John Benjamins Publishing Company

Estimating second language productive vocabulary size

A Capture-Recapture approach

Joy Williams,[1] Norman Segalowitz,[1,2] and Tatsiana Leclair[1]
[1] Concordia University / [2] Queensland University of Technology

This study provides validity evidence for the Capture-Recapture (CR) method, borrowed from ecology, as a measure of second language (L2) productive vocabulary size (PVS). Two separate "captures" of productive vocabulary were taken using written word association tasks (WAT). At Time 1, 47 bilinguals provided at least 4 associates to each of 30 high-frequency stimulus words in English, their first language (L1), and in French, their L2. A few days later (Time 2), this procedure was repeated with a different set of stimulus words in each language. Since the WAT was used, both Lex30 and CR PVS scores were calculated in each language. Participants also completed an animacy judgment task assessing the speed and efficiency of lexical access.

Results indicated that, in both languages, CR and Lex30 scores were significantly positively correlated (evidence of convergent validity). CR scores were also significantly larger in the L1, and correlated significantly with the speed of lexical access in the L2 (evidence of construct validity). These results point to the validity of the technique for estimating relative L2 PVS. However, CR scores are not a direct indication of absolute vocabulary size. A discussion of the method's underlying assumptions and their implications for interpretation are provided.

Keywords: productive vocabulary, second language, capture, recapture, ecological, word association

This paper documents an attempt to assess second language (L2) productive vocabulary size using a novel approach recently advocated by Meara and Olmos Alcoy (2010). This approach is based on a Capture-Recapture (CR) methodology, borrowed from population ecology, that involves taking two samples (a capture and a recapture) from the population whose size is to be assessed and computing

The Mental Lexicon 9:1 (2014), **23–47.** DOI 10.1075/ml.9.1.02wil
ISSN 1871–1340 / E-ISSN 1871–1375

what is known as the Petersen Estimate (Petersen, 1896; Sutherland, 2006). The sampling technique and the Petersen Estimate have traditionally been used in ecological studies to estimate the number of animals of a given species that inhabit a certain area. Meara and Olmos Alcoy (2010) proposed capitalizing on the logic of this technique to estimate the number of words that "inhabit" a person's L2 mental lexicon. The overarching goal of the current work, then, is to gather evidence of the validity of this unconventional approach as an effective means of measuring second language productive vocabulary size.

A distinction is commonly made between two, positively correlated aspects of vocabulary knowledge — receptive or passive knowledge and productive or active knowledge (Laufer, 1998; Webb, 2008). Although there is no consensus on precisely what is meant by these terms, vocabulary researchers generally seem to accept that receptive vocabulary refers to those lexical items that an individual can recognize and understand when listening to speech or reading text, whereas productive vocabulary, which is of interest here, refers to items that an individual can produce accurately when speaking or writing (Milton, 2009; Schmitt, 2010). Additionally, research suggests that receptive vocabulary knowledge develops before and at a faster rate than productive vocabulary, is larger than productive vocabulary and, importantly, is easier and more straightforward to measure or quantify than its productive counterpart (Fitzpatrick, 2003; Laufer, 1998; Laufer & Paribakht, 1998; Milton, 2009; Schmitt, 2010; Webb, 2008; Zimmerman, 2004).

However, despite the prevalence of this distinction and the commonness of these findings, the nature of receptive and productive vocabulary knowledge, and the relationship between them, are not entirely clear (Milton, 2009; Schmitt, 2010). Meara (1997) suggests that receptive and productive abilities may be separate, qualitatively different aspects of word knowledge that differ in the degree to which they are connected in a lexical network. In this view, productive lexical items are those that can be activated from within the network itself, through their many multidirectional links to other lexical items, while receptive items require external stimuli for activation, i.e., encountering the word while reading text or listening to speech. For a receptively known item to gain productive status, new associational links must be made from the rest of items in the lexicon to the receptive vocabulary item in question. Such a view does not imply a natural progression from receptive to productive abilities and maintains that receptive and productive abilities represent two different patterns of connectivity within the lexicon. Melka (1997) acknowledges the possible practical value, for second language pedagogy, of the distinction between receptive and productive vocabulary. However, in her review of the challenges of measuring receptive/productive differences, she comes to the conclusion that it may be best to avoid the distinction altogether because it is "too fuzzy" (p. 99) since any supposed receptive/productive vocabulary

boundary will move as a function of many linguistic and extra-linguistic considerations. She advocates, therefore, viewing the distinction between receptive and productive vocabulary knowledge as reflecting degrees of (or a continuum of) vocabulary knowledge and familiarity, not different vocabulary systems (Melka, 1997).

Regardless of the approach one takes, however, it is clear that no generally accepted boundary or criterion has, as yet, been empirically established to definitively distinguish a word that has receptive status from one that has productive status, nor is there an established threshold at which receptive knowledge becomes productive (Read, 2000; Schmitt, 2010). Furthermore, although competent passive skills in a language often require the listener or reader to display active abilities, in the sense of predicting or anticipating the words that will follow (Milton, 2009), the relation between the two may not necessarily be linear, since acquisition of relatively passive knowledge of vocabulary items does not necessarily imply that learners have productive skill with these items. Thus, receptive vocabulary size cannot be used reliably as an absolute indication of productive vocabulary size. The task of researchers interested in quantifying productive vocabulary size, then, is to design valid ways of measuring this aspect of vocabulary knowledge in its own right. We attempt to do so in the current work by using a novel approach to quantifying productive vocabulary, operationally defined here as those lexical items that participants are able to produce, in written form, while completing a word association task.

Challenges in Measuring Productive Vocabulary

The challenges associated with measuring productive vocabulary size are many. For instance, widely used measures of L2 productive vocabulary knowledge are often time consuming, too controlled and context-dependent. These tests are further limited by the fact that they tend to assess pre-selected targets, restrict test-takers to the production of one correct response, assess receptive abilities concurrently, and elicit insufficient quantities of content vocabulary from which to make meaningful inferences (Clenton, 2008; Fitzpatrick & Clenton, 2010; Meara & Fitzpatrick, 2000; Milton, 2009; Read, 2000). For example, Laufer and Nation's (1995) Lexical Frequency Profile (LFP), designed as an index of lexical richness, requires learners to write short texts, which are then analyzed by the computer program, *VocabProfile* (Cobb, n.d.). This program generates a lexical frequency profile for each learner by computing the proportion of word families in the following categories: the first 1000 most frequent words (the 1K band), the 2K band, the University Word List (UWL), and off-list items (Laufer & Nation,

1995). In addition to being potentially time consuming (e.g., Laufer & Nation, 1995 gave their participants one hour to complete each of two compositions), the LFP requires the production of written context-specific texts, which contain a large proportion of high frequency (function) words and which do not require learners to use vocabulary that is representative of the range of items in their lexicon (Fitzpatrick & Clenton, 2010; Meara & Fitzpatrick, 2000).

