# An Alternative Approach to Measuring Second Language Productive Vocabulary Size: A Validation Study of the Capture-Recapture Methodology

Joy Williams

A Thesis

In

The Department of

Education

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Arts (Applied Linguistics) at

Concordia University

Montreal, Quebec, Canada

September 2012

**CONCORDIA UNIVERSITY**

**School of Graduate Studies**

This is to certify that the thesis prepared

By: Joy Williams

Entitled: An Alternative Approach to Measuring Second Language Productive Vocabulary Size: A Validation Study of the Capture-Recapture Methodology

and submitted in partial fulfillment of the requirements for the degree of

**Master of Arts (Applied Linguistics)**

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____________ Dr. Walcir Cardoso_____________ Chair

____________ Dr. Marlise Horst ______________ Examiner

____________ Dr. Sarita Kennedy_____________ Examiner

____________ Dr. Norman Segalowitz__________ Supervisor

Approved by

_____________Dr. Pavel Trofimovich___________
Chair of Department or Graduate Program Director

_________________Dr. Brian Lewis_____________
Dean of Faculty

Date September 4, 2012

# Abstract

An Alternative Approach to Measuring Second Language Productive Vocabulary Size: A Validation Study of the Capture-Recapture Methodology

**Joy Williams**

This study provides validity evidence for the ecological estimation technique, the Capture-Recapture (CR) method, as an estimate of second language (L2) productive vocabulary size (PVS). Two separate “captures” of productive vocabulary were taken using a word association task (WAT). During the first capture (T1), 47 bilinguals completed different WATs in their first language (L1), English and L2 (French) by providing 4-6 associates to each of 30 high-frequency stimulus words in English and French. A few days later (T2), this procedure was repeated with a different set of stimulus words in each language. Since the WAT was used, data were scored using the traditional Lex30 scoring and using the Petersen formula, which generates a CR estimate of PVS. Participants also completed an animacy judgment task designed to assess the speed and efficiency of lexical access.

The CR’s convergent validity was confirmed by significant positive correlations with Lex30 scores in English and French. The construct validity of the CR was also confirmed by 1) its ability to indicate that L1 PVS was significantly larger than L2 PVS, and 2) its significant correlation with the speed of lexical access. While these results hint at the validity of the technique as an estimate of L2 PVS, the CR scores are not a direct indication of absolute vocabulary size. Instead, it may be more realistic to interpret these estimates as indicative of how much vocabulary is available for task completion. The validity of this interpretation needs to be explored further.

## Acknowledgements

My sincerest thanks and appreciation go out to my supervisor, Dr. Norman Segalowitz, for his direction, patience and input during the course of writing this thesis. I would also like to thank my committee members, Dr. Sarita Kennedy and Dr. Marlise Horst, for their valuable insights and guidance, especially in the planning stages of this project. To the members of the Segalowitz Lab, especially Tatsiana Leclair, thank you, thank you, thank you for the constant support and feedback throughout this process, especially during crunch time. Tatsiana, I definitely could not have completed this project without your hard work and indispensable help over the summer. You've truly helped to make "order out of disorder", as you say, and it is much appreciated!! Last, but definitely not least, I would like to express my gratitude to my family members – my parents and siblings – for their unwavering encouragement and for pushing me when I really had nothing left. Thank you!

# TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

## Introduction

The current work documents our attempt to assess productive vocabulary size using a novel approach recently advocated by Meara & Olmos Alcoy (2010). The approach tested here is the Capture-Recapture (CR) methodology, which involves using a capture-recapture sampling technique to compute what is known as the Petersen Estimate (Petersen, 1896). Both the sampling technique and the Petersen Estimate are traditionally used in ecological studies to accurately estimate how many animals of a given species inhabit a given habitat. The overarching goal of the current work, then, is to validate this unconventional approach as an effective means of assessing second language vocabulary size, a construct that has been difficult to measure. Before we delve deeper into the logic of this proposed methodology and how it was implemented and analysed in our study, it is useful to first discuss the importance of vocabulary size, the components of vocabulary knowledge and the challenges associated with counting words, the construct of productive vocabulary and problems associated with its measurement, the benefits of word association tasks in measuring productive vocabulary and finally the application of the word association test format into the Capture-recapture methodology proposed in the current work.

### Why Vocabulary Size?

> *"Without grammar very little can be conveyed, without vocabulary nothing can be conveyed"* (Wilkins, 1972, p. 111)

This often cited quote speaks to the very practical importance of vocabulary. It is not surprising, then, that language learners and native speakers alike see amassing a large vocabulary as a desirable goal and often equate mastery of the language with being able to understand and use a large number of words (Fitzpatrick, 2003; Read, 2000). This importance of vocabulary to communication and second language acquisition has been reflected in both the renewed interest in examining this construct empirically, as well as in the rapidly growing pool of second language (L2) vocabulary assessment tools designed to estimate various dimensions of vocabulary knowledge, such as size (Fitzpatrick, 2003; Read, 2000).

Indeed, results from these bodies of work provide empirical support for the intuitive notion that individuals with a larger vocabulary size are more effective language users, as evidenced by correlations between vocabulary size and measures of receptive (Belgar & Hunt, 1999; Laufer, 1992) and productive language performance (Laufer & Nation, 1995; Zimmerman, 2004). Laufer (1992), for example, found highly significant positive correlations between reading comprehension and vocabulary size, as measured by both Nation's (1983) Vocabulary Level's Test (VLT; $r = .50$, $p = .0001$) and the Eurocentres Vocabulary Test (Meara & Jones, 1988; $r = .75$, $p = .0001$). Additionally, Belgar and Hunt (1999) measured vocabulary size using two modified versions (A and B) of the 2000 word-level section of the VLT and two modified versions (A and B) of the University Word List (UWL) section of the VLT. Their analyses revealed that TOEFL reading comprehension scores were significantly positively correlated with vocabulary size estimates based on knowledge of the 2000 most frequent words (Version A: $r = .66$) and (Version B: $r = .62$), and knowledge of the words on the UWL, (Version A: $r = .67$) and (Version B: $r = .71$). Zimmerman (2004) also found that the vocabulary size of his participants was significantly positively correlated with their performance on the listening (*r* = .66) and reading (*r* = .60) sections of a placement test.

Results from studies on productive skills also show this pattern. For instance, Laufer and Nation (1995) found significant correlations, ranging from .60 to .80, between vocabulary size and lexical richness, such that learners with larger vocabularies used more sophisticated vocabulary in two written compositions. Belgar and Hunt (1999) also found that individuals with larger vocabulary sizes, as measured by their modified versions of the VLT, performed better on the Structure and Written Expression sections of the TOEFL, with correlation coefficients ranging from .59 to .65. Vocabulary size was also strongly correlated with speaking performance on the placement test (*r* = .66) in Zimmerman's (2004) study.

Taken together, these results suggest that vocabulary size influences proficiency in all four language skills. Since vocabulary size has such significant implications for language use, it is important that researchers develop means of accessing and assessing this construct in valid and reliable ways. Achieving this would not only help to elucidate the nature of the mental lexicon, but would also help inform pedagogical or curriculum-based decisions and allow researchers and professionals to convey to language learners evidence of their language competence in terms of an absolute number, information which is especially attractive to language learners (Fitzpatrick, 2003). Unfortunately, though, arriving at valid and reliable vocabulary size measures has proven to be anything but straightforward, as the very concept of vocabulary knowledge is a complex and multifaceted one, with surprisingly nuanced units of measurement.

## Vocabulary Knowledge

**The units of vocabulary.** Intuitively, the concept of vocabulary knowledge can be seen as referring to the knowledge of words. For researchers, however, the term 'word' is not so straightforward in its meaning. As Milton (2009) points out, researchers interested in vocabulary knowledge "tend to use the word 'word', presumably for ease and convenience, [to refer] to some very specialist definitions of the term, such as *types, tokens, lemmas, word families…*" (p. 7), each of which has implications for the inferences made regarding vocabulary knowledge. Let's take *types* and *tokens* for example. The term *types* is used to refer to the total number of different words in a text or corpus, while *tokens* refers to the total number of words in the text or corpus overall. The following sentence, therefore,

*The girl quickly picked the prettiest flowers*

includes 6 types, since the word *the* is counted only once, and 7 tokens, since *the* is counted each time it appears. This sample sentence raises an important issue associated with counting words for the purpose of coming to meaningful conclusions about an individual's vocabulary knowledge, namely whether the function words (e.g., articles, pronouns, conjunctions, auxiliary verbs, etc.) should be regarded as vocabulary items, in the same way as content words (e.g., nouns, main verbs, adjectives and adverbs) are. In the current work, we take the conventional view that since function words have little or virtually no meaning as isolated lexical items and provide support to the content words in terms of linking them together meaningfully or modifying their meaning (Read, 2000), knowledge of such words is not of primary interest here. As such, all vocabulary size estimates made in the current work will be based on production of content words.

However, while a focus on content words promises to tell the most interesting story of vocabulary development and knowledge, these words exist in a variety of forms, e.g., *pick* and *picked*, *flower* and *flowers*, *quick* and *quickly*. It is crucial that researchers interested in vocabulary knowledge express clearly and explicitly which form of a word their participants will be rewarded for using (Milton, 2009). Most word frequency counts and estimates of vocabulary size are based on counts of *lemmas*, i.e., a group of words consisting of a headword or base form and its most frequent inflected forms (Daller, Milton & Treffers-Daller, 2007; Milton, 2009; Read, 2000). When researchers are interested in counting the number of words in a spoken or written text, a common first step is to lemmatize the content words so that the inflected forms of a base word, provided that they are of the same part of speech as the base word, would all be counted as instances of the same lemma. For example, the verb forms *adapts, adapted,* and *adapting*, would all be counted as instances of the same lemma, identified by the headword *adapt*, while, *adaptation,* which is a noun, would not be considered part of this lemma and would be counted separately. In research focused on second language vocabulary size, estimates based on counts of lemmas are preferred and the use of lemmatized frequency-based wordlists is fundamental to vocabulary tests such as the Vocabulary Levels Test (Nation, 1983; 1990) and the X-Lex (Meara & Milton, 2003).

On the other hand, in some other vocabulary tests, such as Goulden, Nation and Read's (1990) test aimed at estimating first language vocabulary size, the interest is in a much larger unit of measurement, namely the *word family*. Word families include not only the base form of a word and its most frequent inflected forms, but also the derived forms of the base word that are closely related in meaning (Daller, Milton & Treffers-Daller, 2007; Milton, 2009; Read, 2000). Thus, while the noun *adaptation* would not be considered part of the lemma identified by the headword *adapt*, it would be considered part of the same word family as this base word, along with *adapts, adapted,* and *adapting*, and other derivations of the base word, like *adaptable*, *adaptability* and *adaptive*. Counting word families will obviously result in smaller vocabulary size estimates than counting lemmas, since words that would be counted separately in a lemmatized count would be considered instances of the same headword if word families were of interest.

Deciding whether to base vocabulary size estimates on knowledge of lemmas or word families is not at all a trivial matter since “determination of what constitutes a *Word* for counting and analysis…[has] important ramifications not only for the lexical findings themselves, but also for the pedagogical theories and practices that derive from them” (Gardner, 2007, p. 242). In the current work, vocabulary size estimates will be based on counts of lemmas, rather than word families. Counting lemmas may be more valid because it allows us to observe the range of productive knowledge an individual has, since derivations are typically considered as separate lexical items. Counting word families, however, would mask such information since inflected and derived forms of a base word are all considered instances of the same lexical item, even though showing productive knowledge of one or a few items of a word family does not imply that all other members of the family are known (Vermeer, 2004; Nation, 2007).

**Components of vocabulary knowledge.** From a more macro perspective, the complexity of vocabulary is also evident. Vocabulary knowledge is not at all a unitary construct! Nation’s (2001) idealized conceptualization of vocabulary knowledge (see) makes this point very clear.  According to this framework, knowing a word involves being familiar with the component elements of its form, meaning and use. The common element among these different subcomponents of the three elements of word knowledge is that they all have receptive and productive manifestations. This highlights one of the most commonly made distinctions in the field, i.e., that between receptive or passive vocabulary and productive or active vocabulary, which is of interest here.

Table 1: What is involved in knowing a word

| | | | |
|---|---|---|---|
| Form | spoken | R | What does the word sound like? |
| | | P | How is the word pronounced? |
| | written | R | What does the word look like? |
| | | P | How is the word written and spelled? |
| | word parts | R | What parts are recognisable in this word? |
| | | P | What word parts are needed to express the meaning? |
| Meaning | form and meaning | R | What meaning does this word form signal? |
| | | P | What word form can be used to express this meaning? |
| | concept and referents | R | What is included in the concept? |
| | | P | What items can the concept refer to? |
| | associations | R | What other words does this make us think of? |
| | | P | What other words could we use instead of this one? |
| Use | grammatical functions | R | In what patterns does the word occur? |
| | | P | In what patterns must we use this word? |
| | collocations | R | What words or types of words occur with this one? |
| | | P | What words or types of words must we use with this one? |
| | constraints on use (register, frequency…) | R | Where, when, and how often would we expect to meet this word? |
| | | P | Where, when, and how often can we use this word? |

*Note.* In column 3, R = receptive knowledge, P = productive knowledge. From Nation (2001, p. 27).

Receptive vocabulary refers to those lexical items that an individual can recognize and understand when listening to speech or reading text, while productive or active vocabulary refers to the set of vocabulary items that an individual can produce accurately when speaking or writing (Milton, 2009; Schmitt, 2010). The relationship between these two elements of word knowledge is not entirely clear or straightforward (Daller, Milton & Treffers-Daller, 2007; Milton, 2009; Schmitt, 2010). While it is convenient to view receptive and productive vocabulary knowledge as distinct entities, no official boundary or criterion has, as yet, been empirically established that definitively distinguishes a word that has receptive status from one that has productive status (Read, 2000). This concern applies even to Melka's (1997) conceptualization of vocabulary knowledge as a continuum where, with increasing familiarity with and knowledge of a given word, receptive abilities gradually give way to productive knowledge. Nevertheless, the receptive-productive distinction is accepted and widely used and researchers in the field, have been able to show that receptive and productive vocabulary size are at least correlated, such that those who can handle more lexis receptively, can also do so productively, although not necessarily for the same lexical items (Laufer, 1998; Webb, 2008). Empirical evidence also 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 than its productive counterpart (Fitzpatrick, 2003; Laufer, 1998; Laufer & Paribakht, 1998; Milton, 2009; Schmitt, 2010; Webb, 2008; Zimmerman, 2004).

This relative ease of measurement of receptive vocabulary knowledge seems to have had implications for research in the field since, as Meara and Fitzpatrick (2000) note, most of the wealth of vocabulary research claiming links between vocabulary knowledge and more proficient language use are actually based on measures of receptive vocabulary knowledge. From a practical standpoint, this makes sense since, as Fitzpatrick (2003) points out, “Asking a subject “do you know what word x means?” is much more straightforward and time-efficient than any attempt to elicit word x from their mental lexicon” (p. 6). In fact, this strategy of pre-selecting representative vocabulary items to test in this way is a necessary step taken when assessing passive vocabulary knowledge and is a key feature of three well known measures of this construct: (1) The Vocabulary Levels Test (Nation, 1983, 1990), which involves word-definition matching at five levels of word frequency in English; (2) The Eurocentres Vocabulary Size Test (EVST; Meara & Jones 1988, 1990), which is a computer-based checklist test that requires test-takers to indicate whether or not they know words drawn from a range of frequency levels in English; and (3) The Vocabulary Knowledge Scale (Wesche & Paribakht, 1997), which requires test takers to indicate which of 5 categories best represents the degree to which they know a given word. While pre-selecting items to test is necessary for measuring receptive vocabulary knowledge, as we will see in the following section, this strategy is often mentioned as one of the limitations of measures of productive vocabulary size.

