains a large spoken component (50%). In other words, the character of the lists sampled to build the measures differs considerably, and it is possible that this makes the TTV an easier test. There are also important differences in exposure to target language input in the two learning contexts. In Milton's study, the participants were learning French as a foreign language at school while living in an English-speaking milieu. By contrast, the TTV participants live and work in a French-speaking society, and therefore they have a great deal more exposure to target language input. Acquiring proficiency in their new language promises social and economic benefits, so the Québec participants are likely to be motivated learners. There is also research evidence that intensive instruction leads to greater proficiency gains than does a distributed "drip feed" program (Serrano & Muñoz, 2007; White & Turner, 2005), which seems a fair characterization of the classroom situations investigated by Milton. By contrast, Québec *francisation* programs promote rapid integration into the

© *CMLR/RCLV*, doi:10.3138/cmlr.2820

Table 5: Mean correct scores for different language groups (maximum score = 30)

Section	Romance (*SD*) *N* = 67	Asian (*SD*) *N* = 27	Other (*SD*) *N* = 81
2K	25.40 (4.16)	17.85 (7.49)	17.80 (5.68)
3K	24.48 (4.53)	14.30 (7.19)	14.41 (6.06)
5K	21.94 (5.03)	11.78 (8.23)	13.04 (6.76)
10K	13.97 (4.53)	6.37 (5.30)	7.02 (5.11)
Total	85.79 (16.10)	50.30 (25.74)	52.27 (25.74)

French-speaking milieu and clearly qualify as intensive: all of the TTV participants spent at least 12 hours per week in class; most of them spent as many as 30. This may well have given them a vocabulary learning advantage over the British learners, who appear to have attended only two or three hours of class per week during most of their seven years of study. Finally, over a third of the students who took the TTV were speakers of Romance languages and were therefore probably able to recognize many words on the test due to familiarity with cognate equivalents or near-equivalents in their first languages.

To determine the extent to which the TTV might have advantaged participants with a Romance-language background, we divided the 175 participants into three rough first language groups: Romance language speakers, Asian language speakers, and speakers of other languages. The Romance group consisted of 67 speakers of Portuguese, Romanian, and Spanish. The Asian group consisted of 27 speakers of Korean, Mandarin, Teochew, and Vietnamese; these East Asian languages are typologically distant from French and have not been as strongly influenced by Latin as English has been, for instance. The "other" group consisted of 81 speakers of Farsi, Russian, Tagalog, and 11 other languages (see the Participants section above). The means in these three groups were calculated for each of the four frequency sections and for the test as a whole. As can be seen in the first row of Table 5, means on the 2K section were distinctly higher in the Romance group, at 25.40 (maximum score = 30), while the means in the two other groups were both lower, at around 17.8. This pattern is also seen in the other frequency sections, with Romance speakers outperforming the other two groups by substantial margins (and with more consistency, as the smaller standard deviations indicate). When means for total scores in the three groups were tested via a one-way ANOVA, significant differences were found. $F(2, 174) = 59.11$, $p < .0001$. Post hoc pairwise comparisons confirmed a statistically significant advantage for the Romance speakers over both of the non-Romance groups, but there was no statistically significant difference between the two non-Romance groups ($p < .01$). These results confirm

the expected cognate advantage for learners with L1 knowledge of the lexis of another Romance language. They also show that as a group, the Asian language speakers were not at a greater disadvantage than the speakers of other non-Romance languages.