The problem of context dependence is also associated with the Productive Vocabulary Levels Test (PVLT; Laufer & Nation, 1999). This test requires learners to read a sentence and complete it by supplying the missing word. The first letters of the target word are provided to rule out other non-tested, but semantically viable options. The PVLT samples 18 items from each of the 2K, 3K, 5K, and 10K bands and the UWL, and scores for the number of correct items at each word level, and overall, are calculated. Aside from the fact that production is limited to only one correct answer on pre-determined test items, it may not be entirely valid to make inferences about productive vocabulary knowledge as a whole from 18 pre-selected items from five frequency bands. Additionally, the test provides as many initial letters as necessary to effectively disambiguate the target, which means that, at times, most of the word stem is available to test-takers (Read, 2000). This can create considerable variability in the degree of word knowledge, and reliance on contextual information, required to succeed on various items (Read, 2000). It may not, then, be possible to draw conclusions that are specific to productive knowledge since receptive abilities are also required, both to consider the context of the sentence and to make use of the initial letters provided (Fitzpatrick & Clenton, 2010; Read, 2000).

The Capture-Recapture Methodology and Petersen Estimate

Since the nature of what is being measured by these vocabulary tests remains unclear and the results gathered from them are difficult to interpret, Meara and Olmos Alcoy (2010) advocated investigating the construct of productive vocabulary from "different, perhaps unconventional points of view" (p. 223). Along those lines, they examined whether the Capture-Recapture methodology (CR), which is commonly used in population ecology studies to reliably and accurately estimate the size of animal populations in a given area, could be applied to estimate the size of L2 productive vocabulary.

In explaining the logic of the CR methodology, Meara and Olmos Alcoy (2010) provide the example of an ecologist interested in estimating how many fish of a given species live in a river. In order to arrive at such an estimate, the ecologist first selects a section of river that is representative of the conditions that

exist in the entire river and which provides a good chance of sampling the fish of interest. Next, he captures his first sample of fish (Time 1) by using a suitable trapping technique, such as casting a wide net in the chosen section of river. All of the fish captured at Time 1 are counted, marked for easy identification should they return in future captures, and then released to continue moving naturally in the river. After enough time has passed for the population of fish to redistribute itself evenly in the river, the ecologist takes the second sample of fish (Time 2) using the same method as at Time 1. The researcher then makes a count of the total number of fish captured at Time 2, along with a count of the number of marked fish from Time 1, which have been recaptured at Time 2. To summarize, this capture-recapture methodology provides three values: the number of fish captured at Time 1 (x), the number of fish captured at Time 2 (y), and the number of 'repeat' fish (r), i.e., marked fish that were captured at Time 1 and recaptured at Time 2. The estimate of the total population of fish in the river (P) is then calculated by plugging these three values into a formula known as the Petersen Estimate ($P = xy / r$; Petersen, 1896). The basic logic of this formula is that the ratio of r to y should be the same as the ratio of x to P, the unknown whole population size.

In order to explore whether this ecological approach could be applied to the estimation of L2 productive vocabulary size, Meara and Olmos Alcoy (2010) recruited 24 native speakers of English, who were intermediate (n = 11) and advanced (n = 13) learners of Spanish. The trapping procedure used was a single 30-minute writing task in which participants wrote short descriptions of a six-picture cartoon story. This procedure was completed two times, one week apart, each time with the same cartoon story. The data were then transcribed, spelling errors were corrected, grammatical errors ignored, and a computer program calculated the number of word tokens and types in each text. The Petersen Estimate was computed based on the number of word types in the two texts, and a Mann-Whitney U test confirmed that the Petersen estimate of productive vocabulary size reliably distinguished between the intermediate and advanced groups (93.81 vs. 160.37 word types). Meara and Olmos Alcoy (2010) also concluded that the Petersen estimate is able to detect knowledge of more vocabulary items than actually present in the texts since the estimate is far larger than the raw type counts in the first and second narratives.

While these preliminary results suggest that the CR methodology and the Petersen Estimate hold some promise as a measure of productive vocabulary size, Meara and Olmos Alcoy's (2010) trapping instrument may not have been ideal since the use of a picture description writing task may have violated some of the assumptions that need to be met when using the Petersen formula. For example, the assumption that the samples taken are representative of the population as a whole (Lindberg & Rexstad, 2002) may have been violated by the use of the

written picture description task, since the words used to describe the picture story cannot possibly be representative of the items in the lexicon as a whole. Additionally, the use of the story writing task may have also violated the assumption that all members of the population have an equal probability of being sampled (Lindberg & Rexstad, 2002), since the words necessary to describe the events depicted in the picture story have a greater chance of being captured and recaptured than do other items in the individuals' lexicon. Further, the Petersen's estimate may have been lowered simply because participants described the exact same picture story at Time 1 and Time 2. This likely inflated the number of 'repeat' items, which is the denominator in the Petersen formula. Indeed, repeats (i.e., words that were captured at both Time 1 and Time 2) appeared to be quite high in Meara and Olmos Alcoy's (2010) study since 45.50% and 49.31% of the word types produced at Time 2 by the Advanced and Intermediate groups, respectively, were also produced at Time 1. The use of this particular trapping method, therefore, may be responsible for perhaps the most obvious drawback of Meara and Olmos Alcoy's (2010) result, i.e., the fact that "the absolute figures are just ridiculously low, and clearly they cannot be interpreted at face value" (p. 231). By the estimates obtained in that study, the intermediate Spanish speakers have a productive vocabulary size of just over 90 words, while the advanced Spanish speakers have a productive vocabulary size of about 160 words, clearly underestimates of the true values.

Meara and Olmos Alcoy (2010) acknowledged these limitations and suggested that a more appropriate trapping procedure should elicit a fairly large number of words during both captures, without increasing the likelihood of words overlapping across captures. They speculated that the continuous word association format, used in the Lex30 test of productive vocabulary size (Meara & Fitzpatrick, 2000), might be more suitable. Not only are word association tasks relatively quick to construct, administer and score (Meara & Fitzpatrick, 2000; Wolter, 2002), but they also encourage fairly spontaneous production of mostly content words with minimal involvement of receptive skills and little, if any, restriction by context, since participants simply write down the words that come to mind in response to different stimulus words.