### Assessing Productive Vocabulary

Assessing productive vocabulary knowledge has proven to be a more challenging endeavour for a number of reasons. First, the very construct of productive vocabulary seems to be complex and difficult to define, especially for the purposes of measurement, since, intuitively, being able to produce a word could mean anything from having knowledge of the orthographic form of the word to being able to use it competently in context. Fitzpatrick (2007) questions even the validity of the construct itself by cautioning that “many of the studies which use the concept of productive vocabulary are closely linked with the design of vocabulary tests, which encourages us to be wary that the construct is not an artificial one springing from a desire to find attractive and efficient ways of testing” (p. 130). Further, if we are to assume that the construct itself is a valid one, the issue of the scope of what can be considered productive knowledge needs to be considered. Fitzpatrick (2003) suggests that any attempt to investigate productive vocabulary must begin by deciding whether to define the construct in terms of the lexical items an individual actually uses in natural communication, or in terms of the lexical items an individual has the potential to use, but has not chosen to use. Care must be taken in interpreting productive vocabulary size estimates based on either of these operationalizations of the construct. Further, Laufer’s (1998) conceptualization of the construct introduces the idea of degrees of productive ability, i.e., controlled productive ability, which refers to one’s ability to use a given word when prompted or required to do so, and free productive ability, which refers to one’s ability to use a given word at will, without any particular prompts. Additionally, Read (2000), in an attempt to address the inconsistent definitions of productive (and receptive) vocabulary knowledge in the literature, proposes a different way of distinguishing reception and production for the purposes of assessment. As can be seen in Table 2, this conceptualization suggests that productive vocabulary knowledge can manifest itself in two ways, context-independent recall, where “they are presented with some stimulus designed to elicit the target word from their memory” (Read, 2000, pp. 155) and context-dependent use, where “the word occurs in their own speech or writing” (Read, 2000, pp. 156). Certainly, there appear to be interesting views on vocabulary knowledge in general and on productive vocabulary knowledge, in particular, but no clear consensus on exactly how to conceptualize these constructs, a problem that poses a significant challenge to measurement (Fitzpatrick, 2003; Read, 2000; Schmitt, 1997).

Table 2: *Summary of the Types of Vocabulary Knowledge, based on Read*

| | Receptive | Productive |
|---|---|---|
| Context-independent | Recognition[a] | Recall |
| Context-dependent | Comprehension[a] | Use |

Note. Adapted from Read (2000, p. 154-157).
[a] According to Read (2000), recognition involves an individual showing understanding of a word's meaning by, for example, selecting its definition in a multiple choice task, while comprehension refers to being able to understand words in context when listening or reading.

Furthermore, the challenges facing researchers interested in assessing productive vocabulary are not limited to defining the construct. Measures of productive vocabulary knowledge tend to be time inefficient, too controlled, context-dependent, assess pre-selected targets, limit test-takers to the production of one correct response, test receptive abilities also, and elicit insufficient quantities of content vocabulary from which to make meaningful inferences. These problems are best clarified by a discussion of some influential productive vocabulary size measures, the Lexical Frequency Profile (LFP; Laufer and Nation, 1995), the Productive Vocabulary Levels Test (PVLT; Laufer & Nation, 1999) and V_Size (Meara & Miralpeix, 2007).

Although there are tests that make more direct attempts to estimate the size of an individual’s productive vocabulary, most of the investigations into the nature of productive vocabulary seem to be based on analysis of L2 users’ texts in terms of their lexical richness or complexity, as evidenced by type-token ratios or lexical frequency profiles (Meara & Miralpeix, 2007). One such test, designed to assess the lexical richness of learners’ written texts, is Laufer and Nation’s (1995) LFP. Researchers have found this test to be useful and effective at estimating the size of the L2 productive vocabulary (Edwards & Collins, 2011). In order to obtain a lexical frequency profile, Laufer and Nation (1995) asked learners to write two essays of 300-350 words each, one on a general issue, and the other on a controversial issue. A software program, known as VocabProfile, was then used to construct the lexical frequency profile by computing the proportion of word families in the first 1000 most frequent words, the second 1000, the UWL and off-list items (Laufer & Nation, 1995). A disadvantage of this method is that it appears to be fairly time consuming, requiring 1 hour per composition. Further, the LFP requires test-takers to produce fairly lengthy texts, at minimum 200 word tokens per essay if stable results are to be obtained, on two topics that may not require the use of vocabulary that is representative of learners’ lexicon (Fitzpatrick, 2003; Meara & Fitzpatrick, 2000).

The problem of context dependence is also associated with another of Laufer and Nation’s (1999) tests, namely the PVLT, which requires learners to read a sentence and complete the target word. The first letters of the target word are provided to rule out other semantically viable options that are not being tested. In the example below, the word *episodes* is being elicited (Laufer & Nation, 1999, pp. 37).

The book covers a series of isolated epis_________ from history.

This test samples 18 items from each of the 2000, 3000, 5000, University Word List (UWL) and 10 000 word levels and a score for the number of correct items at each word level, and overall, is calculated. An initial concern associated with the PVLT is that it may not be entirely valid to make inferences about productive vocabulary knowledge as a whole from 18 pre-selected items from each of the frequency bands of interest. Additionally, aside from the fact that production is limited by context and to one correct answer, providing as many initial first letters as necessary to effectively disambiguate the cue means that, at times, most of the word stem would be provided for test-takers (Read, 2000). Read (2000) points out that there is considerable variability in the demands placed on the test-taker as a function of how many initial letters are included. “This means that some test items require more word knowledge – and more use of contextual information – than others do, which complicates the issue of what the test as a whole measures” (Read, 2000, p. 125). Thus, although the authors refer to this tool as a test of controlled productive ability, the PVLT may be tapping more than just production. It may not be possible to draw conclusions that are specific to productive knowledge since receptive abilities are also required when considering the context of the sentence and the number of initial letters provided (Fitzpatrick, 2003; Fitzpatrick & Clenton, 2010; Read, 2000).

Like the LFP (Laufer & Nation, 1995), Meara and Miralpeix’s (2007) technique also requires individuals to produce texts from which a lexical frequency profile can be created. Their computer program, V_Size, is then able to produce an estimate of the productive vocabulary size an individual would need to produce such a frequency profile. V_Size does so by comparing the actual lexical profiles of learners’ texts to a series of theoretical profiles generated from calculations based on Zipf’s Law. The goal of this comparison is to find the best match and subsequent vocabulary size estimate for each participant. The limitations associated with testing productive vocabulary in context, through written texts, apply to this technique as well. Additionally, V_size estimates should be interpreted with caution since the "results we get from V_Size vary depending on the dictionary that is used as a comparator for the text, for example, and reclassifying a small number of items can have a surprisingly large effect on the overall vocabulary size estimate" (Meara & Miralpeix, 2007, p. 3).

From this brief review, it can be concluded that productive vocabulary is a complex construct and measuring it in valid ways has proven challenging. Despite their limitations, these tests have merits of their own and the LFP and the PVLT and are actually fairly widely used. However, since the nature of what is being measured by these tests remains unclear and the results gathered from them are difficult to interpret, Meara and Olmos Alcoy (2010) advocate investigating the construct of productive vocabulary from "different, perhaps unconventional points of view" (p. 223). What follows is a review of Meara and Olmos Alcoy's (2010) attempt to do just that in their recent paper titled *Words as species: An alternative approach to estimating productive vocabulary size.*

## Capture-Recapture Methodology and Petersen Estimate

Meara and Olmos Alcoy (2010) point out that the major problem associated with measuring productive vocabulary knowledge is that it is impossible, especially at higher levels of language proficiency, to create a test that elicits all of the words in an individual's lexicon. The solution to this problem has been to estimate the overall vocabulary size from a smaller sample of vocabulary. However, since vocabulary use tends to be highly context-specific, and since it is not easy to create tasks that sample vocabulary in sufficiently large quantities needed for meaningful estimation, even this inferential method has proven problematic in terms of both implementation and interpretation of results (Meara & Fitzpatrick, 2000; Meara & Olmos Alcoy, 2010).

Meara and Olmos Alcoy (2010) attempted to overcome some of the difficulties facing productive vocabulary measurement by using a rather unconventional approach. They borrowed the Capture-Recapture methodology that is commonly used in Ecology to reliably and accurately estimate the size of animal populations living in a given area. Using this method, ecologists take two separate, but representative, samples of the animal population of interest, taking note of the number of animals that appear in both of the samples. For instance, Meara and Olmos Alcoy (2010) provide the example of an ecologist interested in determining how many fish of a given species live in a river. In order to arrive at such an estimate, the ecologist would first select an appropriate section of the river from which to sample the fish. This section of river should be representative of the conditions that exist in the entire river and provide a good chance of sampling the fish of interest. Second, the ecologist will take his first sample of fish (Time 1) by using a suitable trapping technique, such as casting a wide net, in order to capture the fish that swim through the chosen section of river. All of the fish captured at Time 1 will be counted and marked so that they can be easily identified should they return in future captures. These marked fish will then be released to continue moving naturally in the river. Next, after a predetermined period of time, enough to allow the population of fish to redistribute itself evenly in the river, the ecologist will take his second sample of fish (Time 2) using the same method as at Time 1. A count of the fish captured at Time 2 will be obtained along with the number of marked fish that were captured at Time 1, which also appear in the Time 2 sample. 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. In order to estimate the total population of fish in the river (*P*), the ecologist then plugs these 3 values into a formula known as the Petersen Estimate (Petersen, 1896), which is calculated by dividing the product of the Time 1 and Time 2 captures (x*y*) by the number of 'repeat' fish (*r*), such that ($P = xy/r$).

However, in order for the Petersen formula to provide meaningful estimates of population size, a number of assumptions must be met. First, the capture method used should provide a good chance of capturing whatever it is we intend to measure, be it fish in a river or vocabulary in the mental lexicon. Second, in keeping with the fish analogy, the stretch of river from which we choose to sample must be representative in some way of the river as a whole (Meara & Olmos Alcoy, 2010). Third, according to Meara and Olmos Alcoy (2010)

> The mathematics only works in a straightforward way if we assume that the two collection times are equivalent, and if each animal has an equal chance of being counted on both collection times. The population of fish needs to be constant from Day 1 to Day 2 – if half our fish were killed by otters, or died from poisoning overnight, then Petersen's model would simply not apply. (p. 226)

**Meara and Olmos Alcoy's (2010) Study**

Meara and Olmos Alcoy (2010) were interested in determining whether this ecological approach could be adopted to make estimates of productive vocabulary size.

To explore this possibility, they recruited 24 native speakers of English, 11 of whom were intermediate learners of Spanish, while the remaining 13 were advanced learners of Spanish, according to the class teacher. Meara and Olmos Alcoy (2010) chose to 'trap' their participants' vocabulary by using a single 30-minute writing task in which participants produced short texts (they didn't specify whether a word limit was set) describing the six-picture cartoon story, summarized below:

> In the first picture, a man and a boy are playing with a dog beside the sea. The boy throws a stick into the sea for the dog to fetch. The second picture shows this game being observed by a smartly dressed man with an umbrella. In the third picture, this man approaches the dog and shows it his umbrella. The fourth picture shows the smart man throwing his umbrella into the sea. Unfortunately, the dog ignores this. In the fifth picture, the man, the boy, and the dog abandon the smart man, leaving his umbrella floating on the water. The final picture shows the smart man removing his clothes, presumably so that he can swim out to sea and rescue his lost umbrella. (Meara & Olmos Alcoy, 2010, p. 227)

This procedure was completed two times, one week apart. 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. As can be seen from Table 3, at both time points, as well as overall, the advanced group supplied significantly more word tokens and types in their stories than did the intermediate group. Additionally, a Mann-Whitney *U* test confirmed that the Petersen estimate of productive vocabulary size reliably distinguished between the intermediate (*M* = 93.81, *SD* = 31.30) and advanced groups ($M$ = 160.37, $SD$ = 38.51), $U$ = 9.5, $p$ < .01. 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.

Table 3: *Mean Word Token and Type Count, and Petersen Estimate for each group*

| | Group | |
|---|---|---|
| Counts | Advanced<br>*M* (*SD*) | Intermediate<br>*M* (*SD*) |
| | Word tokens | |
| T1 narrative | 190.23 (48.72) | 99.19 (27.16) |
| T2 narrative | 199.15 (63.63) | 133.63 (40.28) |
| Combined | 389.38 (59.81) | 232.81 (89.94) |
| | Word types | |
| T1 narrative | 72.91 (17.00) | 43.36 (8.89) |
| T2 narrative | 73.73 (19.09) | 52.36 (15.09) |
| Repeats | 33.55 (9.11) | 25.82 (6.91) |
| | Petersen Estimate | |
| | 160.37 (38.51) | 93.81 (31.30) |

*Note*. Adapted from Meara and Olmos Alcoy (2010, p. 229).

While these preliminary results suggest that the capture-recapture method and resulting Petersen Estimate may hold some promise as a measure of productive vocabulary size, there are a number of limitations to the procedure adopted by Meara and Olmos Alcoy (2010) that center on their choice of trapping instrument. First, the use of a writing task may have violated the assumption of representativeness in sampling since the context associated with this technique may not be capable of eliciting lexis that is representative of learners' productive vocabulary knowledge as a whole. Additionally, recall that an assumption of the Petersen estimate is that each item has an equal probability of being counted at Time 1 and Time 2. Having participants describe the same picture story twice means that the words necessary to describe the events depicted in that picture story have a greater chance of being captured and recaptured, than the rest of the productive vocabulary in the individuals' lexicon. Indeed, Racine (2011) points out that "by assigning the same task at Time 2, the researchers have essentially fed the fish, increasing the likelihood that they will return to the net at Time 2" (p. 235). Further, the Petersen estimate may have been lowered simply because participants described the exact same picture story at Time 1 and Time 2. This greatly increases the number of 'repeat' items, which is the denominator in the Petersen estimate formula, since it is virtually impossible to tell the story without using function words or content words like "*man*, *boy*, *stick*, *dog*, *throw*, *water*" (Meara & Olmos Alcoy, 2010, p. 231). The data presented in Table 3 hints at this, since, for the Advanced group, 45.50% of the word types produced at Time 2, also occurred at Time 1, while for the Intermediate group, 49.31% of the word types produced in the Time 2 narrative, were also produced at Time 1. The use of an inappropriate 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 their estimates, 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.

The importance of the technique used to elicit or 'trap' vocabulary from participants cannot be understated. Meara and Olmos Alcoy (2010) acknowledged this and suggested that a trapping method in the form of a word association task might be able to elicit more words without increasing the likelihood that participants would repeat words at both time points. This possibility will be explored in the current work.

## The Word Association Format

The word association format may indeed have potential to be a more suitable trapping method for individuals' vocabulary. Typically, a word association test requires participants to write down or say aloud the first related word, or associate, that comes to mind when a given stimulus word is encountered (Meara, 2009; Read, 2000). However, Kruse, Pankhurst and Sharwood Smith (1987) distinguish word association tests based on whether restrictions are placed on the kind of associates given and the number of associates given in response to a stimulus word. For instance, the previously described word association test format would be categorized as a single, free association test because only one response per stimulus word is required and no restrictions are placed on the types of words that can be given. In a more controlled word association test, however, participants are asked to give only associates from a given grammatical or conceptual category (Kruse et al., 1987). Additionally, Kruse and colleagues (1987) distinguish between the *continued* and *continuous* methods of eliciting associates from participants. In continued elicitation, “the stimulus is presented to the subject several times, and each time the subject gives only one response” (Kruse et al., 1987, p. 143), and in continuous elicitation, “the stimulus is presented only once and the subject is asked to give a number of responses in a limited period of time” (Kruse et al., 1987, p. 143). Regardless of the specific format used, however, word association tests have the benefit of being relatively quick to construct, administer and score (Fitzpatrick, 2000; Wolter, 2002).