Both the test and the validation study have several limitations. One design shortcoming of the TTV pertains to the sampling of test words from two different sources. The recent frequency list by Lonsdale and Le Bras (2009) is based on a much larger and more representative French corpus than earlier lists, and ideally, their work would have been used to create all four sections of the TTV. But since they list only the 5,000 most frequent French lemmas, we were able to use it to build only the 2K, 3K, and 5K sections, having to resort to the older list by Baudot (1992), which lists over 16,000 lemmas, to build the 10K section. But would a more recent and comprehensive French corpus such as the one by Lonsdale and Le Bras identify the test words we selected from Baudot as "true" 10K-level items? We are not presently able to answer the question. In piloting the test, however, we discovered that some test words classified by Baudot as 10K were actually fairly frequent according to Lonsdale and Le Bras. The problem items were replaced, and the results reported here testify to the overall quality of the revised test, but the extent to which performance on the 10K section accurately reflects learners' knowledge at this frequency level is difficult to verify. These problems highlight the urgent need for access to good French frequency lists extending beyond the 5K level. In the case of English, lists for 14 frequency levels based on the British National Corpus (BNC) have been available to researchers and teachers of English since 2006 (at Paul Nation's home page, http://www.victoria.ac.nz/lals/about/staff/paul-nation) and 25 lists integrating frequencies from both the BNC and the Corpus of Contemporary American (COCA) English are available there currently. Access to comparable information for a language as important as French is clearly overdue.

Another limitation of the Lonsdale and Le Bras (2009) frequency list (and, by implication, the TTV) pertains to the corpus upon which the list is based. Although we saw this list as the best available resource for developing the TTV, it may be less than ideal for pedagogical use, due to the fact that over 20% of the corpus consists of European and Canadian parliamentary debates (p. 3). This seems likely to have had an effect on the words and word uses that registered as frequent. For instance, we noticed that a rather unusual and formal term *clore* ("to close," as in *clore la session*, "close the session") ranked as a high-frequency lemma (2K). By contrast, *cahier* ("notebook"), a thematic word likely to be learned very early in the language classroom, was ranked at

5K. Designing a pedagogical list to reflect a more representative range of spoken French registers is another avenue for improvement. Bardel et al.'s (2012) development of frequency lists based on a corpus of spoken French is a promising step in this direction.

A third limitation was identified during the interviews. They revealed that two participants proved unable to match the definition *commerce* to the target word *trafic*. Each of them knew that *commerce* meant *magasins* (stores) or *affaires* (business), but when the researcher asked what *trafic* meant, they both answered (in French), "The circulation of cars." The word *trafic* is frequently used in this sense in Québec, but this cars-and-trucks definition did not appear as an answer option on the test. The format of the TTV (and VLT) presents a single main definition of a word (the most frequently used meaning in the corpus upon which the test is based). This clearly results in underestimations of learners' knowledge in cases such as *trafic*, where the interviewees knew a correct but untested meaning of a polysemous word. The example also reveals unexpected complexities in interpreting sources of knowledge. Here it is unclear whether the interviewees were misled by knowledge of the English word *traffic*, or by the English-influenced and characteristically Québec use of the French word *trafic*, or possibly by both.

Finally, we recognize shortcomings of the validation study itself. Our study is not as extensive as the study by Schmitt et al. (2001) that we used as a guide. They tested more students, more questions, and more variously ordered versions of their test than we were able to. In our study, time constraints at the school meant that we were able to pilot a maximum of 48 clusters, of which eight were eventually eliminated. In an ideal scenario, more questions would have been trialled and evaluated. It would also be helpful to test the TTV's usefulness with learners of French at lower and higher ends of the proficiency spectrum, and in learning contexts where French is being taught as a foreign language. There was only one native speaker of English in the participant group, which means that a substantial group of learners of French in Canada (and elsewhere) is underrepresented. We are also aware that it is important to test the test by comparing performance on the TTV to performance on another established vocabulary measure such as the checklist vocabulary size test for French by Meara and his colleagues (1990, 2003). Plans for this validation experiment are currently underway. As improved and expanded frequency lists for French become more available, we envision remodelling the entire TTV and eventually testing it with learners on a much larger scale.

 doi:10.3138/cmlr.2820

Conclusion

There is an imbalance in available corpus-based resources for vocabulary research and pedagogy in the case of L2 French, with a great deal more in the way of frequency lists, size tests, and learning activities available to those interested in English. One of the goals in creating the TTV was to help redress that imbalance by drawing on state-of-the-art corpus work in French to create an updated receptive vocabulary size measure and make it available to the teaching and research community. To this end, the TTV appears in its entirety at the testing link on Cobb's Lextutor website (www.lextutor.ca). The TTV is also intended as a complement to the existing checklist test for French that relies on self-report and assesses vocabulary size only as far as the 5K level. We see the TTV's use of a verifiable answer format and its ability to test word knowledge up to the 10K frequency level position as notable strengths. The study reported here provides initial evidence that the TTV is a viable instrument. Individual items work reasonably well with a high level of internal reliability; the test as a whole identifies plausible vocabulary profiles and distinguishes between different groups of learners. Though hardly perfect and with many future improvements still to come, the test is now ready for practical use. We hope it will be helpful to many.