The Lex30

These benefits of the word association format have been exploited by the Lex30 test of productive vocabulary size (Fitzpatrick, 2003; Meara & Fitzpatrick, 2000). In this test, participants are given a series of 30 stimulus words that do not elicit stereotypical or highly frequent associates, and which are drawn from the first

1000 (1K) most frequent lemmas in Nation's (1984) word list. In keeping with the requirements of a continuous word association format, participants' task is simply to write down at least four words that come to mind in response to each stimulus word encountered. The data are then lemmatized and participants receive one point for each lemma located in Nation's (1984) 2K and beyond word frequency bands. More recent applications of the Lex30 have been constructed and scored using the JACET 8000 wordlist since it is more up-to-date than Nation's (1984) wordlist (JACET, 2003; Fitzpatrick & Clenton, 2010). Regardless of the frequency lists used for scoring, however, higher scores on the Lex30 indicate that an individual can produce a higher proportion of infrequent vocabulary items, which is assumed to indicate an overall larger lexicon (Fitzpatrick, 2003; Meara, 2009; Meara & Fitzpatrick, 2000). Since the Lex30 has been shown to be a reliable and valid index of productive vocabulary size (Fitzpatrick & Clenton, 2010; Fitzpatrick & Meara, 2004; Meara & Fitzpatrick, 2000; Walters, 2012), the Capture-Recapture (CR) methodology will be validated against this already established test.

The Current Work

The goal of the current work is to examine the validity of the CR technique as a measure of L2 productive vocabulary size. Instead of using written picture descriptions to elicit vocabulary, as Meara and Olmos Alcoy (2010) did, a word association task, set-up like the Lex30, was used as our trapping procedure. This allowed us to avoid many of the problems typically associated with measures of productive vocabulary size, as well as some of the methodological limitations of Meara and Olmos Alcoy's (2010) study. Additionally, the word association format allowed us to score the same data based on the logic of both the traditional Lex30 which focuses on the amount of low frequency words supplied, and the CR technique, which focuses on the amount of unique words supplied during the two captures. The results generated from these two scoring methods may then be more comparable since the type of data collected, and the way in which it was collected, is held constant. From this point forward, scores gained from the CR methodology will be termed 'CR scores'.

The proposed CR methodology will be held as valid if convergent and construct validity criteria are met. The convergent validity of a test is established when it correlates with an already validated measure of the same construct (Thorndike & Thorndike-Christ, 2010). Hypothesis 1, therefore, is that the CR and the Lex30 scores will be significantly positively correlated. Additionally, to show construct validity, a measure of productive vocabulary size should distinguish between the

L1, where vocabulary size is larger, and the L2. Thus, Hypothesis 2 is that the CR scores will be significantly larger in the L1 than in the L2. Furthermore, cognitive processing efficiency is a crucial component of fluency, which is also potentially influenced by vocabulary size and the connectivity of one's mental lexicon. Therefore, Hypothesis 3, which also relates to construct validity, is that a significant negative correlation will exist between L2 CR scores and performance on a semantic categorization task that assesses the speed and efficiency of L2 lexical access (Segalowitz, 2010), because speed and efficiency of lexical access should be a reflection of L2 experience and general L2 proficiency, and presumably, therefore, of vocabulary size. A negative correlation is predicted because speed is represented by reaction times in milliseconds (ms) and the efficiency of lexical access is represented by response time stability as reflected by the coefficient of variation (CV; defined as each individual's standard deviation of reaction time divided by that person's mean reaction time). For both of these variables, lower scores represent better performance.

Method

Participants

Participants were 47 English-French bilingual university students (30 females), ranging in age from 19 to 39 years, (M = 23.36, SD = 4.07), with varying degrees of proficiency in their L2. Inclusion criteria were that participants reported English to be their first and native language, with French as their second language, learned at least three years after English. All participants indicated that they have *fluent ability* in English speaking (M = 5, SD = 0) and listening (M = 5, SD = 0), on a 5-point scale ranging from 1 (no ability at all) to 5 (fluent ability), while ratings for English reading (M = 4.94, SD = .32) and writing (M = 4.87, SD = .40) ranged from *moderate* to *fluent ability*. L2 self-ratings of ability were as follows: speaking (M = 3.45, SD = .72), listening (M = 4.32, SD = .81), reading (M = 3.89, SD = .76), and writing (M = 3.09, SD = .88). A Wilcoxon signed-rank test confirmed significant differences between English and French self-ratings of abilities on all four language skills, indicating that participants were indeed more proficient in English, their L1: speaking: T = 0, Z = –5.93, p < .001; listening: T = 0, Z = –4.36, p < .001; reading: T = 1, Z = –5.42, p < .001; writing: T = 0, Z = –5.93, p < .001. Additionally, participants estimated that, on average, 80.26% (SD = 12.77) of their interactions with others occur in English, while only 19.52% (SD = 12.85) of interactions, occur in French. Participants received either course credit or $20 for their participation.

Materials

The Word Association Task

A paper-and-pencil continuous word association task was constructed in both English and French in a manner similar to the set-up of the Lex30 test (Meara & Fitzpatrick, 2000). Specifically, high frequency stimulus words were drawn randomly from within the 2000 most frequent words in English (Davies & Gardner, 2010) and French (Lonsdale & Le Bras, 2009). In contrast to the traditional Lex30, the English frequency list used for stimulus selection and test scoring was based on the 400-million-lemma Corpus of Contemporary American English (COCA) that fairly equally represents spoken texts as well as texts from fiction books, popular magazines, newspapers and academic journals (Davies & Gardner, 2010). The French frequency list used was based on a corpus of 23 million French words that equally represents spoken and written French language use (Lonsdale & Le Bras, 2009). Cross-linguistic homographs (e.g., *table*) and words that differ in the two languages based on only the positioning of one letter (e.g., *tender* in English and 'tendre' in French) were avoided as stimulus words.

Two lists of 30 words each were developed in each language, one for use at Time 1 and the other at Time 2 (counterbalanced across participants). The words were presented in the same fixed random order across participants. See Appendix A for a full list of the stimulus words used. In each language, separate testing booklets were created for use at Time 1 and Time 2, each containing two identical pages of 30 stimulus words. Participants used the second page of stimulus words only if they were able to supply more than the minimum number of associates requested to any of the stimulus words.