Although word association tests have traditionally been used in psychological research and clinical settings, language researchers have adopted this method for examining L2 proficiency, the nature of the associates given by native and non-native speakers, the development and organization of the mental lexicon, changes in the pattern of associates as proficiency increases and depth of word knowledge, i.e., how well words are known (Fitzpatrick, 2007; Kruse et al., 1987; Politzer, 1978; Read, 1998, 2000; Riegel & Zivian, 1972; Söderman, 1993; Sökmen, 1993; Wolter, 2002). As a measure of productive vocabulary size, however, the word association format may be an especially attractive option because of its potential to overcome some of the problems associated with typical measures of productive vocabulary size. For instance, rather than targeting pre-selected items, as the PVLT (Laufer & Nation 1999) does, the word association format encourages fairly spontaneous production with minimal involvement of receptive skills and little restriction by context, since participants simply write down any word or words that come to mind after reading a given stimulus word. Additionally, since the stimulus words in word association tests tend to be open-class content words, it is unlikely that participants would produce closed-class function words if only a single related associate is required. Similarly, if multiple associates are to be given, as in a continuous method, it is unlikely that function words would occur with the same frequency as they do when vocabulary is elicited through written texts. The frequency of occurrence of function words is a major issue for tests like the LFP (Laufer & Nation, 1995), where these types of words account for, according to Nation (2001), approximately 43% of most texts. These features of the word association format were exploited by Meara and Fitzpatrick (2000) and by Fitzpatrick (2003) in her unpublished doctoral thesis that described the development of the Lex30, a test of productive vocabulary size.

**The Lex30**

The Lex30 (Fitzpatrick, 2003; Meara & Fitzpatrick, 2000) is a test of productive vocabulary size that has managed to circumvent a number of the limitations of the other measures of this construct. It is an easily constructed continuous word association task that imposes fewer restrictions on participants' production, requires minimal reliance on receptive resources and elicits fairly large quantities of words in a relatively short amount of time (Meara & Fitzpatrick, 2000; Meara, 2009; Milton, 2009; Fitzpatrick & Clenton, 2010). Additionally, in his comparison of the findings gained from certain measures of productive vocabulary, i.e., the Lex30, LFP and PVLT, Clenton (2008) reports that the Lex30, which calls for no grammatical knowledge and only minimal reliance on semantic knowledge, appears to be the closest approximation to a measure of exclusively productive vocabulary.

In the original Lex30, participants are given a series of 30 stimulus words, drawn from the first 1000 most frequent lemmas in Nation's (1984) word list which do not elicit stereotypical or highly frequent associates. These stimulus words are varied and can activate a wide range of concepts, thereby decreasing the context specificity of the test (Fitzpatrick & Clenton, 2010). In keeping with a continuous response word association format, participants' task is simply to write down at least 4 words that come to mind when they read each of the stimulus words. In such a task, researchers can more effectively gain access to a range of an individual's productive lexicon since there is no predetermined set of correct responses for participants to produce, nor is there one over-arching context for participants to consider when producing words (Meara & Fitzpatrick, 2000; Meara, 2009; Milton, 2009). The data is then lemmatized and scored based on the word frequency of the lemmas such that participants receive one point for each item located in Nation's (1984) 2000 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. This is interpreted as a sign of a larger overall productive vocabulary because the underlying assumption is that frequent vocabulary items are acquired before infrequent ones so that those with larger lexicons are more likely to have access to a greater number of infrequent words (Fitzpatrick, 2003; Meara, 2009; Meara & Fitzpatrick, 2000).

In the initial test of the Lex30, Meara and Fitzpatrick (2000) recruited 46 adult English as a Foreign Language learners, ranging in proficiency from upper-elementary to advanced. Participants were asked to complete the Lex30 and the yes/no Eurocentres

Vocabulary Size Test (EVST), a measure of receptive vocabulary knowledge (Meara & Jones, 1990). For the Lex30, participants were presented with a task sheet on which 30 high frequency stimulus words were written. The entire test lasted a total of 15 minutes, during which time a test administrator called out each word one at a time and participants were given 30 seconds to write down associates to the word that was called out. The stimulus words were presented orally and in written form to increase the chances of participants recognizing the word and to prevent them from spending too much or too little time on a given stimulus word (Fitzpatrick, 2003). For the EVST, participants simply saw a series of words and indicated whether or not they knew those words. Analysis revealed a significant positive correlation between EVST and Lex30 scores, such that participants with a large receptive vocabulary tended to produce a greater number of infrequent items in the Lex30, $r = 0.841$, $p < .01$.

Other tests of the Lex30 have also helped to confirm its reliability and validity. For instance, Fitzpatrick and Meara (2004) found no significant difference between the Lex30 scores of 16 L2 speakers of English who completed 2 separate administrations of the test, three days apart, $t = 1.58$, $p = .135$, and a significant positive correlation between the two sets of scores, $r = .866$, $p < .01$. Similarly, Fitzpatrick and Clenton (2010), who had 103 low-intermediate to advanced learners take the Lex30 test twice, one week apart, found Lex30 scores at Time 1 (M = 21.30, SD = 11.75) and Time 2 (M = 23.90, SD = 10.51)[*] to be similar and highly correlated, $r = .84$, $p < .0001$. Further, Lex30 scores appear to be stable even though the actual associates provided at Time 1 and Time 2 are different. Fitzpatrick and Meara (2004), as well as Fitzpatrick and Clenton (2010), found

[*] Statistical tests of the significance of this difference were not reported.

that all participants tended to produce different words at Time 2 regardless of the fact that they were associating to the same stimulus words they encountered at Time 1. The Lex30 score, however, which is an indication of the number of infrequent words provided, remained the same. This was interpreted as an indication that the Lex30 elicits lexis that is fairly representative of the current state of an individual's mental lexicon, since the proportion of infrequent words that an individual is capable of supplying, is constant, regardless of the fact that different words are supplied across time (Fitzpatrick & Meara, 2004).

Parallel forms reliability tests were also conducted with the Lex30. To do so, Fitzpatrick and Clenton (2010) constructed a parallel form of the Lex30, called Lex30b, which featured stimulus words drawn from the 1000 most frequent English words according to the JACET 8000 word list (JACET, 2003). The Lex30b was contrasted with the traditional Lex30 where stimulus words are drawn from Nation's (1984) word frequency list. Forty (40) Japanese learners of English completed, in written form, both versions of the Lex30 just 5 minutes apart. Analyses indicated that parallel forms of the Lex30 behave similarly since scores were significantly positively correlated, $r$ = .692, $p$ <. 01, and there were no significant differences in scores on the Lex30 ($M$ = 24.3, $SD$ = 8.514) and Lex30b ($M$ = 23.5, $SD$ = 7.923), $t$ = 0.806, $p$ = .425. Similarly, Fitzpatrick and Clenton (2010) found no significant difference between Lex30 scores when the test was administered in the written format ($M$ = 16.6, $SD$ = 8.104) and again 6 weeks later in spoken format ($M$ = 15.6, $SD$ = 7.088), ($t$ = 0.751, $p$ = 0.457), where participants read the cue word and then spoke their responses.

In addition, evidence of the Lex30's validity came from another one of

Fitzpatrick and Meara's (2004) studies in which the Lex30 scores of 46 native English speakers and 46 non-native speakers of English were compared. Results indicated that the Lex30 was able to consistently distinguish these two groups of participants, with the native speakers ($M = 44$, $SD = 7.62$) supplying a higher percentage of infrequent words than the non-native speakers ($M = 30$, $SD = 9.34$), $t = 7.5$, $p < .001$ (Fitzpatrick & Meara, 2004). More recently, Walters (2012) also showed that the Lex30 was able to distinguish between advanced ($n = 32$, $M = 55.84$, SD = 11.71), intermediate ($n = 25$, $M = 36.72$, SD = 10.05) and high beginning ($n = 30$, $M = 27.23$, SD = 5.72) users of English, $F(2,84) = 72.59$, $p < .001$, $\omega = .99$, with post hoc Scheffé analyses confirming that the means of all groups were significantly different from each other , $p < .01$.

The concurrent validity of the Lex30 was also confirmed by Fitzpatrick and Meara (2004) who examined the nature of the relation between Lex30 scores and scores on other measures of productive vocabulary knowledge. Fifty-five (55) Chinese learners of English (intermediate to advanced) completed the Lex30, the PVLT and an L1 Mandarin to L2 English translation test. They found moderate positive correlations between the Lex30 scores and scores on the PVLT ($r = .504$, $p < .01$) and translation test ($r = .651$, $p < .01$), indicating that the Lex30 is capable of tapping the construct of productive vocabulary. However, since the correlations were modest, Fitzpatrick and Meara (2004) suggest that the Lex30 may be assessing a different aspect of this complex construct than the translation test and PVLT since the correlation between scores on these two tests were much larger ($r = .843$, $p < .01$). Walters (2012) replicates this result with even stronger correlations between Lex30 scores and the PVLT ($r = .772$, $p < .001$), and a Turkish-English translation test ($r = .745$, $p < .001$).

Taken together, these results help to establish the reliability and validity of the Lex30 as a test of productive vocabulary knowledge. As such, we intend to use the Lex30 as our comparison measure of productive vocabulary size in our attempt to validate the CR as a measure of the same construct.

## General Problem Statement

Productive vocabulary has proven to be a complex and multifaceted construct, one that has been challenging to both define and measure empirically (Fitzpatrick, 2003; Read, 2000; Schmitt, 1997). Part of the difficulty associated with the measurement of productive vocabulary knowledge stems from the fact that, as yet, no clear consensus exists on how exactly to conceptualize or operationalize this construct. As a result, estimates of productive vocabulary size vary considerably, and researchers use a variety of subtly different aspects of vocabulary knowledge, such as lexical richness (Laufer & Nation, 1995; Meara & Miralpeix, 2007) or knowledge of or access to infrequent words (Laufer & Nation, 1999; Meara & Fitzpatrick, 2000), to make inferences about productive vocabulary size as a whole.  Difficulties also arise from the widely used measures of productive vocabulary knowledge, such as the PVLT (Laufer & Nation, 1999) and the LFP (Laufer & Nation, 1995), which tend to be time inefficient, too controlled, context-dependent, assess pre-selected targets, limit test-takers to the production of one correct response, test receptive abilities also, and elicit insufficient quantities of content vocabulary from which to make meaningful inferences (Clenton. 2008; Fitzpatrick & Clenton, 2010; Read, 2000; Meara & Fitzpatrick, 2000; Milton, 2009).

The Lex30 (Meara & Fitzpatrick, 2000) has been able to overcome some of these measurement challenges since its easily constructed word association format imposes fewer restrictions on participants' production, requires minimal reliance on receptive resources or semantic knowledge and elicits fairly large quantities of words in a relatively short amount of time (Meara & Fitzpatrick, 2000; Meara, 2009; Milton, 2009; Fitzpatrick

& Clenton, 2010). However, this test may be limited by its heavy reliance on lexical frequency information in estimating productive vocabulary size, which may not always be available in many languages. It may be worthwhile, then, to focus our efforts on developing a valid and reliable test of productive vocabulary size that capitalizes on the benefits of the word association task, as the Lex30 does, but which does not rely on lexical frequencies in estimating productive vocabulary size. The Capture-Recapture (CR) methodology, borrowed from Ecology and recently advocated by Meara and Olmos Alcoy (2010), may be such a test. The goal of the current work is to investigate this possibility. What follows in the next chapter is a manuscript-based account of one study designed to examine the validity of the CR technique as a measure of productive vocabulary size.

## The Current Work

The goal of the current work is to examine the validity of the Capture-Recapture (CR) technique as a measure of productive vocabulary size. Instead of using written texts to elicit vocabulary from our participants, as Meara and Olmos Alcoy (2010) have done, we propose as our trapping procedure, a continuous word association task that is similar in setup to the Lex30. This decision is advantageous for a number of reasons. First, it allows us to avoid a number of the problems typically associated with the administration of productive vocabulary tests. Secondly, the use of the word association task in the current work also allows us to score the data based on the logic of (1) the traditional Lex30, which rewards participants for the amount of low frequency words given, and (2) the CR technique, which rewards participants for the amount of unique words given during both captures. A third benefit of using the continuous word association task as our trapping procedure is that it appears to stimulate participants to produce a variety of different words each time they complete the task (Fitzpatrick & Clenton, 2010; Fitzpatrick & Meara, 2004). Indeed, even though Fitzpatrick and Clenton's (2010) participants associated to the same stimulus words during the two separate administrations of the Lex30 word association task, only about 41% of words were repeated. While, for the purposes of the CR technique, this percentage of repeated words is high, perhaps changing the stimulus words at Time 2 will serve to decrease the number of associates that are common to both 'captures' by encouraging participants to access a range of items in their lexicon. If that is the case, the number of repeats that go into the denominator of the Petersen formula will not be artificially high and lead to a lower vocabulary size estimate. This will be explored in the current work.

Finally, a further benefit of using the continuous word association task for eliciting vocabulary in the current work is that results based on the traditional Lex30 scoring and the CR scoring will be more comparable. Fitzpatrick (2003) makes the point that it is difficult to compare the results of different tests of productive vocabulary since they can all claim to measure different aspects of this complicated construct. Further, exploration of an individual's lexicon can involve analysis of at least three different aspects of that lexicon, i.e., the quantity of the items it contains, the extent to which those items are known and the nature of their organization in the lexicon (Fitzpatrick, 2003). Attempting to validate a measure of the quantity of the lexicon by comparing its performance to a measure that assesses how well lexical items are known may be misleading. Thus, by using the word association format to elicit vocabulary and just scoring the data in two different ways, we can be more confident that we are assessing and comparing performance on the same aspects of productive vocabulary knowledge, and making inferences about the same aspect of the lexicon, in this case the quantity of items it contains.

The proposed CR methodology will be deemed valid if the following validity criteria are met (Thorndike & Thorndike-Christ, 2010):

- Convergent validity – Since the data will be scored using both the traditional Lex30 scoring and the CR scoring, we will be able to determine the extent to which the CR correlates with the widely used Lex30. Our first hypothesis (H1) is that the CR and the Lex30 scores will be significantly positively correlated.
- Construct validity – Since the CR is intended as a measure of productive vocabulary size, scores obtained from this method should distinguish between the

L1, where vocabulary size should be larger, and the L2. As such, our second hypothesis (H2) is that the CR scores will be larger in the L1 (English) than in the L2 (French). Additionally, since cognitive efficiency is a crucial component of fluency, our third hypothesis (H3), which is also related to construct validity, states that a significant negative correlation is expected between the CR vocabulary size measure and performance on a semantic categorization task designed to assess the speed and efficiency of lexical access (Segalowitz, 2010).

It should be noted that since the word association and semantic classification tasks were completed in the L1 and L2, we were able to use residualized L2 scores in all our analyses (except for the construct validity test described in H2 above, since residualized L2 scores cannot be compared with unresidualized L1 scores). These residualized scores reflect second language performance that is statistically independent of first language performance and give a purer indication of second language vocabulary size and efficiency (Segalowitz, 2010). To our knowledge, using participants' own L1 scores as baseline measures, or controlling for them in this way, has never been done in previous vocabulary studies.

# CHAPTER 2

## Manuscript

Researchers often distinguish between two, positively correlated aspects of vocabulary knowledge - receptive or passive vocabulary and productive or active vocabulary (Laufer, 1998; Webb, 2008), which is of interest here. Receptive vocabulary refers to those lexical items that an individual can recognize and understand when listening to speech or reading text, while productive or active vocabulary refers to the set of vocabulary items that an individual can produce accurately when speaking or writing (Milton, 2009; Schmitt, 2010). Empirical evidence 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). The unique challenge associated with measuring productive vocabulary size, in particular, has prompted researchers to investigate the construct in increasingly creative ways. Accordingly, the current work attempts to estimate L2 productive vocabulary size using a novel approach, known as the Capture-Recapture (CR) methodology.