Correspondence should be addressed to Roselene Batista. Email: roselene.ds.batista@gmail.com.

References

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production: Vocabulary acquisition and use in L2 French and Italian. *Studies in Second Language Acquisition*, *34*(2), 269–290. http://dx.doi.org/10.1017/S0272263112000058

Baudot, J. (1992). *Les fréquences d'utilisation des mots en français écrit contemporain*. Montréal: Presses de l'Université de Montréal.

Bogaards, P. (2000). Testing L2 vocabulary knowledge at a high level: The case of the *Euralex French Tests*. *Applied Linguistics*, *21*(4), 490–516. http://dx.doi.org/10.1093/applin/21.4.490

Cobb, T. (2000). Compleat lexical tutor. [Website]. Retrieved from http://www.lextutor.ca/

Cobb, T., & Horst, M. (2004). Is there room for an academic word list in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). Philadelphia: John Benjamins. http://dx.doi.org/10.1075/lllt.10.04cob.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. http://dx.doi.org/10.2307/3587951

© *CMLR/RCLV*, doi:10.3138/cmlr.2820

David, A. (2008). Vocabulary breadth in French L2 learners. *Language Learning Journal*, *36*(2), 167–180. http://dx.doi.org/10.1080/09571730802389991

Eyckmans, J., van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behaviour in Yes/No Vocabulary Tests. In H. Daller, M. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 59–76). Cambridge: Cambridge University Press.

Forsberg Lundell, F., & Lindqvist, C. (2014). Lexical aspects of very advanced L2 French. *The Canadian Modern Language Review*, *70*(1), 28–49. http://dx.doi.org/10.3138/cmlr.1598

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*(4), 341–363. http://dx.doi.org/10.1093/applin/11.4.341

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 191–208). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/lllt.10.14gre.

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French*. New York: Routledge.

Meara, P., & Jones, G. (1990). *Eurocentres Vocabulary Size Tests 10KA*. Zurich: Eurocentres Learning Service.

Meara, P., & Milton, J. (2003). *X_Lex, The Swansea Levels Test*. Newbury, UK: Express.

Milton, J. (2006). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, *16*(02), 187–205. http://dx.doi.org/10.1017/S0959269506002420

Milton, J. (2007). Lexical profile, learning styles and construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 45–58). Cambridge: Cambridge University Press.

Milton, J. (2008). French vocabulary breadth among learners in the British school and university system: Comparing knowledge over time. *Journal of French Language Studies*, *18*(3), 333–348. http://dx.doi.org/10.1017/S0959269508003487

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Nation, I.S.P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. http://dx.doi.org/10.3138/cmlr.63.1.59

Nation, I.S.P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

 doi:10.3138/cmlr.2820

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *Language Teaching*, *31*(7), 9–13.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511732942.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26–43. http://dx.doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System*, *35*(3), 305–321. http://dx.doi.org/10.1016/j.system.2007.02.001

Stæhr, L.S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*(2), 139–152. http://dx.doi.org/10.1080/09571730802389975

Tidball, F., & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 133–149). Cambridge: Cambridge University Press.

Verlinde, S., & Selva, T. (2001). Corpus-based versus intuition-based lexicography: Defining a word list for a French learners' dictionary. In *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 594–598). Lancaster University.

White, J., & Turner, C. (2005). Comparing children's oral ability in two ESL programs. *The Canadian Modern Language Review*, *61*(4), 491–517. http://dx.doi.org/10.3138/cmlr.61.4.491

© *CMLR/RCLV*, doi:10.3138/cmlr.2820