Living-Nonliving Task (LNL; Segalowitz, 2010)

The LNL is a computerized semantic classification task that measures the cognitive fluency of lexical access in English and French, that is, the speed and stability with which word meaning is processed. Following a brief training session in their native language (English), participants completed the main task in separate English and French testing blocks in counterbalanced order. A series of single words was presented one at a time in the center of a 12-inch computer screen and participants simply pressed the appropriate button on a controller to indicate whether the word referred to a living (e.g., *a dog*) or a nonliving thing (e.g., *a bed*). The stimulus words used were also drawn from the English (Davies & Gardner, 2010) and French (Lonsdale & Le Bras, 2009) frequency lists, but were different from those used in the word association task to avoid priming effects from the word association task. Each word was presented until a response was made or for a maximum time of 3000 ms, after which a new word appeared on the screen.

Participants were instructed to respond as quickly and as accurately as possible to each word and received audible feedback when an error was made. There were a total of 60 trials in both the English and French tasks, comprised of 12 warm-up trials and 48 experimental trials. Response times for correct trials were recorded and the coefficient of variability (CV), a measure of (within-subject) stability and efficiency of responses, was computed (CV = SD / RT). A low mean response time and CV coefficient indicate faster and more efficient responses on the LNL, which are interpreted as an indication of better cognitive fluency in lexical access.

Procedure

Participants completed two separate one-hour testing sessions, an average of 4.26 (*SD* = 2.56) days apart. At Time 1 (T1), participants completed the word association task first in their L1, English, followed by the task in their L2, French. They were given 15 minutes in each language to write down at least 4 associates to each of 30 high frequency stimulus words. However, ample space was available in the testing booklet in the event that participants were able to provide more than the minimum 4 associates to any of the stimulus words. Participants then completed the living-nonliving task in English and French, in counterbalanced order, by pressing the appropriate button to indicate whether the word in the center of the computer screen was a living or a non-living thing. This task took approximately 5 minutes in each language. Participants then filled out only half of a language background questionnaire (LBQ) to end the T1 testing session.

A few days later, at Time 2 (T2), participants completed the 15-minute word association task, again in English followed by French, each with a different set of 30 stimulus words. They also completed the other half of the LBQ to end the T2 testing session.

Data Analysis

Lemmatization

All associates provided in English were lemmatized according to the procedure outlined in Meara and Fitzpatrick (2000), which is based on Bauer and Nation's (1993) criteria for level 2 and 3 affixes. Words with the following affixes were treated as instances of their base lemmas: Level 2 (inflectional suffixes): plural, 3rd person singular present tense, past tense, past participle, comparative, *-ing,* superlative, possessive; Level 3 (most frequent and regular derivational affixes): *-able* (not when added to nouns; e.g., *teachable*), *-er, -ish, -less, -ly, -ness, -y* (adjectives from nouns; e.g., *fruity*), *non-,* and *un-*. Words with affixes that do not

appear in these lists were not lemmatized, and were treated as separate words (Meara & Fitzpatrick, 2000). It should be noted, however, that unlike in Meara and Fitzpatrick (2000), numbers were not included in lemma counts in the current work.

In the absence of information on the frequency of French affixes, equivalent French lemmatization rules were adapted from the English rules. As such, French plurals (*-s*, *-x*), third person singular present tense, past tense (*passé composé*, *imparfait*), and *-ing* form (*-ant*) were all lemmatized. Other French affixes that were lemmatized include *-able* (when added to verbs, e.g., *habitable* to *habiter*), *-eur* (e.g., *travailleur* to *travailler*), -âtre (e.g., *rougeâtre* to *rouge*), *-ment* (e.g., *doucement* to *doux*), and those affixes that form negatives or opposites (*in-, im-, mal-, dé(s), il-, non-*) in French. All feminine forms were converted to the masculine form.

Commonly used abbreviations were converted to their long forms, e.g., *tv* to *television*, *bday* to *birthday*, and ideas that were expressed using multiple words, were broken down into separate items, e.g., *wood panel* was treated as *wood* and *panel* and counted separately. In both French and English, proper nouns, function words, acronyms and onomatopoeia were excluded from the T1 and T2 counts and from all analyses.

Scoring

The data gathered from the word association task was scored based on the logic of the traditional Lex30 and the CR technique. Scoring based on the logic of the Lex30 rewards participants for each infrequent lemma provided. As such, one point was assigned to each English and French word located beyond the bands from which the stimulus words were selected, i.e., beyond the 2K word frequency band in English, as given in Davies and Gardner (2010), and in French, as given in Lonsdale and Le Bras (2009). For CR scoring, the number of unique lemmas at T1 (x) and T2 (y) were recorded, along with the number of lemmas common to both captures (r). The Petersen Estimate formula (xy / r) was then applied to the data to give an estimate of productive vocabulary size, i.e., a CR score.

Results

Preliminary Analyses

Since the word association and semantic classification tasks were completed in the L1 and L2, we were able to use residualized L2 scores in our analyses. That is, in order to statistically control for L1 performance and other nuisance variables

(individual differences related to writing speed, overall ability to generate associates, etc.), all L2 scores were residualized, i.e., regressed against their equivalent L1 score. These residualized scores give a purer indication of L2 vocabulary size and efficiency (Segalowitz, 2010). As such, wherever possible, results based on residualized L2 scores are reported. Additionally, non-parametric tests were used to analyze data that were not normally distributed and, as convention dictates, medians, rather than means, are reported with these results. Effect sizes, in the form of Pearson *r*, are also reported.

Descriptive statistics of the number of lemmas generated at Time 1 and 2 in English and French are included in Table 1. This Table suggests that the word association task encouraged participants to access a range of items in their mental lexicon since only 18.15% and 17.83% of the words supplied at Time 1 in English and French, respectively, were also supplied at Time 2 in response to different stimulus items. Additionally, the word association format itself is capable of distinguishing between languages since, at both times, participants supplied more lemmas in their L1 than in their L2, as seen in Table 1. The non-parametric Wilcoxon signed-rank test confirmed that significantly more lemmas were generated in English than in French at Time 1, $T = 1$, $Z = -5.94$, $p < .001$, $r = -.61$, and at Time 2, $T = 0$, $Z = -5.97$, $p < .001$, $r = .62$. Table 1 also suggests that, in both the L1 and L2, participants generated more lemmas at Time 2 than at Time 1. Analyses indicated that the number of lemmas supplied at Time 2 was significantly higher than at Time 1 in English, $T = 9$, $Z = -4.58$, $p < .001$, $r = -.47$ and in French,[1] $t(46) = -4.31$, $p < .001$, $r = .54$, respectively. As such, we chose to report L1 and L2 Lex30 scores based on performance at Time 2, under the assumption that producing more lemmas may increase the likelihood of scoring highly on the Lex30.