### Challenges in Measuring Productive Vocabulary

The difficulty in measuring productive vocabulary knowledge stems partly from the lack of consensus surrounding a conceptualization of the construct. This has also contributed to challenges in interpreting and comparing test results, since a variety of techniques have been used, e.g., translation tests, gap-fill tasks, or word association tests, to collect subtly different information about productive vocabulary knowledge e.g., lexical richness (Laufer & Nation, 1995; Meara & Miralpeix, 2007) or knowledge of and access to infrequent words (Laufer & Nation, 1999; Meara & Fitzpatrick, 2000), that may not be directly comparable,. Furthermore, the measures of L2 productive vocabulary knowledge that are widely used tend to be time inefficient, too controlled, context-dependent, assess pre-selected targets, limit test-takers to the production of one correct response, assess receptive abilities also, and elicit insufficient quantities of content vocabulary from which to make meaningful inferences (Clenton, 2008; Fitzpatrick & Clenton, 2010; Read, 2000; Meara & Fitzpatrick, 2000; Milton, 2009).

For instance, Laufer and Nation's (1995) Lexical Frequency Profile (LFP), designed to assess lexical richness, requires learners to write two essays of 300-350 words each, one on a general issue, and the other on a controversial issue. A lexical frequency profile for each learner is then created by computing the proportion of word families in the first and second 1000 most frequent words, the University Word List (UWL) and off-list items (Laufer & Nation, 1995). In addition to being fairly time consuming (1 hour per composition), the LFP requires the production of fairly lengthy texts on two topics that may not encourage participants to use vocabulary that is representative of learners' lexicon (Fitzpatrick, 2003; Meara & Fitzpatrick, 2000).

The problem of context dependence is also associated with the Productive Vocabulary Levels Test (PVLT; Laufer & Nation, 1999), which requires learners to read a sentence and complete the target word. The first letters of the target word are provided to rule out other semantically viable options that are not being tested. This test samples 18 items from each of the 2000, 3000, 5000, UWL and 10 000 word levels and a score for the number of correct items at each word level, and overall, is calculated. Aside from the fact that production is limited to only one correct answer, 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, providing as many initial letters as necessary to effectively disambiguate the target 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 be possible to draw conclusions that are specific to productive knowledge since receptive abilities are also required to consider the context of the sentence and the number of initial letters provided (Fitzpatrick, 2003; Fitzpatrick & Clenton, 2010; Read, 2000).

## Capture-Recapture Methodology and Petersen Estimate

Since the nature of what is being measured by these tests remains unclear and the results gathered from them are difficult to interpret, Meara and Olmos Alcoy (2010) advocate investigating the construct of productive vocabulary from "different, perhaps unconventional points of view" (p. 223). Along those lines, they investigated whether the Capture-Recapture methodology (CR), which is commonly used in Ecology 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. Second, the ecologist will capture 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 will be 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 will take his second sample of fish (Time 2) using the same method as at Time 1. A count of the total number of fish captured at Time 2 will be obtained, along with a count of the number of marked fish from Time 1, which also appear in the Time 2 capture. 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. In order to estimate the total population of fish in the river ($P$), the ecologist then plugs these 3 values into a formula known as the Petersen Estimate ($P = xy/r$; Petersen, 1896).

In order for the Petersen formula to provide meaningful estimates, a number of assumptions must be met. First, the capture method used should provide a good chance of capturing whatever it is we intend to measure, be it fish in a river or vocabulary in the mental lexicon. Second, in keeping with the fish analogy, the stretch of river from which sample are taken must be representative of the river as a whole. Third, conditions at the two captures should be equivalent and each animal should have an equal chance of being captured at both times (Meara & Olmos Alcoy, 2010).

### Meara and Olmos Alcoy's (2010) Study

In order to explore whether this ecological approach could be adopted to estimate L2 productive vocabulary size, Meara and Olmos Alcoy (2010) recruited 24 native speakers of English, who were intermediate (*n* = 11) learners 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 about an incident by the sea involving a lost umbrella, two men, a boy and a dog (See Figure 1). This procedure was completed two times, one week apart.

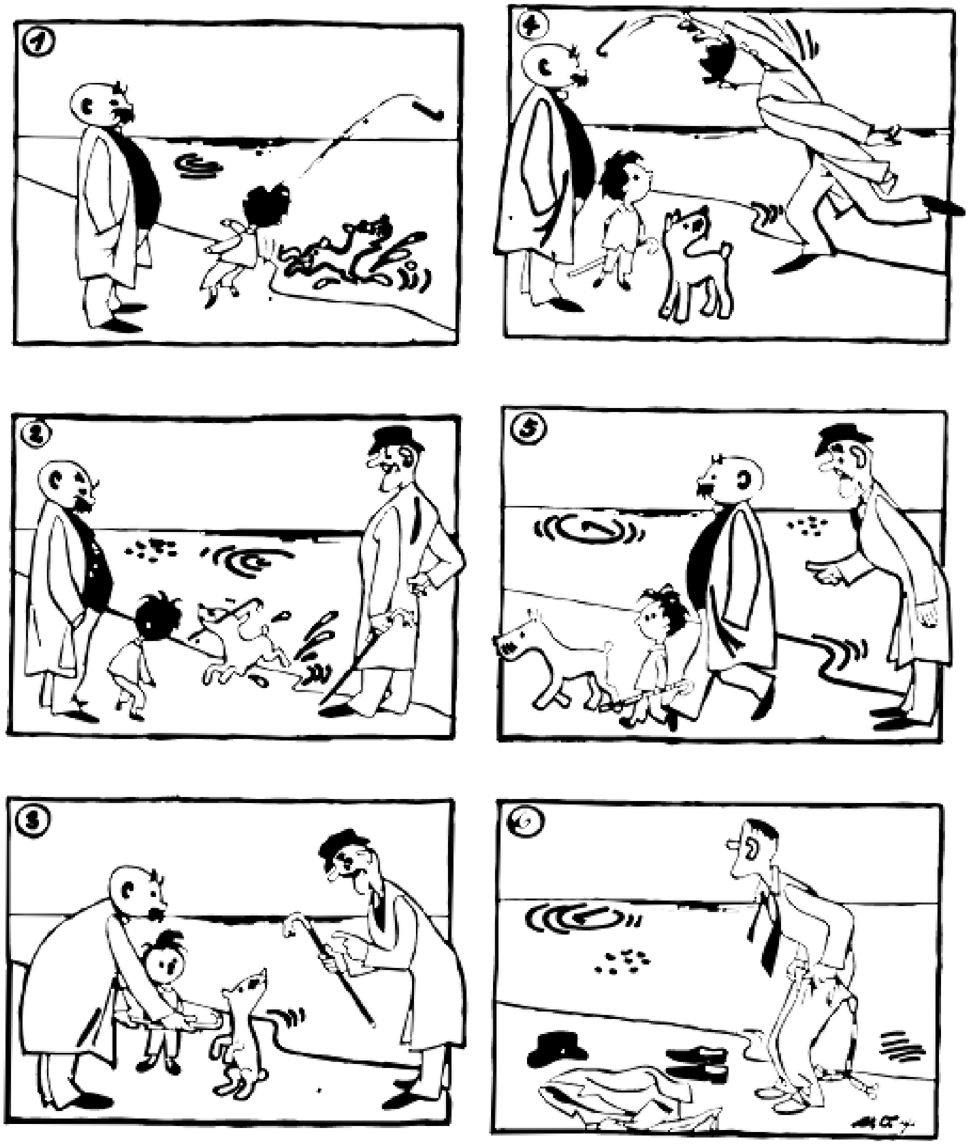

*Figure 1.* Picture Story used by Meara and Olmos Alcoy (2010)

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. Meara and Olmos Alcoy (2010) found that, at both time points, as well as overall, the advanced group supplied significantly more word tokens and types in their stories than did the intermediate group. Additionally, a Mann-Whitney *U* test confirmed that the Petersen estimate of productive vocabulary size reliably distinguished between the intermediate (*M* = 93.81, *SD* = 31.30) and advanced groups (*M* = 160.37, *SD* = 38.51), *U* = 9.5, *p* < .01. 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 holds some promise as a measure of productive vocabulary size, Meara and Olmos Alcoy's (2010) choice of trapping instrument may not have been ideal. The writing task likely violated the assumptions of representativeness in sampling and items having equal probabilities of being sampled since the context may not elicit lexis that is representative of learners' productive vocabulary as a whole, and the words necessary to describe the events depicted in the picture story have a greater chance of being captured and recaptured, than 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 greatly increases the number of 'repeat' items, which is the denominator in the Petersen estimate formula. Indeed, repeats were quite high in Meara and Olmos

Alcoy's (2010) study since, for the Advanced group, 45.50% of the word types produced at Time 2, also occurred at Time 1, while for the Intermediate group, 49.31% of the word types produced in the Time 2 narrative, were also produced at Time 1. The use of an inappropriate 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 their estimates, 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.

Meara and Olmos Alcoy (2010) acknowledge these limitations and suggest that a more appropriate trapping procedure would elicit a fairly large number of words during both captures, without increasing the likelihood of words overlapping across captures. They speculate that the continuous word association format, used in the Lex30 (Meara & Fitzpatrick, 2000) test of productive vocabulary size, might be able more suitable. Not only are word association tasks relatively quick to construct, administer and score (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). Participants are given a series of 30 stimulus words which do not elicit stereotypical or highly frequent associates, and which are drawn from the first 1000 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 4 words that come to mind in response to each stimulus word encountered. The data is then lemmatized and participants receive one point for each lemma located in Nation's (1984) 2000 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, 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 measure 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 as we investigate its validity as a measure of productive vocabulary size.

It should be noted, however, that it is possible that the Lex30's reliance on lexical frequency information may limit its applications since this information is not always readily available in many languages. There is need, then, for a valid and reliable test of productive vocabulary size that capitalizes on the benefits of the word association task, as the Lex30 does, but which does not rely on lexical frequencies in estimating productive vocabulary size. The possibility that the Capture-Recapture (CR) methodology is such a test will be investigated in the current work.

## The Current Work

The goal of the current work is to examine the validity of the CR technique as a measure of productive vocabulary size. Instead of using written texts 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 allows us to avoid many of the problems associated with measures of productive vocabulary size. Additionally, the word association format allows us to score the same data based on the logic of (1) the traditional Lex30, which rewards participants for the amount of low frequency words given, and (2) the CR technique, which rewards participants for the amount of unique words given during both captures. Results will then be more comparable since the difference between the two is in scoring, not in the type of data collected, or the way in which it was collected.

The proposed CR methodology will be held as valid if convergent and construct validity criteria are met (Thorndike & Thorndike-Christ, 2010). Specifically, the convergent validity of a test is established when it correlates with an already validated measure of the same construct. As such, hypothesis 1 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. As such, hypothesis 2 is that the CR scores will be larger in the L1 than in the L2. Furthermore, since cognitive efficiency is a crucial component of fluency, which is also undoubtedly influenced by vocabulary size, hypothesis 3, which also relates to construct validity, is that a significant negative correlation exists between CR scores and performance on a semantic categorization task that assesses the speed and efficiency of lexical access (Segalowitz, 2010). A negative correlation is predicted because speed is represented by reaction times in milliseconds and the efficiency of lexical access is represented by the coefficient of variation (CV), defined as the standard deviation divided by 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 report 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 that the differences between English and French self-ratings of abilities on all four language skills were found to be significant, 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: $T = 0$, $Z = -5.93$, $p < .001$. 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 stimuli 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.

**Living-Nonliving task (LNL; Segalowitz, 2010).** The LNL is a computerized semantic classification task that measures English and French cognitive fluency, which refers to the ease and stability with which cognitive processes are conducted. Following a brief training session, participants completed the main task in both English and French 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/ un chien) or a *nonliving* thing (e.g., a bed/ un lit). 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. Each word was presented until a response was made or for a maximum time of 3000 milliseconds (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. In both languages, the stimulus words were presented on the screen with the appropriate definite or indefinite articles (English: *the, a*; French: *le, la, un, une*). There were a total of 60 trials in both the English and French tasks, the first 12 of which were warm up trials, while the remaining 48 were the experimental trials. Response times for correct trials were recorded and the coefficient of variability (CV), a measure of the stability and efficiency of responses, was computed using the formula, *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.

## 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 and second, in their L2, French. They were given 15 minutes in each language to write down at least 4-6 associates to each of 30 high frequency 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 in English and 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.

### 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 affixes included in Table 4 were treated as instances of their base lemmas. Words with affixes that do not appear in Table 4 were not lemmatized, and were treated as separate words (Meara & Fitzpatrick, 2000).

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 compose*, *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.

Table 4: *Level 2 and 3 Affixes for Lemmatization*

| Level 2: Inflectional Suffixes | Level 3: Most Frequent and Regular Derivational Affixes |
|---|---|
| • Plural | • -able not when added to nouns |
| • 3rd person singular present tense | • -er |
| • past tense | • -ish |
| • past participle | • -less |
| • -ing | • -ly |
| • comparative | • -ness |
| • superlative | • -th cardinal-ordinal only |
| • possessive | • -y adjectives from nouns |
|  | • non- |
|  | • un- |

*Note.* Adapted from Meara and Fitzpatrick (2000)

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* would be 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 word provided. As such, one point was assigned to each English and French word that falls beyond the 2000 most frequent words in English, according to Davies and Gardner (2010), and in French, according to 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. In order to statistically control for L1 performance and other nuisance variables, all L2 scores were residualized, i.e., regressed against their equivalent L1 score. These residualized scores give a purer indication of second language vocabulary size and efficiency (Segalowitz, 2010). As such, wherever possible, results based on residualized L2 scores will be 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.

Descriptive statistics of the number of lemmas generated at Time 1 and 2 in English and French are included in Table 5. This table suggests that the word association format itself is capable of distinguishing languages since, at both times, participants supplied more lemmas in their L1 than in their L2. The non-parametric Wilcoxon signed-rank test confirmed that significantly more lemmas were generated in English than in French at both 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 5 also suggests that, relative to Time 1, participants generated more lemmas at Time 2 in their L1 and L2. 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, $t(46) = -4.31$[‡], $p < .001$, $r = .54$, respectively. As

[†] The Pearson *r* will be used as the effect size statistic in the current work.

[‡] 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.

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.

Additionally, Table 5 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.

Table 5: *Descriptive Statistics for the Raw Lemma counts, Repeats, Reaction Times and CV scores in English and French*

| | English | | | French | | |
|---|---|---|---|---|---|---|
| Variables | *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 | 666.89 | 78.41 | 701 | 728.36 | 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 in milliseconds on the LNL task in English were compared to their French equivalents to determine whether the expected pattern of results (lower RTs in the L1) would be found. A dependent *t*-test revealed that RTs were indeed lower in the L1 ($M$ = 666.89, $SD$ = 78.41), relative to the L2 ($M$ = 728.36, $SD$ = 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 ($r_s = .57$, $p < .0001$) 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.

## Testing 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 6.

Table 6: *Descriptive Statistics of the CR and Lex30 Vocabulary Size Estimates*

| | English | | French | |
|---|---|---|---|---|
| Variables | *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.

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 ones, like the L1 vocabulary scores. As can be seen in Table 6, participants' average CR scores were indeed higher in English than they were in French, and the Wilcoxon signed-rank test confirmed that this L1 (*Mdn* = 889.97)-L2 (*Mdn* = 606.67) difference in CR scores 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 (*M* = 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 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 lexical access. This hypothesis was partially supported. Spearman correlations revealed that, in English, CR scores were not correlated with speed ($r_s$ = -.25, $p$ = .09) or efficiency ($r_s$ = -.001, $p$ = .86) of lexical access. However, 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).

We also examined whether Lex30 scores would correlate positively with the speed and efficiency of lexical access. Spearman's rho indicated that the English Lex30 scores were not correlated with RTs on the LNL task ($r$ = -.24, $p$ = .10) or CV scores ($r$ = -.05, $p$ = .73). However, 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).

## Discussion

In this section, discussion of the results of this study will center 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 Productive Vocabulary Size?

The goal of the current work was to provide validity evidence for the CR methodology as a measure of 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, will also have access to a greater number of infrequent words, as measured by the Lex30[§]. However, the magnitude of the relation between these two measures of productive vocabulary size suggests that they may be giving different, but complimentary, 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 (1983) Vocabulary Level's

[§] This finding also helps to confirm the assumption of the Lex30 that individuals with larger lexicons are more likely to have access to a greater number of infrequent words.

Test, a precursor to the PVLT, and correlation coefficients reported ranged from .6 to .8 (with *p* values below .0002), 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 distinguish between the L1, where vocabulary size is larger, and the L2. 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 groups are distinct from each other, such as between native speakers and language learners (Fitzpatrick & Meara, 2004), or individuals with clearly different amounts of experience in their second language (Walters, 2012). Capturing intraindividual differences may pose a challenge for the Lex30 because of individuals' stable response tendencies across time. More specifically, we found significant positive correlations between English and French Lex30 scores (*r* = .51, *p* < .0001), 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 in their L1 and L2 with regards to acquiring infrequent vocabulary items and/or producing them as associates in word association tests, both of which may influence the Lex30's ability to pick up intraindividual L1-L2 differences. The CR, on the other hand, which shows only a trending relation between L1 and L2 scores (*r* = .26, *p* =.08) has shown itself to be sensitive to L1-L2 differences in productive vocabulary size despite any response tendencies common to both languages and despite the moderate differences participants report in their first and second language abilities.

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 English, almost no relation between CR scores and the speed (RTs) and efficiency (CV) of lexical access was observed, indicating that the size of an individual's productive vocabulary does not influence how fast semantic decisions are made or how efficiently cognitive processes are carried out in the L1. This same pattern was observed with the English Lex30 scores. These results are unexpected since cognitive fluency is a crucial component of overall fluency in a language, which is itself also influenced by vocabulary size. It is possible that our inability to find this result in English was due to ceiling effects or range restrictions operating in the L1, where participants' performance tended to be less variable than their performance in their L2. On the other hand, since both the CR and Lex30 show no relation to the cognitive fluency measures in English, we can also speculate that there is something fundamentally different about how vocabulary size, and the speed and efficiency of lexical access operate in the L1 system that prevents a relation among these three variables from being observed. Perhaps the native language is so rehearsed that the size of the L1 lexicon is not actually a crucial component of cognitive fluency. Unfortunately, though, the

[**] Negative correlations were predicted because greater CR scores were expected to correlate with lower RTs and CV scores, indicating faster and more efficient cognitive processing.

methods used in the current work were not sensitive enough to allow us to make definitive conclusions of this nature.

In French, on the other hand, a significant negative relation was observed between residualized CR scores and the speed of lexical access, indicating that participants with higher productive vocabularies tended to respond faster, as evidenced by lower RTs, on the lexical decision task. Interestingly, however, CR scores were not correlated with CV scores. French Lex30 scores showed a similar significant negative correlation with speed, but no relation to the efficiency of lexical access. Since both 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 associated 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 cognitive efficiency.

### 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, which states 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 was preferable to the continuous writing 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 our study supplied far more usable data in their first and second language at Time 1 and Time 2 (see Table 5) 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 their writing 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 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, a large significant positive correlation between the raw number of lemmas supplied at Time 1 and Time 2 ($r_s$ = .81, $p$ < .0001) was observed, indicating that the number of words participants can generate on the word association task is similar across time, despite the fact that different stimulus words were used at both times. 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, although significant. When the correlation coefficients are squared, we see that roughly 10% and 23% of the variance in CR scores is accounted for by the number of words given at Time 1 and Time 2, respectively. Similarly, in French, a large significant positive correlation between the raw number of lemmas supplied at Time 1 and 2 ($r_s = .87$, $p < .\ 0001$) was also observed, and the correlations between the French CR score and the raw number of lemmas supplied at Time 1 ($r_s = .72$, $p < .0001$) and 2 ($r_s = .73$, $p < .0001$) in French indicated that about 52% and 53% of the variance in CR scores was accounted for by the number of lemmas generated at Time 1 and Time 2, respectively. The CR, then, appears to be more than just the sum of its parts and may 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. So, what exactly does the CR score tell us, then?

Since the CR estimates are far larger than either raw count, it tells us that participants 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. Nation (2001) cites the results of two recent studies (Goulden, Nation & Read, 1990; Zechmeister, Chronis, Cull, D'Anna & Healy, 1995) which estimate that educated adult native speakers of English (like the university students who participated in this study) know, in a primarily receptive sense, around 20 000 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 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 the 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 a 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.

## 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 and was ideal for the CR technique since 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 complimentary, 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. However, as far as interpretation of the CR estimate is concerned, it may be more appropriate for us to limit the scope of our generalizations. Specifically, instead of interpreting the CR estimate as a direct indication of the size of an individual's productive vocabulary, perhaps it should be interpreted as an indication of the number of words an individual has available to them to complete a certain task under the specific task requirements encountered. This may be the indication of an overall larger productive vocabulary size. Further study is needed to explore the validity of this interpretation, replicate these results and refine the paradigm.

# CHAPTER 3

## Extended Methodology

### Participants

In total, 52 English-French bilingual university students participated in this study. However, the results presented in the manuscript were based on a final sample size of 47 because five participants were excluded from analysis, four because of high error rates on the LNL task, and one because of numerous outlier scores. In keeping with the multicultural nature of Montreal, which makes it difficult to recruit exclusively bilingual participants, 22 of the final 47 participants reported having basic to intermediate knowledge of a third language – Spanish (*n* = 10), Italian (*n* = 4), Hebrew (*n* = 1), Mandarin/Cantonese (*n* = 2), Greek (*n* = 1), Yiddish (*n* = 1), Polish (*n* = 1), German (*n* = 1), and Arabic (*n* = 1). Participants were recruited from Concordia University through the Psychology department's participation pool website, and from McGill University through ads posted on the McGill Classifieds website. Psychology students were given course credit for their participation, while McGill students, or those answering the McGill classified ad, received $10 after each testing session. The consent form signed by participants is included in Appendix A.

### Materials

**Language Background Questionnaire.** This is a paper-and-pencil questionnaire (see Appendix B) designed to establish participants' eligibility for the study, and gather demographic information, language learning history and self-reported estimates of the percentage of time spent interacting with people in the first and the second language. Participants also rated their proficiency in English and French speaking, reading, writing, and listening using Likert type scales ranging from 1 (*no ability at all*) to 5 (*fluent ability*).

**The Continuous Word Association Task.** Four versions of the word association task were created in both English (Appendix C) and French (Appendix D), each with a different set of 30 high frequency stimulus words. One group of participants completed Versions A and B at Time 1 and Time 2, respectively, while another group completed Versions C and D at Time 1 and 2, respectively. Instructions for the French and English word association tasks are included in Appendix E.

**Living-Nonliving task (LNL; Segalowitz, 2010).** Word stimuli for the Living-Nonliving task (see Appendix F) were presented on a 12 inch iMac computer (1024 X 768 resolution; 700 MHz Power PC G4) and displayed on a white background using MATLAB® (MathWorks, 2007) software. Before the main task, participants completed a 54-trial training session in their L1, English, to familiarize themselves with the mechanics of the task. On each training trial, participants were presented with a word in the center of the computer screen and pressed the appropriate button on a controller to indicate whether the word was a color word (e.g., "red") or a number word (e.g., "two"). Following training, participants completed the main task described in the manuscript. When an error was made, participants received audible feedback from the computer (a beep), and an additional 450 ms interval was inserted before the next stimulus was presented.

## Data Treatment

Commonly used abbreviations were converted to their long forms, e.g., *tv* to *television*, *bday* to *birthday* and *intro* to *introduction*. If participants produced the same word more than once *within* a given testing session, only one of those words were counted. In both French and English, the following words were excluded from the T1 and T2 counts and from all analyses:

- function words, including prepositions (e.g., *in*/*dans*), pronouns (e.g., *he*/*il*), conjunctions (e.g., *that*/*que*)
- proper nouns, including months of the year (January/janvier), days of the week, (e.g., Monday/lundi), cities, countries or nationalities (e.g., *Paris*, *Canada* or *American/américain*), and names of religions (e.g., Christian/chrétien)
- acronyms (e.g., *USA*, *SIDA*),
- onomatopoeia (e.g., *ouch*),
- number names (e.g., *nine/neuf*),
- holidays (e.g., *Christmas*/*Noël*), and
- brand names and Games (e.g., *Microsoft* or *Monopoly*).

# Extended Results

## Data Integrity

**Dealing with outliers.** For each participant, reaction times on any given trial in the LNL task that were 3 standard deviations or more above their individual mean were excluded from analysis. Additionally, data from participants who made errors on 20% or more of the English or French LNL trials were excluded from all analyses. Four participants fit this description and were removed.

For the purposes of analysis, an outlier was defined as any value that was found to be 3 standard deviations or more above or below the mean of the variable in question. Any score, or scores, that fit this description were transformed to the next highest score plus or minus one unit. For example, if a participant's Lex30 score of 90 was found to be 3 standard deviations *above* the mean, and the next highest Lex30 score was 76, the outlier score would be replaced with a value of 77, one unit above the next highest, non-outlier score on this variable. Similarly, if a participant's Lex30 score of 12 was found to be 3 standard deviations *below* the mean, and the next lowest non-outlier score was 20, the outlier score would then be replaced with a value of 19. The English CV score of one participant, the French CV score of a different participant, and the Lex30-T2 score from yet another participant required this transformation. Data from one other participant were excluded from all analyses because his scores were identified as outliers on almost all of the variables of interest, i.e., mean reaction time on the LNL task in English, English and French CR scores, and English and French CV scores.

**Checking for normality.** The Shapiro-Wilk (*W*) test was used to examine whether the assumption of normality was met in the data. This test compares the scores gained from our sample to those of a normal distribution that has the same mean and standard deviation as that observed in our sample (Field, 2009). Thus, significance on the Shapiro-Wilk test indicates substantial deviations from normality. As can be seen in Table 7, this analysis revealed that a number of the variables relevant to our hypotheses were significantly non-normal. Skewness and kurtosis values for these variables were converted into *z*-scores (see Table 7) and compared against the known values of the normal distribution, such that an absolute value greater than 1.96 represents significant skew or kurtosis at the $p < .05$ level. This was observed in our sample.

Rather than transforming these data, we chose to conduct analyses using non-parametric statistical tests that do not assume normality and which are also robust in the presence of outliers (Field, 2009). As such, instead of conducting Pearson correlation tests, non-parametric Spearman correlations ($r_s$) were used to test whether Lex30 and CR scores are positively correlated, as well as how these variables relate to cognitive efficiency, i.e., CV scores. Additionally, in place of the paired samples *t*-test, its non-parametric equivalent, the Wilcoxon signed-rank Test (Wilcoxon, 1945), was used to determine whether CR scores were capable of distinguishing between participants' L1 and L2. As convention dictates, median (*Mdn*) values are reported along with the results of any non-parametric test performed (Field, 2009).

Table 7: *Skewness, Kurtosis and Shapiro-Wilk Statistic for the Variables of Interest*

| Variables | Standardized Skewness | Standardized Kurtosis | *W*[a] |
|---|---|---|---|
| | English (L1) Scores | | |
| Lex30-T1[b] | 1.71 | -0.25 | .96 |
| Lex30-T2[c] | 2.05 | -0.69 | .92** |
| CR | 5.08 | 7.09 | .86*** |
| Speed (RT) | 1.69 | -0.31 | .95 |
| Efficiency (CV) | 2.64 | 0.80 | .93** |
| | Raw French (L2) Scores | | |
| Lex30-T1[b] | 0.67 | -1.43 | .96 |
| Lex30-T2[c] | 1.29 | -0.08 | .98 |
| CR | 5.89 | 8.59 | .83*** |
| Speed (RT) | 1.52 | -0.52 | .95 |
| Efficiency (CV) | 4.17 | 3.49 | .88*** |
| | Residualized L2 Scores | | |
| Lex30-T1[b] | 0.29 | -1.49 | .97 |
| Lex30-T2[c] | -0.08 | -1.00 | .97 |
| CR | 5.86 | 9.12 | .84*** |
| Speed (RT) | 1.80 | 1.18 | .97 |
| Efficiency (CV) | 3.67 | 4.52 | .90** |

[a] In all cases, *df* = 47. [b] Lex30 score at Time 1. [c] Lex30 score at Time 2
* *p* < .05. ** *p* < .01. *** *p* < .001.

## Preliminary Analyses

**Days between testing.** There was considerable variability in the number of days between the first and second testing session. Twenty-nine (29) participants completed the two testing sessions between 1 and 5 days apart (Group 1), while the remaining 18, completed the two testing sessions between 6 and 11 days apart (Group 2). Mann-Whitney *U* tests revealed no significant differences between these two groups of participants in their performance on the word association task at Time 2 (See Table 8). This result is especially important for CR scoring, since it suggests that participants who completed the two word association tasks within a short period of time, were not more likely to repeat words across time than those who had had more time between testing sessions.

Table 8: *Results of the Mann-Whitney U Test Comparing the Performance of Group 1 (1-5days) and Group 2 (6-11 days) on the Variables Relevant to Vocabulary Size*

| | Group 1 | | Group 2 | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *M* (*SD*) | *Mdn* | *M* (*SD*) | *Mdn* | *U* | *Z* | *p* |
| | | | L1 Scores | | | | |
| Raw Lemmas-T2[a] | 159.59 (33.11) | 150 | 155.89 (30.92) | 150.50 | 242.50 | -.41 | .69 |
| Lex30-T2 | 53.62 (20.68) | 46 | 49.56 (20.87) | 42.50 | 224 | -.81 | .43 |
| Repeats [b] | 26.14 (11.33) | 26 | 25.06 (8.00) | 24.50 | 245.50 | -.34 | .74 |
| CR | 998.13 (477.81) | 889.97 | 950.24 (314.16) | 872.57 | 253 | -.18 | .87 |
| | | | L2 Scores | | | | |
| Raw Lemmas-T2[a] | 114.55 (34.22) | 110 | 107.50 (30.34) | 114.50 | 235 | -.57 | .58 |
| Lex30-T2 | 50.62 (21.09) | 45 | 45.50 (18.32) | 48.50 | 233.50 | -.60 | .55 |
| Repeats [b] | 18.03 (8.19) | 16 | 18.17 (6.22) | 18 | 247 | -.31 | .77 |
| CR | 756.81 (430.61) | 647.78 | 631.18 (345.51) | 518.47 | 203 | -1.27 | .21 |
| | | | Residualized L2 Scores | | | | |
| Lex30-T2 | 0.07 (.91) | .10 | -.11 (1.12) | .18 | 246 | -.31 | .77 |
| CR | 0.11 (1.07) | -.09 | -.18 (.84) | -.37 | 204 | -1.25 | .22 |

*Note*. Twenty-nine (29) participants completed the two testing sessions between 1 and 5 days apart (Group 1); 18 participants completed the two testing sessions between 6 and 11 days apart (Group 2).

[a] Raw number of lemmas generated at Time 2. [b] Number of lemmas that occurred in both captures.

**Versions of the word association task.** Recall that four versions of the word association task were created in both English and French, each with a different set of 30 high frequency stimulus words. One group of participants (Group AB; *n* = 25) completed Versions A and B at Time 1 and Time 2, respectively, while another group (Group CD; *n* = 22) completed Versions C and D at Time 1 and 2, respectively. The average frequency rank of the stimulus items presented in each version is displayed in Table 9. Mann-Whitney *U* tests revealed no significant differences in the frequency ranks of the stimulus items encountered by Group AB and Group CD in English, $U = 1607$, $Z = -1.01$, $p =.31$, or in French, $U = 1703$, $Z = -.51$, $p =.61$. However, as suggested in Table 9, the overall frequency ranks of the stimulus words used in the French word association tasks were significantly higher than those used in English, $T = 3000$, $Z = -7.81$, $p < .001$, even though all words were drawn randomly from the first 2000 most frequent words in English and French.