Table 1. Descriptive Statistics for the Raw Lemma counts, Repeats, Reaction Times and CV scores in English (first language) and French (second language).

Variables	English (L1)			French (L2)		
	Mdn	M	SD	Mdn	M	SD
Raw Lemmas-T1	138	141.72	26.28	100	101.45	29.76
Raw Lemmas-T2[a]	150	158.17	32.00	110	111.85	32.63
Repeats	25	25.72	10.11	17	18.09	7.43
Speed (RT)[b]	646	667.00	78.41	701	728.00	98.68
Efficiency (CV)[b]	.19	.20	.07	.19	.20	.06

[a] These values representing the average number of lemmas supplied at Time 2 include the repeat items. When those repeats are removed from the Time 2 lemma count, the average becomes 132.45 (SD = 28.55) in English, and 93.77 (SD = 29.37) in French.

[b] Unresidualized values for the French speed and efficiency are reported. The means of the standardized residuals are zero, and the standard deviation for both of these variables is .99.

Speed and Efficiency of Lexical Access

Reaction times on the LNL task in English were compared to their French equivalents to determine whether the expected pattern of results (lower (faster) RTs in the L1) would be found. A dependent *t*-test revealed that RTs were indeed lower in the L1 (M = 667, SD = 78.41), relative to the L2 (M = 728, SD = 98.68), $t(46) = -5.90$, $r = .66$, indicating that participants were faster at making lexical decisions about words in their L1. Additionally, in both English and French, correlations between the speed (RT) and efficiency (CV) of lexical access were examined. As expected, we found significant positive correlations between the RTs and CV scores in English ($r_s = .57$, $p < .001$) and between the residualized RTs and CV scores in French ($r_s = .32$, $p = .03$), indicating that those who responded faster were also more efficient responders with less noise and instability in their cognitive processing (Segalowitz, 2010).

Testing the Hypotheses

Hypothesis 1

It was hypothesized that the CR vocabulary size estimate would be positively correlated with Lex30 scores, as an indication of the CR's convergent validity. In support of Hypothesis 1, Spearman correlations (r_s) revealed that, in English, the CR estimate of vocabulary size was significantly positively correlated with Lex30 scores ($r_s = .66$, $p < .001$), and in French, residualized CR scores were also significantly positively correlated with residualized Lex30 scores ($r_s = .66$, $p < .001$).

Hypothesis 2

Descriptive statistics for the vocabulary variables of interest are presented in Table 2.

In testing the construct validity of the CR technique, it was hypothesized that CR scores would distinguish between participants' L1 and L2. Only unresidualized French scores were used in these analyses since residualized scores cannot be compared with unresidualized scores (N.B., the L1 vocabulary scores cannot be

Table 2. Descriptive Statistics of the CR and Lex30 Vocabulary Size Estimates.

Variables	English (L1)		French (L2)	
	M	SD	M	SD
CR	979.79	419.52	708.69	400.99
Lex30	52.06	20.63	48.66	20.03

Note. N = 47.

residualized because there is no appropriate baseline for them). As can be seen in Table 2, participants' average CR scores were indeed higher in English than they were in French, and the Wilcoxon signed-rank test confirmed that this L1 – L2 difference in median CR scores (889.97 and 606.67 respectively) was significant, $T = 12$, $Z = -3.79$, $p < .001$, $r = -.39$.

We also examined whether the Lex30 scores would distinguish between L1 and L2. Results indicate that Lex30 scores were unable to distinguish between participants' L1 and L2. The Wilcoxon signed-rank test revealed a non-significant difference between the English (*Mdn* = 43) and unresidualized French (*Mdn* = 45) Lex30 scores, $T = 22$, $Z = -1.06$, $p = .30$, $r = -.11$.

Hypothesis 3

As an additional test of the CR's construct validity, it was hypothesized that L2 CR scores would be significantly negatively correlated with two aspects of cognitive fluency, i.e., the speed (reaction times on the LNL task) and efficiency (CV scores) of L2 lexical access. In the residualized French data, the expected negative correlations were observed between the CR scores and performance on the LNL task. Specifically, residualized CR scores were found to correlate significantly, and in the expected negative direction, with residualized RTs on the LNL task ($r_s = -.44$, $p = .002$), but not with CV scores ($r_s = -.16$, $p = .28$). Although not specifically hypothesized, we also examined the relationship between the L1 English CR scores and cognitive fluency of L1 lexical access. Spearman correlations revealed that, in English, unresidualized CR scores were not correlated with speed ($r_s = -.25$, $p = .09$) or efficiency ($r_s = -.001$, $p = .86$) of lexical access.

We also examined whether Lex30 scores would correlate negatively with speed and efficiency of lexical access. Spearman correlations of the residualized French data revealed that Lex30 scores were significantly negatively correlated with the speed ($r = -.48$, $p = .001$) of lexical access, but not with efficiency ($r = -.06$, $p = .70$). Additionally, Spearman's rho indicated that the English Lex30 scores were not correlated with RTs on the LNL task ($r = -.24$, $p = .10$) nor with CV scores ($r = -.05$, $p = .73$).

Discussion

Discussion of the results of this study will focus on three main questions: (1) Is the CR a valid measure of productive vocabulary size? (2) Is the word association task an appropriate trapping procedure? (3) Does the CR estimate give information beyond that which is available in the raw lemma counts?

Is the CR a Valid Measure of L2 Productive Vocabulary Size?

The goal of the current work was to investigate the validity of the CR measure of L2 productive vocabulary size. Evidence for the three validity criteria was observed in the current work. First, the CR's convergent validity was confirmed by significant positive correlations (.66 in English, and .66 in French) between the CR vocabulary size estimate and scores on the validated Lex30 test of the same construct. This result indicates that individuals who have access to a greater number of words in their lexicon, as measured by the CR estimate, also have access to a greater number of infrequent words, as indicated by the Lex30 scores.[2] However, the magnitude of the relation between these two measures of productive vocabulary size suggests that they may be giving different, but complementary, information about productive vocabulary knowledge, namely about the quantity (CR) and quality (Lex30) of the lexicon. Additionally, the correlation coefficients for the relation between the CR and Lex30 are comparable to those of past studies that have also sought convergent validity evidence for new measures of productive vocabulary size. Laufer and Nation's (1995) LFP, for example, was validated against the active version of Nation's (1984) Vocabulary Level's Test, a precursor to the PVLT, and correlation coefficients reported ranged from .6 to .8 (with p values below .001), and the Lex30's convergent validity was established with correlation coefficients of .50 ($p < .01$) and .65 ($p < .01$) with the PVLT (Laufer & Nation, 1999) and a translation test, respectively. In light of these considerations, the correlations observed between the CR and the Lex30 were deemed sufficient to establish the convergent validity of the proposed method.