We also examined whether participants performed differently on the vocabulary-related measures as a function of the versions of the word association task completed. As seen in Table 10, the non-parametric Mann-Whitney *U* test confirmed that there were no statistically significant differences in performance on the variables of interest between participants who completed Versions A and B, and those who completed Versions C and D of the word association task. Participants in Group AB and Group CD generated equivalent number of lemmas, and had similar Lex30 and CR scores in both English and French. As such, the version of the word association task that participants completed was not considered in future analysis.

Table 9: *Means and Standard Deviations of the Word Frequency Ranks in each Version of the Word Association Task*

| Versions | *M* | *SD* |
|---|---|---|
| | English | |
| A | 514.37 | 558.57 |
| B | 357.47 | 497.18 |
| C | 363.60 | 396.94 |
| D | 267.30 | 357.37 |
| | French | |
| A | 1068.70 | 620.92 |
| B | 728.33 | 503.94 |
| C | 786.27 | 602.39 |
| D | 889.43 | 530.63 |

*Note*. A different set of 30 high frequency stimulus words were used in each version of the word association task in English and French.

Table 10: *Results of the Mann-Whitney U Test comparing Participants' Performance on Versions A and B vs. Versions C and D of the Word Association Task*

| | Versions A and B | | Versions C and D | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *M* (*SD*) | *Mdn* | *M* (*SD*) | *Mdn* | *U* | *Z* | *p* |
| | | | L1 Scores | | | | |
| Raw Lemmas-T1[a] | 135.12 (23.38) | 135 | 149.23 (27.89) | 147.50 | 205.5 | -1.48 | .14 |
| Raw Lemmas-T2[b] | 153.20 (29.49) | 148 | 163.82 (34.45) | 163 | 226 | -1.05 | .30 |
| Lex30-T1 | 42.68 (15.36) | 39 | 50.50 (16.79) | 47 | 207.50 | -1.44 | .15 |
| Lex30-T2 | 48.12 (19.30) | 43 | 56.54 (21.60) | 49 | 213.50 | -1.31 | .19 |
| CR | 880.26 (307.41) | 760 | 1092.90 (502.31) | 930.14 | 196 | -1.68 | .09 |
| | | | L2 Scores | | | | |
| Raw Lemmas-T1[a] | 100.96 (31.21) | 100 | 102 (28.74) | 101 | 269 | -.13 | .90 |
| Raw Lemmas-T2[b] | 107.56 (33.49) | 105 | 116.73 (31.69) | 117 | 238 | -.79 | .44 |
| Lex30-T1 | 43.44 (16.76) | 44 | 43.86 (17.32) | 41 | 273.50 | -.03 | .98 |
| Lex30-T2 | 46.08 (18.91) | 42 | 51.59 (21.28) | 48.50 | 224 | -1.09 | .28 |
| CR | 715.70 (433.92) | 629.42 | 700.73 (370.01) | 564.87 | 264 | -.24 | .83 |
| | | | Residualized L2 Scores | | | | |
| Lex30-T1 | 0.10 (.97) | .17 | -.11 (1.02) | -.48 | 234.50 | -.86 | .39 |
| Lex30-T2 | -0.03 (.94) | .10 | .03 (1.06) | .15 | 267 | -.17 | .87 |
| CR | 0.06 (1.07) | -.21 | -.07 (.90) | -.29 | 248 | -.58 | .58 |

*Note*. 25 participants completed versions A and B at Time 1 and Time 2, respectively; 22 participants completed versions C and D at Time 1 and Time 2, respectively.

[a] Raw number of lemmas generated at Time 1. [b] Raw number of lemmas generated at Time 2.

**Practice effect.** Means and standard deviations of the number of lemmas generated at Time 1 and Time 2 and Lex30-T1 and -T2 scores in English and French are included in Table 11. Evidence of a possible practice effect was observed in these variables in both languages.

Table 11: *Means and Standard Deviations of the Raw Lemma Count and Lex30 scores at Time 1 and Time 2 in English and French*

| | English | | French | |
|---|---|---|---|---|
| Variables | *M* | *SD* | *M* | *SD* |
| Raw Lemmas-T1 | 141.72 | 26.28 | 101.45 | 29.76 |
| Raw Lemmas-T2 | 158.17 | 32.00 | 111.85 | 32.63 |
| Lex30-T1 | 46.34 | 16.35 | 43.64 | 16.84 |
| Lex30-T2 | 52.06 | 20.63 | 48.66 | 20.03 |
| Lex60[a] | 95.53 | 35.08 | 87.17 | 33.45 |

[a] Pooled Time 1 and Time 2 Lex30 scores.

***Raw number of lemmas generated.*** As seen in Table 11, more lemmas were supplied at Time 2 than at Time 1 in English. Since the raw number of English lemmas generated at Time 1 ($W$ = .98, $df$ = 47, $p$ = .68) was normally distributed, while the number of lemmas generated at Time 2 ($W$ = .94, $df$ = 47, $p$ = .03) was non-normally distributed, the non-parametric Wilcoxon signed-rank test was used to test the significance of the observed mean differences. Results indicated that, in English, participants generated significantly more associates at Time 2 (*Mdn* = 150) than at Time 1 (*Mdn* = 138), $T$ = 9, $Z$ = -4.58, $p$ < .001, $r$ = -.47.

Table 11 also shows that, in French, participants generated more lemmas when they did the word association task at Time 2. Normality was observed in the raw number of French lemmas generated at Time 1 (*W* = .99, *df* = 47, *p* = .97) and Time 2 (*W* = .99, *df* = 47, *p* = .97), as well as in the difference scores between these two variables (*W* = .98, *df* = 47, *p* = .41). As such, a paired-samples *t*-test was used to compare mean differences and indicated that participants also generated significantly more lemmas in French at Time 2 (*M* = 111.85, *SD* = 32.63), than they did at Time 1 (*M* = 101.45, *SD* = 29.76) , *t*(46) = -4.31, *p* < .001, *r* = .53.

***Lex30 scores at Time 1 and Time 2.*** In both English and French, a Lex30 score at Time 1 and Time 2 was calculated for each participant and analyses were conducted to determine whether these scores differed across time. Results indicated that, in both English and French, Lex30 scores were significantly higher at Time 2, which may also be indicative of a practice effect in our data. As seen in Table 11, English Lex30 scores at Time 2 indicate that participants supplied roughly 6 more infrequent words than they did at Time 1, a difference which was found to be statistically significant, *T* = 14, *Z* = -3.08, *p* < .002, *r* = -.32. A similar pattern was observed in French, where participants supplied roughly 5 more infrequent words at Time 2 than they did at Time 1, a difference that also reached significance, *t*(46) = -2.83, *p* = .007, *r* = .38.

**Speed and efficiency of lexical access.** Reaction times in milliseconds on the LNL task and the CV scores in English were compared to their French equivalents to determine whether the expected pattern of results (lower RTs and CV scores in the L1) would be found. These lower RTs (meaning faster responses) and CV scores indicate more efficient cognitive processing, which would be expected in the more fluent L1. The Wilcoxon signed-rank test revealed that RTs were indeed lower in the L1 (*Mdn* = 646), relative to the L2 (*Mdn* = 701), *T* = 7, *Z* = -4.83, *p* < .001, *r* = -.50, indicating faster L1 performance. On the other hand, a comparison of the English CV scores and the unresidualized French CV scores revealed no difference in efficiency of cognitive processing in the L1 (*Mdn* = .19) and L2 (*Mdn* = .19), $T = 23$, $Z = -.39$, $p = .70$, $r = -.04$.

**Testing Hypotheses**

In the manuscript, wherever possible, only residualized CR and Lex30-T2 scores, residualized RTs and CV scores, and English Lex30-T2 scores were reported in tests of our hypotheses. However, we also conducted tests of our hypotheses using unresidualized CR, RT and CV scores, unresidualized Lex30-T1 and Lex30-T2 scores, residualized Lex30-T1 scores, English Lex30-T1 scores, as well as a pooled Lex60 score based on the Time 1 and Time 2 scores. These additional analyses are reported here.

**Hypothesis 1.** It was hypothesized that the CR scores would be positively correlated with Lex30 scores, as an indication of the CR's convergent validity. In support of Hypothesis 1, Spearman correlations revealed that, in English, the CR estimate of vocabulary size was significantly positively correlated with Lex30 scores at Time 1 ($r_s$ = .65, $p < .001$). Unresidualized French scores show the same pattern of results, i.e., a significant positive correlation between CR and Lex30 scores at Time 1 ($r_s$ = .72, $p <$ .001) and Time 2 ($r = .74$, $p < .001$) scores. In the residualized French data, the same pattern was observed, i.e., a significant positive correlation between residualized CR scores and Lex30 scores at Time 1 ($r = .71$, $p < .001$).

Additionally, in both French and English, we combined the two separate Lex30 scores to come up with a 'Lex60' score (see Table 11), as if participants had associated to 60 high frequency stimulus words. This was done to create a 'Lex' score that takes into account all of the data provided by participants, as the CR does. If an infrequent word was repeated at Time 1 and Time 2, a point was assigned to only one of those words. Spearman correlations indicated that the CR scores were also significantly positively correlated with Lex60 scores in English ($r_s$ = .71, $p$ < .001) and in French ($r_s$ = .79, $p$ < .001), and residualized CR and Lex60 scores were also significantly positively correlated ($r$ = .76, $p$ < .001). Overall, these analyses also confirmed the CR's convergent validity.

**Hypothesis 2.** This hypothesis was primarily concerned with establishing whether the CR technique can distinguish between the L1 and L2, and these analyses are reported in the manuscript. However, we were also interested in whether Lex30-T1 and Lex60 scores would be able to distinguish first and second languages. Analyses indicated that at Time 1, Lex30 scores were unable to distinguish between participants' L1 and L2. The difference between the English Lex30 score at Time 1 (*M* = 46.34, *SD* = 16.35) and the French Lex30 score at Time 1 (*M* = 43.64, *SD* = 16.84) was not significant, $t$(46) = 1.05, $p$ = .30, $r$ = .15. Lastly, a Wilcoxon signed-rank test indicated that Lex60 scores were also unable to distinguish between participants' L1 (*Mdn* = 89) and L2 (*Mdn* = 81), $T$ = 21, $Z$ = -1.46, $p$ = .15, $r$ = -.15. Thus, while CR scores were able to distinguish between participants' L1 and L2, none of the 'Lex' measures were able to do so,

**Hypothesis 3.** As an additional test of the CR's construct validity, it was hypothesized that CR scores would be significantly negatively correlated with the speed and efficiency of lexical access. Analysis of the unresidualized data reveals that this hypothesis was partially supported. Spearman correlations show that, in French, unresidualized CR scores were found to correlate significantly, in the expected negative direction, with speed of lexical access ($r_s$ = -.44, $p$ = .002), but not at all with efficiency of lexical access ($r_s$ = -.20, $p$ = .18).

We also examined whether the ‘Lex’ scores would correlate positively with the speed and efficiency of lexical access. Spearman correlations indicated that Time 1 Lex30 scores in English were not significantly correlated with speed ($r_s$ = -.10, $p$ = .51) or efficiency ($r_s$ = .07, $p$ = .64) of lexical access. English Lex60 scores showed the same pattern of results, i.e., no relation to speed ($r_s$ = -.18, $p$ = .22) or cognitive efficiency ($r_s$ = .02, $p$ = .90) of lexical access. Spearman correlations of the residualized French data, however, revealed a significant negative correlation between the Lex30 scores at Time 1 and speed of lexical access ($r_s$ = -.32, $p$ = .03), but no significant relation between the Lex30-T1 scores and the efficiency of lexical access ($r_s$ = .05, $p$ = .72). Similarly, in the unresidualized French data, Lex30-T1 scores were found to correlate negatively with RTs on the LNL ($r_s$ = -.35, $p$ = .02), but not with CV scores ($r_s$ = -.02, $p$ = .90). The same pattern emerged from the Time 2 data, i.e., a significant negative correlation between unresidualized French Lex30-T2 scores and RTs on the LNL($r_s$ = -.48, $p$ = .001), but no relation to CV scores ($r_s$ = -.10, $p$ = .52). The residualized and unresidualized Lex60 scores in French showed the same pattern of results, i.e., a significant negative correlation with speed ($r_s$ = -.39, $p$ = .007) and  ($r_s$ = -.44, $p$ = .002), respectively, but no relation to CV scores ($r_s$ = .03, $p$ = .85) and ($r_s$ = -.05, $p$ = .73), respectively. A summary of these results, and those presented in the manuscript, are provided in Table 12.

Table 12: *Summary of the Results of the Test of Hypothesis 3*

| Variables | Is the vocabulary measure significantly correlated with RTs? | Is the vocabulary measure significantly correlated with CV scores? |
|---|---|---|
| | **L1 Scores** | |
| CR | No | No |
| Lex30-T1 | No | No |
| Lex30-T2 | No | No |
| Lex60 | No | No |
| | **Unresidualized L2 Scores** | |
| CR | Yes, negatively. | No |
| Lex30-T1 | Yes, negatively | No |
| Lex30-T2 | Yes, negatively. | No |
| Lex60 | Yes, negatively. | No |
| | **Residualized L2 Scores** | |
| CR | Yes, negatively. | No |
| Lex30-T1 | Yes, negatively. | No |
| Lex30-T2 | Yes, negatively. | No |
| Lex60 | Yes, negatively. | No |

# CHAPTER 4

## General Discussion

In the current work, we examined the validity of the Capture-Recapture (CR) methodology as a measure of L2 productive vocabulary size. We found that this technique, which is traditionally used to reliably estimate the size of animal populations, holds some promise as a measure of L2 productive vocabulary size. The word association format was used to elicit vocabulary from participants. As such, the test characteristics of the CR, as outlined by Read (2000), are virtually identical to those of the Lex30. For instance, the CR, like the Lex30, can be described as measure of productive vocabulary size that is[††]

- *discrete*, in that it assesses productive vocabulary knowledge "as a distinct construct, separated from other components of language competence" (Read, 2000, p. 8);
- *comprehensive*, in that it "takes account of all the vocabulary content of a …written text" (Read, 2000, p. 11) to generate vocabulary size estimates, perhaps even more than the Lex30 does; and
- *context-independent*, in that the word association format "neither presents prompt items in context nor requires responses to be contextualised" (Fitzpatrick, 2003, p. 225).

However, despite these similarities in the design of the CR methodology and the Lex30,

[††] The three dimensions proposed by Read's (2000) are discrete--embedded, selective--comprehensive and context-independent--context-dependent. Embedded tests are just part of a wider assessment of a larger construct; selective tests target pre-selected items; and, context-dependent tests require participants to consider the context in production of a correct response.

these tests did not behave identically in the current work. Unlike the Lex30, the CR was shown to be sensitive enough to distinguish participants' L1 and L2, and was found to correlate with speed and efficiency of lexical access. Nonetheless, we believe that the CR and Lex30 are giving different, but complimentary, information about productive vocabulary knowledge, the former indicating perhaps the size of the current lexical pool from which an individual can draw, and the latter indicating the quality of that pool, specifically in terms of the infrequency of words.

While the validity evidence obtained in this study for the CR is promising, no review of its performance is complete without careful examination of whether the assumptions of the Petersen formula were actually met. Violations of the following assumptions place considerable limitations on the validity of, and conclusions based on, the CR estimates.

### Assumption 1

Since the CR technique and Petersen's formula are inferential procedures, in that they require a sample to make inferences about a population, an initial assumption that must be met if valid estimates of population size are to be made, is that the sample taken should be representative of the population as a whole. While it is unclear how exactly we should define representativeness as it relates to items in the lexicon, other researchers have reasoned that the word association format encourages production that is representative of the current state of an individual's lexicon, at least where lexical frequencies are concerned (Fitzpatrick & Clenton, 2010; Fitzpatrick & Meara, 2004). Fitzpatrick and colleagues came to this conclusion after observing that, even though their participants supplied mostly different words during two administrations of the Lex30 test, the proportion of infrequent words supplied remained the same. In other words, participants' frequency profile remained constant, even though they supplied different words across time. However, in our study, we found significant differences between Lex30 scores at Time 1 and Time 2 in both French and English, with significantly more infrequent words generated at Time 2. It would be reasonable to speculate that we were unable to find the stability in Lex30 scores observed by Fitzpatrick and colleagues because our participants associated to different words at Times 1 and 2. This result casts doubt on the ability of the word association task to sample representatively from the lexicon, even where word frequencies are concerned, since under different conditions, participants showed an ability to produce additional infrequent words. The issue of how to define representativeness in terms of vocabulary items is not easily resolved. Lexical frequencies aside, representativeness of vocabulary items can be conceptualized in terms of even the parts of speech, registers and genres that an individual can handle. It is not yet obvious how one would go about constructing an elicitation method that can capture truly 'representative' samples of any, or all, of these characteristics of the lexicon.

Furthermore, the fact that lexical data needs to be lemmatized and processed once collected, may be undermining our attempts at representativeness. In the current work, we excluded function words, including prepositions, pronouns and conjunctions, proper nouns, and number names. If any sample is to come close to being truly representative of an individual's lexicon, it is possible that we would have to include these valid lexical items in future calculations of CR estimates of productive vocabulary size.

**Assumption 2**

A second assumption of the CR methodology is that all items in the population can be trapped and have the same probability of being trapped; if not, “population size will be seriously underestimated” (Sutherland, 2006, p. 99). This assumption was clearly violated in Meara and Olmos Alcoy’s (2010) study, where the items necessary to describe the picture story had a much greater probability of being captured and recaptured than other items in participants’ lexicon, while other vocabulary items were simply not trappable with their writing task. Although we used a trapping procedure in the current work that places far fewer restrictions on participants’ production, it is still possible that some priming of certain vocabulary items occurred since we used words to stimulate the production of other words. Indeed, a common finding in the literature is that the words supplied as associates are influenced by the features of the stimulus word, with nouns and verbs, encouraging other nouns and verbs, respectively, as associates, and adjectives often encouraging nouns (Fitzpatrick, 2006; Sökmen, 1993). The risk we face with the word association task “is that each response in the list acts as a stimulus for the next response, and so on, resulting in an association chain rather than a collection of associations” (Fitzpatrick, 2006, p. 3). Furthermore, unlike fish in a river, the same words can be trapped multiple times *within* a testing session, such as when a participant supplies the word *dream* four times in response to four different stimulus words at Time 1. This phenomenon is likely indicative of the fact that these within-session repeats are, for whatever reason, more trappable than other lexical items.

It appears that, in language, it may not be possible to construct an elicitation method that will not influence the probabilities of certain words being captured over others, because of the nature of the links between items in the mental lexicon. If an ecological estimation technique is to be applied to the estimation of productive vocabulary size, then it is necessary to find an ecological model that allows for unequal probabilities of capture and recapture in the population, since this appears to be the case in language.

**Assumption 3**

There is also the related issue of whether the mere fact of having given a word as an associate at Time 1, increases (even slightly) its probability of being generated at Time 2. This would be a violation of an additional assumption of the CR methodology, which specifies that the process of marking the captured items should not affect their behaviour or fate (Lindberg & Rexstad, 2002; Sutherland, 2006). Within-session repeats were present in the current work and we can argue that the likelihood of participants supplying these words as associates increased their capturability, either in the same session, or across time. Again, this is a difficult challenge to overcome in language since words are necessarily linked to each other in the lexicon.

**Assumption 4**

The fourth, and likely most important, assumption of the CR is that "[r]esampling is instantaneous; that is, birth, death, immigration and emigration do not occur during the resampling process" (Lindberg & Rexstad, 2002). In other words, the CR paradigm assumes that the population to be measured is 'closed', i.e., "that there are no gains (births or immigration) or losses (deaths or emigration) during the course of the study" (Sutherland, 2006, p. 98). Presumably, this assumption is violated when the CR is used to assess vocabulary size, especially in the second language, since the emerging lexicon is not stable enough to be considered a closed system (Bell, 2009; Racine, 2011). New vocabulary items are being added to the lexicon fairly regularly, some of which may be forgotten shortly after they're learned. Other lexical items that are not used often may even lose their productive status all together (Racine, 2011). Thus, it is likely that the mental lexicon as a whole is more of an open system, subject to regular fluctuations in the total population size because of losses and gains to the population. Sutherland (2006) points out that the two-sample Capture-Recapture technique explored in the current work is the most basic of the estimation techniques and is appropriate *exclusively* for closed populations. "If there are losses from the population, the estimate obtained is for the size of the population at the time of the first catch; if there are gains, the estimate corresponds to the population size during the second catch; if there is turnover (gains and losses) [arguably the case for language] the estimate is biased" (Sutherland, 2006, p. 100).

If an ecological sampling procedure is to be applied to estimating productive vocabulary size, then an open population sampling technique should be used, such as the Jolly-Seber models outlined by Sutherland (2006). Unfortunately, this is easier said than done! Indeed, the simplicity of the CR technique and the Petersen's estimate is what makes it an attractive option. Open population models, on the other hand, are much more complex, both in their procedures and in the mathematics required to calculate estimates of population size. In open population models at least three capture occasions seem to be necessary, and both the rate of loss from and gains to the population have to be estimated (Sutherland, 2006); these calculations may be especially difficult to perform with vocabulary. Additionally, open population models have their own set of assumptions, in addition to those of closed models, and seem to be calculated using specialized statistical programs (Sutherland, 2006). Nevertheless, if ecological estimation techniques are to be used to make valid estimates of productive vocabulary size as a whole, then we must focus our efforts on exploring the feasibility of applying open population models to language.

**An Alternative Interpretation of the CR Estimate**

In the manuscript, we proposed the idea of limiting the scope of the generalizations we make based on the CR estimates. Specifically, we speculated that instead of interpreting the CR estimate as a direct indication of the absolute size of an individual's productive vocabulary, perhaps it should be interpreted as an indication of the number of words an individual has available to them to complete a certain task under the specific task requirements encountered. Under this interpretation, the number of words available for task completion would be the indication of an overall larger productive vocabulary size, since the actual values obtained seem more realistic as estimates of size on a much smaller scale. Indeed, limiting the scope of generalizations of the CR might be the way to proceed since there is the possibility that the number of words an individual has available to complete a given task can be considered a closed system. To the extent that this alternative interpretation is valid, perhaps then the CR estimate is actually suitable in this context. Presumably, no new words are being added to or lost from the lexical pool during the 15 minutes it takes participants to complete a word association task. If there are two separate trapping sessions, however, gains to and loss from the population become an issue.

It may be possible to design a study in which both captures are done during one testing session, such as if participants complete two word association tasks after taking a 5 or 10 minute break. This may limit the likelihood of fluctuations to the population between captures. Of course, there are other assumptions that would need to be considered, such as the representativeness of sampling and the capturability assumptions. It appears that the formula for calculating the unbiased estimate of population size, $N$ = [(n1 +1)(n2 +1)/(m2 +1)] – 1, may account for violations of the representativeness, since this formula eliminates bias arising from statistical issues (Sutherland, 2006). As far as the assumption related to equal probabilities of capture and recapture is concerned, however, Lindberg and Rexstad (2002) state that in closed models there is actually "a relaxation of the general assumption of equal probabilities of capture of all individuals in our population at every sampling occasion" (p. 4). Additionally, there are models for estimating the size of closed populations that allow for heterogeneity in capture probabilities, due to a wide range of factors, that can be applied under these conditions. Thus, if it is reasonable to interpret the CR as being an indication of how many words an individual has available for task completion at a given time, and if that pool represents a closed system, there is potential for the CR to be used appropriately to estimate population size within these parameters.

There are a number of techniques used to estimate the size of open and closed populations in ecology, some of which are far more complex and refined than the CR methodology (Sutherland, 2006). Applying one of the many estimation techniques involves carefully deciphering the ecological literature to find the formula or methodology that best applies to the mental lexicon, as well as the correct psycholinguistic analogues for the corresponding ecological variables. Fish and words may be considered capturable to some extent, but the factors that influence their capturability may be different and operate in distinctive ways. For instance, a set of environmental and behavioural factors influence the capturability of fish, but the capturability of a word, in the sense of accessibility or retrievability from memory, may be influenced by completely different factors, which may interact in ways that ecological factors do not (e.g., a retrieved word may prime other words affecting their retrievabiity). It is absolutely necessary, therefore, to consider these issues closely if we hope to come to valid conclusions about productive vocabulary size using a technique derived from population ecology.

## Final Conclusions

The goal of the current work was to provide validity evidence for the Capture-Recapture technique as a measure of second language productive vocabulary size. The convergent and construct validity of the CR technique was confirmed since it showed significant positive correlations with the validated Lex30 measure of productive vocabulary size, distinguished between the L1 and L2, and was related to the speed of L2 lexical access. At first glance, these results suggest that the CR is indeed a valid measure of productive vocabulary size. Although it is tempting to interpret the CR as an estimate of absolute population size, as ecologists do, we propose interpreting these estimates as an indication of how many vocabulary items and individual has available to him for completing a specific task at a given time and under the specific conditions set up by the task. Perhaps this can be taken as the indication of an overall larger vocabulary size. Further, if this interpretation of the CR is valid, perhaps we can consider that the number of lexical items available for task completion is a closed system, like those the CR is intended for.

Additionally, we used a continuous word association task to elicit vocabulary from participants. This allowed us to circumvent a number of problems commonly associated with measures of productive vocabulary size that would have artificially lowered the CR estimates. More specifically, the word association task places few, if any, restrictions on participants' production, so they are not required to supply the same words or to consider a context in order to complete the task. Furthermore, the word association format allowed participants to generate a fairly large number of content words from which to estimate productive vocabulary size, in a relatively short amount of time. These features made the word association format an ideal trapping method for use within the CR paradigm.

However, before any ecological estimation techniques can be implemented as measures of productive vocabulary size, it is critical that we borrow the method that best suits the conditions of vocabulary and the mental lexicon. It can be argued that the CR method and the Petersen formula don't apply to estimating productive vocabulary size as a whole, since it is difficult to meet a number of their assumptions when dealing with vocabulary. Further, and more importantly, the overall mental lexicon may be an open system that requires estimation techniques suited for estimating the population size in systems that are subject to regular gains and loses between captures. The CR is not appropriate for estimating population size under these conditions. Conversely, if it is reasonable to interpret the CR as being an indication of how many words an individual has available for task completion at a given time, and if that pool represents a closed system, there is potential for the CR to be used appropriately to estimate population size within these parameters, since an unbiased estimator can be used to deal with lack of representativeness in sampling and the assumption of equal probabilities of capture "need not be strictly adhered to in closed populations" (Lindberg & Rexstad, 2002).

## Future Directions

The application of ecological sampling techniques has potential to become a fruitful area of research. An appropriate starting point would be to outline the scope of the generalizations we can make from these estimates. If we want to estimate to the productive vocabulary as a whole, then it may be worthwhile to survey the open population estimation models and investigate whether they can be applied to language. Some open population models appear to require estimation of a number of elements, such as the rate of loss and gains from the population (Sutherland, 2006). It would be interesting to see whether the rate at which vocabulary items in the L1 and L2 are lost from and added to the mental lexicon is something that can be estimated. Additionally, further study is needed to explore the validity of the interpretation of the CR as an indication of the vocabulary available for task completion at a given time. This would allow us to use the much simpler, and more statistically accessible, closed population models to estimate productive vocabulary size. Of course, if we can find an appropriate way of applying an ecological technique to the estimation of productive vocabulary size, further validity and reliability studies would have to be conducted to determine whether the test would be able to distinguish separate groups, such as language learners and native speakers, or learners with varying degrees of language proficiency. Would such a test be able to show improvement as learners progress in their language learning? Would the scores show stability over time especially among native speakers or advanced language learners whose vocabulary size may already be relatively stable? These are all interesting questions that can be pursued in future research.

Another interesting line of research could tackle the issue of sampling representatively from an individual's lexicon. We need to be able to define what representativeness means in language and develop tasks that can truly generate a representative sample from which to estimate vocabulary size. In the current work, we used a word association task featuring stimulus words drawn from the first 2000 most frequent words in the lexicon. However, if we define representativeness in terms of the frequencies contained in the lexicon, then perhaps it is worthwhile to investigate the frequencies of the associates themselves. Since we have information from the Lex30, we know the proportion of infrequent words supplied, but we do not know the range of infrequent words supplied, since this test rewards a point for words in any frequency band above that of the stimulus words. Future research can explore whether selecting stimulus words from a range of frequency bands would stimulate participants to produce associates that are more representative of their lexicon, at least where the frequencies of items are concerned.

# REFERENCES

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253-279.

Beglar, D., & A. Hunt. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing, 16*(2), 131-162.

Bell, H. (2009). The messy little details: a longitudinal case study of the emerging lexicon. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners* (pp. 111–127). Bristol, UK: Multilingual Matters.

Clenton, J. (2008). Investigating the construct of productive vocabulary: Comparing different measures. *Proceedings of the BAAL Annual Conference*, *27*.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.

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

Edwards, R., & Collins, L. (2011). Lexical frequency profiles and Zipf's law. *Language Learning*, *61*(1), 1-30.

Field, A. (2009). *Discovering statistics using SPSS (3rd ed.)*. Thousand Oaks, CA: Sage Publications, Inc.

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

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook*, *6*, 121-145.

Fitzpatrick, T. (2007). Productive vocabulary tests and the search for concurrent validity. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 116-132). Cambridge, UK: Cambridge University Press.

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

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241-265.

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

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

Kruse, H., Pankhurst, J., & Sharwood Smith, M. (1987) A multiple word association probe in second language acquisition research *Studies in Second Language Acquisition, 9*(2), 141-154.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In Arnaud, P.J.L., & Béjoint, H. (eds), *Vocabulary and Applied Linguistics.* London: Macmillan. pp. 126–132.

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

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307-322.

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

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

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

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

MathWorks. (2007). MATLAB, version 2007b. Natick, Massachusetts: The MathWorks Inc.

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

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

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., & Jones, G. (1988). Vocabulary size as a placement indicator. In: Grunwell, P. (Ed.), *Applied Linguistics in society.* CILT, London. pp. 80-87.

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

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

Meara, P. M., & Miralpeix, I. (2007). *Vocabulary size estimator.* Swansea: Lognostics.

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 Schmitt, N., & M. McCarthy (Eds), *Vocabulary: Description, acquisition and pedagogy.* Cambridge: Cambridge University Press.

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

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

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

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

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 116-132). Cambridge, UK: Cambridge University Press.

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.

Politzer, R. L. (1978). Paradigmatic and syntagmatic associations of first year French students. In Honsa, V and J. Hardman-de-Bautista (Eds), *Papers in linguistics and* *child language.* The Hague: Mouton.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In Kunnan, A. (Ed), *Validation in language assessment.* Mahwah, NJ: Lawrence Erlbaum.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Riegel, K., & I. Zivian. (1972). A study of inter- and intralingual associations in English and German. *Language Learning, 22*(1), 51-63.

Schmitt, N. (1997). Vocabulary learning strategies. In Schmitt, N. and McCarthy, M. (eds), *Vocabulary: Description, acquisition, and pedagogy.* Cambridge: Cambridge University Press.

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

Seber, G. A. F., Huakau, J. T., & Simmons, D. (2000). Capture-Recapture, epidemiology, and list mismatches: Two lists, *Biometrics, 56*, 1227–1232. doi: 10.1111/j.0006-341X.2000.01227.x

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

Söderman, T. (1993) Word associations of foreign language learners and native speakers – different response types and their relevance to lexical development. In Hammarberg, B. (Ed.) *Problems, process and product in language learning.* Abo: AfinLA.

Sökmen, A. (1993) Word association results: A window to the lexicons of ESL students. *JALT Journal,* 1I5(2), 135-150.

Sutherland, W. J. (2006). *Ecological Census Techniques (2nd ed.): A handbook.* Cambridge: Cambridge University Press.