Secondly, the CR's construct validity was confirmed by the finding that the estimates of productive vocabulary size generated by the CR methodology successfully distinguished between the L1, where vocabulary size is larger, and the L2, in a within-subject design. Interestingly, the Lex30 test was unable to do so, as evidenced by a non-significant difference between Lex30 scores in English and French. Perhaps the number of infrequent words an individual is capable of supplying is better suited for capturing differences in productive vocabulary size when the speaker groups are distinct from each other, such as between native speakers and language learners (Fitzpatrick & Meara, 2004), or individuals with widely different amounts of experience in their second language (Walters, 2012). Capturing intraindividual L1-L2 differences may pose a challenge for the Lex30 because of individuals' stable response tendencies across time. Indeed, we found significant positive correlations between English and French Lex30 scores ($r = .51$, $p < .001$), indicating that individuals who give many infrequent words in their L1 tend to also give many infrequent words in their L2. It is possible, then, that individuals have similar tendencies or response strategies in their L1 and

L2 with regards to producing infrequent words as associates in word association tests, and/or acquiring infrequent vocabulary items, both of which may influence the Lex30's ability to pick up intraindividual L1-L2 differences. The CR, on the other hand, showed only a weak relation between L1 and L2 scores (r = .26, p = .08). To the extent that this relationship may be reliable, it might reflect individual differences in peoples' general tendencies to produce more or less word associates under test conditions regardless of language. Note, however, that this relationship was statistically weak, as would be expected given that, in general, vocabulary size in the L1 should not predict how much experience a person has with an L2.

Lastly, the other test of the CR's construct validity, i.e., negative correlations with the speed and efficiency of lexical access, provided only partial validity evidence for the proposed methodology. In French, a significant negative relation was observed between residualized CR scores and the speed of lexical access, that is, the L2-specific measures after controlling for L1 performance. This result indicates that participants with higher productive L2 vocabularies tended to respond faster, as evidenced by lower RTs, on the semantic classification (living-nonliving) task. These CR scores, however, were not correlated with CV scores. French Lex30 scores showed a similar significant (negative) correlation with RT but not with CV. The lack of a significant relationship with the CV might be considered unexpected because both large vocabulary (high CR scores) and highly efficient lexical access (low CV scores) should be related to overall level of L2 proficiency. On the other hand, the CV scores came from the LNL task, a task requiring participants to converge on the correct meaning of a stimulus word (e.g., the meaning of *dog* to determine that it is animate). In this context, inefficient processing (instability in the RT that would be reflected in the CV) might result from accessing words and meanings that are incorrect for the LNL task but not necessarily inappropriate as word association responses. Thus, participants with relatively inefficient L2 lexical access processing (higher CVs) might still be able to generate many word associates, suggesting that perhaps a CV score based on the LNL task is not a relevant measure of cognitive fluency for performance in the word association task. Since both the CR and Lex30 vocabulary measures show the same pattern of results with the cognitive fluency measures, it is possible that the number of items in the mental lexicon, regardless of whether that quantity is estimated by the CR technique or by an index of access to infrequent words, is truly not related to how efficiently the cognitive processes underlying language use are conducted. Alternatively, it may be possible that vocabulary size is related to cognitive efficiency at a certain point in L2 development, which our participants may have already passed. While the methods used in the current work don't allow us to make these conclusions definitively, future research is necessary to truly explore the exact

nature of the relation between vocabulary size and different aspects of cognitive fluency.

In English, almost no relation between CR scores and the speed (RT) and efficiency (CV) of lexical access was observed, indicating that the size of an individual's productive vocabulary is not related to the fluency with which cognitive processes involved in lexical access are carried out in the L1. This same pattern was observed with the English Lex30 scores. The lack of a finding in the L1 may be due to ceiling effects or range restrictions operating in the L1, the participants' most fluent language, where performance tended to be less variable than in the L2. It is also possible that for adults the native language is so well practised that the size of the L1 lexicon is no longer a crucial component of cognitive fluency.

Is the Word Association Task an Appropriate Trapping Procedure?

Following Meara and Olmos Alcoy's (2010) suggestion, we used a continuous word association task to elicit vocabulary from participants under the supposition that it would meet a basic assumption of the CR methodology, namely that the capture method used should provide a good chance of capturing whatever it is we intend to measure. We feel that the word association format has shown itself to be a reasonable means of trapping relatively large quantities of content words in a fairly short amount of time. Certainly this method of elicitation has advantages over the continuous written picture description task used by Meara and Olmos Alcoy (2010), whose advanced and intermediate Spanish learners supplied an average of only 73.32 and 47.86 word types, respectively, after two 30-minute writing sessions. These values, which represent the usable data, were less than half of the average number of word tokens supplied by each group in their narratives (Advanced: $M = 194.69$; Intermediate: $M = 116.41$). On the other hand, participants in the present study supplied far more usable data in their L1 and L2 at Time 1 and Time 2 (see Table 1) after only 15 minutes of providing associates to high frequency stimulus words. Thus, in half the time, participants in our study were able to generate roughly twice as many content words in the word association task than Meara and Olmos Alcoy's participants did in the written picture description tasks. In so far as participants engage actively with the task, we feel that the word association task provides a good chance of capturing fairly large quantities of meaningful lexical data from which to estimate productive vocabulary size.

Does the CR Estimate Give Information beyond that which is Available in the Raw Lemma Counts?