Thorndike, R. M., & Thorndike-Christ, T. (2010). Qualities desired in any measurement procedure: Validity. *Measurement and evaluation in Psychology and Education* (8th ed.). Boston, MA: Pearson

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch LI and L2 children. In Bogaards, P., & Laufer, B. (eds.) *Vocabulary in a second language.* Amsterdam and Philadelphia, PA: John Benjamins, 173-89.

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

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

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

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1,* 80–83.

Wilkins, D. (1972). *Linguistics and Language Teaching.* London: Edward Arnold.

Wolter, B. (2002). Assessing proficiency through word associations: Is there still hope? *System*, *30*, 315-329.

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

# APPENDIX A

## CONSENT FORM TO PARTICIPATE IN RESEARCH

This is to state that I agree to participate in a program of research being conducted by Joy Williams (joyawilliams@gmail.com) of the Department of Psychology at Concordia University as a requirement for completion of the Master's Thesis, under the supervision of Dr. Norman Segalowitz.

**A. PURPOSE** I understand that the purpose of this research is to study processes underlying second language development.

**B. PROCEDURES** I understand that this study will take place at Concordia University, in the laboratory of Dr. Segalowitz. I understand that I will be asked to fill out a word association test, in which I will write a minimum of four word associates to a total of 60 stimulus words. I will also be asked to identify word stimuli that will appear on a computer screen by responding on a keypad. I am aware that my responses for these two tasks will be timed. I am also aware that I will have to answer a questionnaire concerning my use of my first and second languages. I understand that I will take part in two sessions, with a total testing time of approximately 1 hr per session.

**C. CONDITIONS OF PARTICIPATION**

- I agree to participate in this study, which is expected to last about 2 hours in total.
- I understand that I am free to withdraw my consent and discontinue my participation at any time without negative consequences.
- I understand that my participation in this study is confidential (i.e., the researcher will know but will not disclose my identity).
- I understand that the data from this study may be published. In this case, my identity and my personal data will not be revealed in a way that can be associated with me.
- I will be paid $10 per hour or participation credits upon completion of my participation.

I HAVE CAREFULLY STUDIED THE ABOVE AND UNDERSTAND THIS AGREEMENT.
I FREELY CONSENT AND AGREE TO PARTICIPATE IN THIS STUDY.

NAME (please print): ______________________________

SIGNATURE: ______________________________

RESEARCHER SIGNATURE: ______________________________

DATE: ______________________________

Please indicate if you are willing to participate in other studies conducted by our research group:

YES ____ e-mail: ______________________________ NO _______

**For further information about this study, either before or after it is completed, please contact: Dr. Norman Segalowitz by telephone at (514) 848.2424 x2239 or by e-mail at <norman.segalowitz@concordia.ca>. If you have questions about your rights as a research participant, please contact Adela Reid, Research Ethics and Compliance Officer, Concordia University at (514) 848.2424 x7481 or by e-mail <Adela.Reid@concordia.ca>.**

# APPENDIX B[‡‡]

## LANGUAGE BACKGROUND QUESTIONNAIRE

Name : ________________________________ Date ________________________

Age : __________________ Sex: M ___ F___

1. If you are a student:

   What is your field of study? ____________________________________________

   What degree are you pursuing? College/Cégep ___ Bachelor ___ MA/PhD ___

2. Where were you born? City:___________________ Country: _______________________

3. What do you consider to be your **first learned language**?

   English ___ French ___ Other ________________

4. What do you consider to be your **second learned language**?

   English ___ French ___ Other ________________

5. At what age did you learn your **second language**? ______________________________

6. What language do you consider your dominant language?

   English ___ French ___ Other ________________

7. What language do you speak at home now? _______________________________________

8. What is the first language of your: Mother? ____________ Father? ________________

9. In what language did you attend school? (Please check the appropriate one):

   - Elementary school: English ___ French ___ French Immersion ___ Other ______
   - Middle/High school: English ___ French ___ French Immersion ___ Other ______
   - College/Cégep: English ___ French ___ Other ____________
   - University: English ___ French ___ Other ____________

10. If you are not currently a student, what is the highest level of education you have completed:

    High school ___ College ___ University (Bachelor) ___ University (MA/PhD) ___

[‡‡] Data from Items 16 to 31 could not be used because of interpretation difficulties that were discovered after the fact.

11. Have you received second language instruction in school at any of the levels listed below, and for how long? YES ___ NO ____
If YES, specify each language, starting with your main second language.

MAIN SECOND LANGUAGE: _______________

- Elementary School: less than 1 year ___ 1-2 years ___ more than 2 years ___
- Middle/High School: less than 1 year ___ 1-2 years ___ more than 2 years ___
- College/Cégep/University: less than 1 year ___ 1-2 years ___ more than 2 years ___
- Other: less than 1 year ___ 1-2 years ___ more than 2 years ___

Please specify: ___________________________________________________________

THIRD LANGUAGE (if any): _______________

- Elementary School: less than 1 year ___ 1-2 years ___ more than 2 years ___
- Middle/High School: less than 1 year ___ 1-2 years ___ more than 2 years ___
- College/Cégep/University: less than 1 year ___ 1-2 years ___ more than 2 years ___
- Other: less than 1 year ___ 1-2 years ___ more than 2 years ___

Please specify: ___________________________________________________________

Any other special school-related learning experiences (e.g., intensive French in Grade 6):

_________________________________________________________________

12. Do you have any visual impairment NOT corrected by wearing glasses or contact lenses? Yes ___ No ___

13. Do you have a known hearing impairment? Yes ___ No ___

14. Do you have a known reading or attention disability? Yes ___ No ___

15. Please rate your level of ability for each of the four skills listed below by using the following rating scheme and circling the appropriate number in the boxes below:

**1** = no ability at all **2** = very little **3** = moderate **4** = very good **5** = fluent ability

| **Language** | **Speaking** | **Reading** | **Writing** | **Listening** |
|---|---|---|---|---|
| English | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 |
| French | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 |
| Other<br>____________ | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 |

**In general, when I use French, my second language,**

| | TRUE | FALSE |
|---|---|---|
| 16. I can enter into and speak fairly well in full-length conversations about simple and familiar topics, *only if I am prepared in advance*. | ______ | ______ |
| 17. I cannot communicate easily *even when the topic is simple and familiar.* | ______ | ______ |
| 18. I have to *speak more slowly, repeat and correct my speech more often than in my first language,* even in ordinary conversations. | ______ | ______ |
| 19. I can easily and correctly use all or nearly all expressions that native speakers typically use. | ______ | ______ |
| 20. I can express myself easily on a wide variety of different topics, feelings and opinions, *pausing only occasionally to find the appropriate words and expressions*. | ______ | ______ |
| 21. I can easily and smoothly find other ways to say things when I don't know a particular word *without noticeably searching for words or avoiding saying certain things*. | ______ | ______ |
| 22. I can discuss work, family, travel, and personal interests in ordinary, full-length conversations, *as long as I use simple language*. | ______ | ______ |
| 23. I can handle short conversations *only if the topics are simple and familiar.* | ______ | ______ |

**In general, when I use French, my second language,**

| | TRUE | FALSE |
|---|---|---|
| 24. I can understand most native speakers *if they occasionally repeat individual words or short expressions when I need help*." | ______ | ______ |
| 25. I do not understand simple questions and instructions, even when people speak clearly and slowly to me in standard spoken language. | ______ | ______ |
| 26. I can usually manage normal conversations with native speakers *if they speak slowly and directly to me using simple language and frequently explain things to me*." | ______ | ______ |
| 27. I can easily understand most native speakers talking about unfamiliar topics, even when I am not used to their accent or way of speaking. | ______ | ______ |
| 28. I can easily understand most native speakers, talking about unfamiliar topics, *but I may have to ask for explanations and repetition more often than in my first language.* | ______ | ______ |
| 29. I can understand most native speakers *if they use standard spoken language and talk about familiar topics*. | ______ | ______ |
| 30. I can usually manage normal conversations with native speakers *if they repeat the message in full or in part when I need help.* | ______ | ______ |
| 31. I can understand simple questions and instructions from most native speakers*, if they speak clearly, slowly and directly to me, and they repeat when necessary.* | ______ | ______ |

a) In what situations do you tend to speak in English with other people? (check all that apply)

___ When one on one ___ At home ___ With friends ___ With family

___ When out (shopping, etc.) ___ Other (please specify) ____________________________

b) In what situations do you tend to speak in French with other people? (check all that apply)

___ When one on one ___ At home ___ With friends ___ With family

___ When out (shopping, etc.) ___ Other (please specify) ____________________________

c) What percentage of your interactions with other people are in: English __ %? French __%?

*Please answer the following questions, considering how you speak when interacting with other people. Please circle a number to indicate how much you agree with each statement.*

d) I often start a sentence in English and then switch to speaking French

1 2 3 4 5 6 7
Very true Somewhat true Not at all true

e) I often start a sentence in French and then switch to speaking English

1 2 3 4 5 6 7
Very true Somewhat true Not at all true

f) I often use a French word when speaking English

1 2 3 4 5 6 7
Very true Somewhat true Not at all true

I do this in situations when (check all that apply):
___ I'm not sure of the English word
___ No translation or only a poor translation exists for the word
___ The English word is hard to pronounce
___ None of the above / not sure

g) I often use an English word when speaking French

1 2 3 4 5 6 7
Very true Somewhat true Not at all true

I do this in situations when (check all that apply):
___ I'm not sure of the French word
___ No translation or only a poor translation exists for the word
___ The French word is hard to pronounce
___ None of the above / not sure

h) In general, I often mix English and French with the people I speak to

1 2 3 4 5 6 7
Very true Somewhat true Not at all true

# APPENDIX C

## STIMULI USED IN THE ENGLISH WORD ASSOCIATION TASKS

Table 13: *Stimuli used in the Four Versions of the Word Association Task in English*

| Version A | Version B | Version C | Version D |
|---|---|---|---|
| 1. see | water | buy | make |
| 2. help | father | program | country |
| 3. news | can | finger | room |
| 4. family | Talk | call | little |
| 5. live | believe | food | ground |
| 6. home | eye | sit | friend |
| 7. become | leave | go | mother |
| 8. provide | new | try | speak |
| 9. student | tip | turn | hit |
| 10. story | let | show | rush |
| 11. hour | woman | fact | company |
| 12. write | head | break | time |
| 13. state | right | begin | like |
| 14. work | day | happen | have |
| 15. soldier | good | include | mean |
| 16. child | stand | life | feel |
| 17. game | take | keep | hold |
| 18. grow | school | nurse | man |
| 19. name | bad | book | old |
| 20. run | kind | look | people |
| 21. lose | send | health | find |
| 22. skin | give | government | wall |
| 23. start | soil | house | ask |
| 24. aim | building | hand | think |
| 25. holiday | political | corner | offer |
| 26. tell | way | door | use |
| 27. side | win | say | number |
| 28. improve | high | night | want |
| 29. get | guest | study | part |
| 30. consider | know | need | play |

## APPENDIX D

## STIMULI USED IN THE FRENCH WORD ASSOCIATION TASKS

Table 14: *Stimuli used in the Four Versions of the Word Association Task in French*

| Version A | Version B | Version C | Version D |
|---|---|---|---|
| 1. autoriser | gestion | comprendre | prouver |
| 2. bataille | fil | sérieux | catégorie |
| 3. posséder | entendre | partager | voisin |
| 4. secrétaire | niveau | bouger | vif |
| 5. remarquer | couvrir | triste | contraire |
| 6. devoir | législatif | rencontrer | perte |
| 7. préparer | frère | ciel | expliquer |
| 8. réussite | secteur | cours | usine |
| 9. diriger | éviter | souligner | disponible |
| 10. monde | répondre | compter | choisir |
| 11. inquiétant | doute | allié | témoigner |
| 12. découverte | règle | suivre | produire |
| 13. évoluer | paraître | nombreux | moyen |
| 14. souci | chercher | lendemain | individu |
| 15. assemblée | poste | intéressant | coûter |
| 16. élire | fermer | sembler | avenir |
| 17. juger | identité | représenter | valoir |
| 18. particulier | gérer | mériter | interdire |
| 19. relever | rentrer | annoncer | goût |
| 20. vitesse | vigueur | pouvoir | conduire |
| 21. prétendre | pratique | soir | amener |
| 22. animer | venir | transmettre | entretenir |
| 23. empêcher | attendre | dépendre | dur |
| 24. prestation | concerner | célèbre | apporter |
| 25. caractère | vente | apprendre | financement |
| 26. terminer | arriver | utile | foule |
| 27. prier | considérer | oublier | mener |
| 28. bonheur | mise | maladie | souhaiter |
| 29. feu | descendre | avocat | exiger |
| 30. poche | espérer | arrêter | monnaie |

# APPENDIX E

## ENGLISH AND FRENCH INSTRUCTIONS FOR THE WORD ASSOCIATION TASKS.

### INSTRUCTIONS

On the next page, you will see some words on the left side. Next to each word, **write down any other words in English** that it makes you think of. Write down **as many as you can** (at least **four**, if possible). It doesn't matter if the connections between the word and your words are not obvious; there are no right or wrong answers. Simply write down words as you think of them. You will have 15 minutes to complete this task. Try to fill the entire page if possible.

If you manage to fill the whole page before the time is up, then continue on the **second page**, which has the **same list of words**.

Please write as clearly as possible, one word per box, so that we will not have difficulty reading what you have written.

Example:

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 1 | **white** | *black* | *colour* | *snow* | *dress* | *chalk* | *clean* |

Thank you very much for participating in this study!

## INSTRUCTIONS

Vous verrez une série de mots alignés à gauche sur la page suivante. À droite de chaque mot, **écrivez les premiers mots qui vous viennent à l'esprit en Français** dans les boîtes prévues à cet effet. Écrivez **le plus de mots possibles** (au moins **quatre**). Vous êtes libre d'écrire les mots que vous voulez, même si le lien entre le mot donné et vos réponses n'est pas évident; il n'y a pas de bonnes ou de mauvaises réponses. Vous aurez 15 minutes pour effectuer cette tâche.

<u>Si vous finissez de remplir la page avant la fin des 15 minutes</u>, vous pouvez continuer l'exercice sur la **page suivante** qui contient à nouveau **les même mots**.

S'il vous plaît, écrivez le plus lisiblement possible, un mot par boîte.

Exemple:

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 1 | **blanc** | *noir* | *couleur* | *neige* | *robe* | *craie* | *propre* |

Merci beaucoup de votre participation!

## APPENDIX F

## FRENCH AND ENGLISH LIVING-NON-LIVING STIMULI

Table 15: *Training Stimuli and Test Stimuli used in the Living-Non-living Task in English and French*

| Training | English | | French | |
|---|---|---|---|---|
| black | grandfather | judge | acteur | jeu |
| blue | worker | journalist | adulte | jouet |
| brown | wolf | ice | animal | journaliste |
| eight | witness | hospital | appareil | lit |
| five | window | hat | arbre | livre |
| four | weapon | glass | armer | maison |
| green | visitor | friend | avocate | médaille |
| nine | truck | flower | billet | moteur |
| red | tool | fish | bouteille | navire |
| three | table | farmer | cadeau | officier |
| two | student | drawing | cahier | ordinateur |
| white | stone | door | ceinture | passager |
| | sister | dog | chanteur | patron |
| | shirt | doctor | chat | peinture |
| | secretary | desk | cheval | père |
| | roof | dancer | clé | piano |
| | rifle | citizen | couteau | pont |
| | president | chair | directeur | président |
| | plate | building | échelle | princesse |
| | plane | boy | écran | professeur |
| | piano | box | écrivain | reine |
| | person | boat | église | siège |
| | passenger | blanket | enfant | soldat |
| | paper | bird | femme | table |
| | officer | belt | fille | toit |
| | motor | ball | fils | usine |
| | mother | bag | frère | vache |
| | mirror | author | grand-père | vêtement |
| | man | artist | infirmière | voisin |
| | king | animal | invitée | voiture |