Meara and Olmos Alcoy (2010) concluded that the CR gives valuable information above and beyond that which is available in the raw lemma counts, since the

estimates of vocabulary size generated by the Petersen's formula is far greater than both the Time 1 and Time 2 counts. In the current work, we found additional evidence in support of this conclusion. For instance, in English, there was a large significant positive correlation between the raw number of lemmas supplied at Time 1 and Time 2 ($r_s = .81$, $p < .001$), indicating that the number of words participants can generate in the word association task is similar across time, despite the fact that different stimulus words were used at each time. However, the correlations between the English CR score and the raw number of lemmas supplied at Time 1 ($r_s = .32$, $p = .03$) and 2 ($r_s = .48$, $p = .001$) in English are much smaller, and indicate that the raw lemma counts at Time 1 and Time 2 respectively accounted for roughly 10% and 23% of the variance in CR scores. Similarly, in French, there was a large significant positive correlation between the raw number of lemmas supplied at Time 1 and 2 ($r_s = .87$, $p < . 0001$). The correlations between the French CR score and the raw number of lemmas supplied at Time 1 ($r_s = .72$, $p < .001$) and 2 ($r_s = .73$, $p < .001$) in French indicated that the number of lemmas generated at Time 1 and Time 2 accounted for about 52% and 53% of the variance in CR scores, respectively. The CR, then, appears to be more than just the sum of its parts. It may indeed be giving more information about productive vocabulary size than the raw lemma counts give, since the raw counts do not explain all of the variance in CR scores. What exactly does the CR score tell us, then?

Since the CR estimates are far larger than either raw count, they tell us that participants likely have access to, or know, far more words than they were able to supply. However, we are unable to say anything about what those words are and the extent to which participants actually know and can produce them. Furthermore, although the estimates of productive vocabulary size generated in the current work (L1: $M = 979.79$; L2: $M = 708.69$) are far larger than those reported by Meara and Olmos Alcoy (2010), the estimates are still not large enough to be taken at face value. In population ecology research, when the Petersen formula is used, the estimate generated applies to the population as a whole and can truly be taken as an indication of how many animals live in a given area. This cannot be the case in language, where the CR estimates don't reflect the several thousand L1 and L2 words participants likely know to be able to claim the high language proficiencies they reported in the current work. Indeed, results of two separate studies (Goulden, Nation, & Read, 1990; Zechmeister, Chronis, Cull, D'Anna, & Healy, 1995) suggest that educated adult native speakers of English (like the university students who participated in this study) know, in a primarily receptive sense, around 17,000 word families, and Fitzpatrick (2003) estimates that for non-native speakers to function effectively in everyday situations in their L2, they should know at least 2000 words, while 5000–7000 may be needed to function effectively

in an undergraduate English-speaking environment. The CR estimates, then, may have seriously underestimated the absolute L1 and L2 vocabulary size.

Consequently, it may be constructive for us to consider limiting the scope of our generalizations based on the vocabulary size estimates produced by the CR method. Along those lines, we speculate that the estimates generated from the CR methodology reflect the amount of vocabulary an individual has available to complete a task at a given time and under the conditions set up by that task. This may be an indication of an overall larger vocabulary size. If it is valid to interpret the CR score in this way, then the nature of the connections between lexical items, as well as the speed of lexical access are also implicated in the CR score, since individuals with many or stronger links between items in their lexicon may also be able to access those lexical items quickly, even under time pressure, supply them as associates, and subsequently earn high CR scores. Future research into the validity of this interpretation of CR scores is needed.

There are some limitations to the present use of the CR methodology that will deserve attention in the future. These concern underlying assumptions of the technique which may or may not have serious implications for understanding what the CR score reflects. For example, the CR method assumes that the capturability of items is not influenced by which items have already been captured (Lindberg & Rexstad, 2002; Sutherland, 2006). This is probably a reasonable assumption in studies of animal populations; with words, however, one can imagine that there may be effects of priming where one word-associate primes a particular region of the mental lexicon rendering some words more available than others. Another assumption is that an item (animal) can only be captured once in a given capture episode; however, as was observed in the present study, words can be generated more than once within a given capture session. Yet another assumption is that the selection of stimulus words will yield word-associates that are broadly representative of the mental lexicon as a whole (Lindberg & Rexstad, 2002; Sutherland, 2006). This is almost certainly not true in this study, as can be inferred from Tables 1 and 2. Over half the associates given in the L1 (158.17) and L2 (111.85) at Time 2 were within the 1K and 2K bands as evidenced by the relatively low L1 (52.06 items beyond 2K) and L2 (48.66) Lex30 scores at Time 2. Finally, another important assumption of the CR logic used here is that the population under study represents a closed system, in which population size is fixed and stable, rather than an open system, in which population size fluctuates. In population ecology, a closed system approach (such as the CR method) assumes that there are no gains to (births or immigration) or losses from (death or emigration) the population during the course of the study. When this assumption does not hold, ecologists draw on open-system sampling and computational approaches that account for fluctuations in population size (Lindberg & Rexstad, 2002;

Sutherland, 2006). The question arises, then, do the issues raised by open systems in animal populations apply, by analogy, in some way to the case of word populations in the mental lexicon and, if so, how should this shape decisions about the appropriate sampling and computational methods to use? We see this as an empirical question.

It could be argued, on the one hand, that it might be reasonable to view the mental lexicon as a relatively *closed* system in that we do not expect a person's L1 or L2 vocabulary to fluctuate significantly over the short span during which most vocabulary research takes place (usually one or two sessions within a week or so). On the other hand, perhaps the psychological equivalents of birth, immigration, death and emigration are those mental processes that affect accessibility to the words one knows, including priming, one's emotional state, or the mental strategies used to accomplish the task at hand. Additionally, Sutherland (2006) points out that the computational methods chosen can vary in terms of precision and the degree of bias in the estimations they yield. Methods appropriate for closed systems (such as the CR method) can yield biased estimates when the population is in fact open. Methods appropriate for open systems do not yield biased estimates even when applied to a closed system, but the estimates are less precise. Future research will need to address the extent to which the mental lexicon should be considered an open system and what the consequences of this might be for how vocabulary samples are taken. Clearly, it is not possible to address the issues raised by all of these assumptions within the confines of the present study, but in principle it should be possible in future studies to test these assumptions in order to better understand the possibilities and limits of the CR measure.

Conclusions

We set out to investigate whether the CR methodology can be considered a valid measure of productive vocabulary size in a second language. An easily constructed word association task was used to elicit fairly large quantities of content words from participants in a short amount of time, with the advantage that it did not restrict participants' production or artificially raise the number of repeat items. Additionally, convergent validity of the CR methodology was established based on significant positive correlations between CR and Lex30 scores. These two tests may be tapping different, but complementary, aspects of productive vocabulary knowledge. Although the CR outperformed the Lex30 in a number of ways, our intention was not to pit the two tests against each other, since we feel that together they have the potential to be a rich source of information about productive vocabulary knowledge. Indeed, since the word association format is used to elicit

data for both the CR and Lex30 estimates, professionals will be able to score the same data in two ways, and convey to language learners an index of their progress in the language in terms of both an estimate of approximately how many words they may know or have access to and what proportion of those words tend to be infrequent. Whether or how we can use the CR and Lex30 scores together to give more information about productive vocabulary size than either of them can give alone, remains an open empirical question.

The CR technique, as implemented in the current work, also displayed good construct validity, as evidenced by its ability to distinguish participants' L1 from their L2, and by its significant relation to the speed of lexical access. Taken together, these findings suggest that the CR methodology holds promise as a valid means of measuring the complex construct of productive vocabulary size. At present, however, interpretation of the CR estimate requires some caution until more is known about how the measure behaves under different testing conditions, with different stimulus materials, etc. For now it may be more prudent to limit the scope of our generalizations and interpret the CR estimate as a reflection — not an absolute estimate — of the size of an individual's L2 productive vocabulary available under certain task conditions. Whether or not this index ultimately proves to provide a broadly useful measure of overall productive vocabulary size remains for future research to uncover.

Authors' Note

The authors would like to acknowledge the support for this research from McGill University's Training and Retention of Health Professionals project, funded by Health Canada, through a subaward to Norman Segalowitz of the project's Health-Care Access for Linguistic Minorities (H-CALM) research team. Correspondence with the authors of this paper should be directed to Joy Williams at <joyawilliams@gmail.com> or to Norman Segalowitz at <norman.segalowitz@concordia.ca>.

Notes

1. A dependent *t*-test was used to compare the number of lemmas generated in French across time because this variable did not violate the assumption of normality.

2. This finding also helps to confirm the assumption underlying use of the Lex30 test, namely that individuals with larger lexicons are more likely to have access to a greater number of infrequent words.

References

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279. DOI: 10.1093/ijl/6.4.253

Clenton, J. (2008). Investigating the construct of productive vocabulary: Comparing different measures. In M. Edwardes (Ed.), *Taking the measure of Applied Linguistics: Proceedings of the 41st Annual Meeting of the British Association for Applied Linguistics*, 11–13 September 2008 (pp. 27). Swansea University. Retrieved from: http://www.baal.org.uk/proc08/clenton.pdf

Cobb, T. (n.d.). Web VocabProfile [accessed 31 July 2013 from: http://www.lextutor.ca/vp/eng/], an adaptation of Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved from: http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. New York: Routledge.

Fitzpatrick, T. (2003). Eliciting and measuring productive vocabulary using word association techniques and frequency bands. Ph.D. dissertation, University of Wales Swansea.

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, 27, 537–554. DOI: 10.1177/0265532209354771

Fitzpatrick, T., & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *Vigo International Journal of Applied Linguistics*, 1, 55–73. Retrieved from: http://webs.uvigo.es/vialjournal

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

JACET Basic Words Revision Committee (Eds.). (2003). *JACET list of 8000 words (JACET 8000)*. Tokyo: JACET.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 19, 255–271. DOI: 10.1093/applin/19.2.255

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307–322. DOI: 10.1093/applin/16.3.307

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16, 33–51. DOI: 10.1177/026553229901600103

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, 48(3), 365–391. DOI: 10.1111/0023-8333.00046

Lindberg, M., & Rexstad, E. (2002). Capture-recapture sampling designs. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics*, Volume 1 (pp. 251–262). Chichister: John Wiley & Sons.

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

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 109–121). Cambridge: Cambridge University Press.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins. DOI: 10.1075/lllt.24

Meara, P., & Fitzpatrick, T. (2000). Lex 30: An improved method for assessing productive vocabulary in an L2. *System*, 28, 19–30. DOI: 10.1016/S0346-251X(99)00058-5

Meara, P. M., & Olmos Alcoy, J. C. (2010). Words as species: An alternative approach to estimating productive vocabulary size. *Reading in a Foreign Language*, 22, 222–236. Retrieved from: http://nflrc.hawaii.edu/rfl

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84–102). Cambridge, MA: Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters. DOI: 10.1057/9780230242258

Nation, I. S. P. (Ed.). (1984). Vocabulary lists: Words, affixes and stems. *English Language Institute Victoria University of Wellington Occasional Paper*, 12.

Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Linsfjord from the German Sea. *Report of the Danish Biological Station*, 6, 5–84.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511732942

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave. DOI: 10.1057/9780230293977

Segalowitz, N. (2010). *The cognitive bases of second language fluency*. New York: Routledge.

Sutherland, W. J. (2006). *Ecological census techniques: A handbook* (2nd ed.). Cambridge, MA: Cambridge University Press. DOI: 10.1017/CBO9780511790508

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson.

Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly*, 9(2), 172–185. DOI: 10.1080/15434303.2011.625579

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, 39, 79–95.

Wolter, B. (2002). Assessing proficiency through word associations: Is there still hope? *System*, 30, 315–329. DOI: 10.1016/S0346-251X(02)00017-9

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior*, 27(2), 201–212.

Zimmerman, K. J. (2004). *The role of vocabulary size in assessing second language proficiency*. Ph.D. dissertation, Brigham Young University.

Appendix A

Table A. Words used in the English and French Word Association Tasks at Time 1 and Time 2.

English		French	
Time 1	**Time 2**	**Time 1**	**Time 2**
aim	bad	animer	arriver
become	believe	assemblée	attendre
child	building	autoriser	chercher
consider	can	bataille	concerner
family	day	bonheur	considérer
game	eye	caractère	couvrir
get	father	découverte	descendre
grow	give	devoir	doute
help	good	diriger	entendre
holiday	guest	élire	espérer
home	head	empêcher	éviter
hour	high	évoluer	fermer
improve	kind	feu	fil
live	know	inquiétant	frère
lose	leave	juger	gérer
name	let	monde	gestion
news	new	particulier	identité
provide	political	poche	législatif
run	right	posséder	mise
see	school	préparer	niveau
side	send	prestation	paraître
skin	soil	prétendre	poste
soldier	stand	prier	pratique
start	take	relever	règle
state	talk	remarquer	rentrer
story	tip	réussite	répondre
student	water	secrétaire	secteur
tell	way	souci	venir
work	win	terminer	vente
write	woman	vitesse	vigueur

Note. These words were presented to participants in a fixed randomized order at Time 1 and Time 2.

Corresponding Address

Joy Williams
Psychology Department, Concordia University
Room SP-244
7141 Sherbrooke Street West
Montréal (Québec) H4B 1R6

joyawilliams@gmail.com